# A COMPARISON OF DATA MINING METHODS FOR PREDICTION AND CLASSIFICATION TYPES OF QUALITY PROBLEMS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZEYNEP ANAKLI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
INDUSTRIAL ENGINEERING

DECEMBER 2009

Approval of the thesis:

**A COMPARISON OF DATA MINING METHODS FOR PREDICTION AND CLASSIFICATION TYPES OF QUALITY PROBLEMS**

submitted by **ZEYNEP ANAKLI** in partial fulfillment of the requirements for the degree of **Master of Science in Industrial Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Gradute School of **Natural and Applied Sciences** _____

Prof. Dr. Nur Evin Özdemirel
Head of Department, **Industrial Engineering** _____

Prof. Dr. Gülser Köksal
Supervisor, **Industrial Engineering Dept., METU** _____

Assoc. Prof. Dr. Esra Karasakal
Co-Supervisor, **Industrial Engineering Dept., METU** _____

**Examining Committee Members:**

Prof. Dr. Sinan Kayalıgil
Industrial Engineering, METU _____

Prof. Dr. Gülser Köksal
Industrial Engineering, METU _____

Assoc. Prof.Dr. Esra Karasakal
Industrial Engineering, METU _____

Assoc. Prof. Dr.İnci Batmaz
Dept. of Statistics, METU _____

Assoc. Prof. Dr.Murat Caner Testik
Industrial Engineering, Hacettepe University _____

**Date**: 10 December 2009

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Last name : Zeynep Anaklı

Signature :

# ABSTRACT

A COMPARISON OF DATA MINING METHODS FOR PREDICTION AND CLASSIFICATION TYPES OF QUALITY PROBLEMS

Anaklı, Zeynep
M.S., Department of Industrial Engineering
Supervisor : Prof.Dr. Gülser Köksal
Co-Supervisor: Assoc. Prof.Dr. Esra Karasakal

December 2009, 182 Pages

In this study, an Analytic Network Process (ANP) and Preference Ranking Organization MeTHod for Enrichment Evaluations (PROMETHEE) based approach is developed and used to compare overall performance of some commonly used classification and prediction data mining methods on quality improvement data, according to several decision criteria.

Classification and prediction data mining (DM) methods are frequently used in many areas including quality improvement. Previous studies on comparison of performance of these methods are not valid for quality improvement data. Furthermore, these studies do not consider all relevant decision criteria in their comparison. All relevant criteria and interdependencies among criteria should be taken into consideration during the performance evaluation.

In this study, classification DM methods namely; Decision Trees (DT), Neural Networks (NN), Multivariate Adaptive Regression Splines (MARS), Logistic Regression (LR), Mahalanobis-Taguchi System (MTS), Fuzzy Classifier (FC) and Support Vector Machine (SVM); prediction DM methods DT, NN, MARS, Multiple Linear Regression (MLR), Fuzzy Regression (FR) and Robust Regression (RR) are

prioritized according to a comprehensive set of criteria using ANP and PROMETHEE.

According to results of this study, MARS is found superior to the other methods for both classification and prediction. Moreover, sensitivity of the results to changes in weights and thresholds of the decision criteria is analyzed. These analyses show that resulting priorities are very insensitive to these parameters.

Keywords: Classification, Prediction, Analytic Network Process, PROMETHEE, Data mining

# ÖZ

TAHMİN ETME VE SINIFLANDIRMA KALİTE PROBLEMLERİ ÖZELİNDE VERİ MADENCİLİĞİ METOTLARININ KARŞILAŞTIRILMASI

Anaklı, Zeynep
Yüksek Lisans, Endüstri Mühendisligi Bölümü
Tez Yöneticisi: Prof. Dr. Gülser Köksal
Ortak Tez Yöneticisi: Doç. Dr. Esra Karasakal

Aralık 2009, 182 Sayfa

Bu çalışmada, veri madenciliği literatüründeki en yaygın tahmin etme ve sınıflandırma metotlarının performansları Analitik Ağ Süreci (AAS) ve Zenginleştirme Değerlendirmeleri için Sıralama Organizasyon Metodu (PROMETHEE) kullanılarak, birden çok kritere göre kalite iyileştirme verisi özelinde değerlendirilmektedir.

Veri madenciliğinde sıklıkla kullanılan tahmin etme ve sınıflandırma metotlarının kalite iyileştirme de olmak üzere bir çok alanda uygulamaları bulunmaktadır. Çok sayıda tahmin etme ve sınıflandırma metodu vardır. Ancak, daha önce bu metotların karşılaştırılması için yapılan çalışmalar kalite iyileştirme verileri için geçerli olmayabilir. Üstelik bu çalışmalarda bütün karar kriterleri değerlendirilmemiştir. Metotların performansları değerlendirilirken bu kriterler ve aralarındaki etkileşimler de hesaba katılmalıdır.

Bu çalışmada sınıflandırma metotlarından karar ağaçları (DT), sinir ağları (NN), MARS, lojistik regresyon (LR), Mahalanobis-Taguchi Sistemi (MTS), bulanık

sınıflandırma (FC) ve destek vektör makinaları (SVM); tahmin etme metotlarından da DT, NN, MARS, çoklu doğrusal regresyon (MLR), bulanık regresyon (FR) ve robust regresyon (RR)  AAS ve PROMETHEE yöntemleri kullanılarak kapsamlı bir karar kriteri kümesi değerlendirilerek önceliklendirilmiştir.

Bu çalışmada elde edilen sonuçlara göre, hem sınıflandırma hem de tahmin etme metotları içinde, MARS en iyi metottur. Ayrıca, elde edilen sonuçların karar kriterlerinin ağırlıklarına ve metotlar arasındaki performans farkını ölçecek olan eşik değerine olan duyarlılığı değerlendirilmiştir. Analiz sonuçları, elde edilen önceliklendirme sonuçlarının  bu parametlere duyarlılığı olmadığını göstermiştir.

Anahtar Kelimeler: Sınıflandırma, Tahmin etme, Analitik Ağ Süreci,  PROMETHEE Veri Madenciliği

To My Parents

# ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my supervisor Prof. Dr. Gülser Köksal and my cosupervisor Assoc.Prof.Dr. Esra Karasakal for their guidance, advice, criticism, encouragements and insight throughout the research.

I would also like to thank Prof. Dr. Sinan Kayalıgil, Assoc. Prof. Dr. İnci Batmaz, Fatma Yerlikaya Özkurt, Elçin Kartal, Berna Bakır for their suggestions and comments. I am grateful for their interest, their explanations, and patiently answering all of my questions.

Barış Yenidünya, Dilber Ayhan, Gizem Özer, Ezgi Avcı, Tuna Kılıç and Süreyya Özöğür are gratefully acknowledged for providing a part of the data for the analysis.

I would also thank to my colleague Emrah Öz for his invaluable help and support. I am also grateful to my co-worker Burak Işıktan for his patience and understanding through the preparation of this manuscript.

Finally, I would like to thank to my brother Mete Anaklı for his encouragement, help and patience throughout my study.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| AHP | : | Analytic Hierarchy Process |
| AAS | : | Analitik Ağ Süreci |
| AN | : | Abductive Network |
| ANN-based | : | Artificial Neural Network-based |
| ANP | : | Analytic Network Process |
| AUC | : | Area Under (Receiver Operating Characteristics) Curve |
| AUROC | : | Area Under Receiver Operating Characteristics Curve |
| BNN | : | Bayesian Neural Network |
| BB | : | Backpropagation |
| C.I. | : | Consistency Index |
| C.R. | : | Consistency Ratio |
| CART | : | Classification and Regression Trees |
| CBR | : | Case-Based Reasoning |
| CHAID | : | Chi-Squared Automatic Interaction Detection |
| CI | : | Confidence Interval |
| CN | : | Customer Need |

| | | |
|---|---|---|
| CompetANN | : | Competitive ANN |
| DEMATEL | : | Decision Making Trial and Evaluation Laboratory |
| DM | : | Data Mining |
| DR | : | Design Requirements |
| DT | : | Decision Tree |
| DVR | : | Digital Video Recorder |
| ELECTRE | : | Elimination et choix traduisant la realite |
| EN | : | Entropy Network |
| FAN | : | Fuzzy Adaptive Network |
| FC | : | Fuzzy Classifier |
| FF | : | Fuzzy Functions |
| FR | : | Fuzzy Regression |
| FST | : | Fuzzy Set Theory |
| GAM | : | Generalized Additive Model |
| GLZ | : | Generalized Linear Models |
| IFN | : | Info-Fuzzy Network |
| KDD | : | Knowledge Discovery From Databases |
| KNN | : | K-Nearest-Neighbours |

LM : Linear Models

LMS : Least Median Squares

LPE : Large Prediction Error

LWR : Locally Weighted Regression

LR : Logistic Regression

MAPE : Mean Absolute Percentage Error

MARS : Multivariate Adaptive Regression Splines

MAUT : Multi Attribute Utility Theory

MCDM : Multiple Criteria Decision Making

MCR : Misclassification Rate

MDA : Multiple Discrimant Analysis

MLP : Multi-Layer Perceptron

MLR : Multiple Linear Regression

MSE : Mean Square Error

MTS : Mahalanobis-Taguchi System

NBC : Naïve Bayesian Classifier

NLR : Nonlinear Regression

NMI : Normalized Mutual Information

| | | |
|---|---|---|
| NN | : | Neural Networks |
| PCC | : | Percent of Correctly Classified |
| PLR | : | Penalized Logistic Regression |
| PNN | : | Probabilistic NN |
| PRESS | : | Prediction Sum Of Squares |
| PROMETHEE | : | Preference Ranking Organization MeTHod for Enrichment Evaluations |
| PWI | : | Proportion of Plots Within Some User-Specified Range |
| QI | : | Quality Improvement |
| R.I. | : | Random Consistency Index |
| RBF | : | Radial Basis Function |
| RecBFN | : | Rectangular Basis Function Network |
| RMSE | : | Root Mean Square Error |
| ROC | : | Receiver Operating Characteristic |
| RR | : | Robust Regression |
| RST | : | Rough Set Theory |
| SVM | : | Support Vector Machine |
| SWOT | : | Strengths, Weaknesses, Opportunities and Threats |
| TSA | : | Time Series Analysis |

# CHAPTER 1

# INTRODUCTION

In data mining (DM) literature, there are several applications of classification and prediction methods in a variety of areas. Quality Improvement (QI) and control is one of these areas, tasks such as parameter design, tolerance design, inspection and screening, quality monitoring and quality analysis. Köksal et al. (2008) review the DM applications in many of these quality tasks in manufacturing industry. According to Köksal et al. (2008), classification and prediction are the most frequently used DM functions used to perform these tasks, in the literature.

This study investigates which classification and prediction method should be preferred for specific QI and control problems. In this study, we focus on commonly used classification and prediction methods applied in performing the QI tasks selected and analyzed by Köksal et al. (2008).

The aim of this study is to comprehensively evaluate and compare performance of the selected classification and prediction methods on the selected quality problems to guide the QI practitioners and researchers.

An important part of a successful classification and prediction is selection of the most appropriate method. Even though several methods are available for these purposes, none of them has been labeled as the best one. According to Bradley (1997), there is no universally accepted ranking of these methods. For each problem or case, suggested method may change owing to the nature of the problem.

In the literature, there are many studies comparing and suggesting the classification and/or prediction methods for different problem areas (Manel, Dias and Ormerod,

1999; Brazdil and Soares, 2000; Moisen and Frescino, 2002; Bradley, 1997, Dhar and Stein, 1997, Patel, 2003, Köksal et al., 2008). Most of these comparisons are mainly based on the accuracy performance of the methods. Evaluation of a method cannot be limited with its one aspect such as accuracy performance.

Indeed, Dhar and Stein (1997) define different aspects such as explicability, flexibility, response speed and scalability to evaluate performance of several DM tools. However, these evaluations for each aspect are not aggregated within a context and methods are not evaluated according to their overall performances on those aspects.

Patel (2003) also defines several aspects to evaluate some DM methods' performance, but does not aggregate these evaluations to see the overall performance of the compared methods.

Moreover, evaluations in Dhar and Stein (1997) and Patel (2003), are not specific to a problem area. These evaluations represent methods' average performance on the defined aspects. These performances may change with different problem areas.

In this study we are interested in prediction or classification based on data collected for certain types of QI and control tasks. As Rokach and Maimon (2006) state, quality related data in manufacturing has its own characteristics. Thus, evaluating the applied data mining methods according to conventional ways is ineffective. Most significant characteristics of the quality related data are imbalanced classes (such as defective, non defective), curse of dimensionality (small sized of data, large number of variables) and mixed type of data.

The performance of the methods, according to aspects such as those defined by Dhar and Stein (1997) and Patel (2003), may change when applied on quality related data. Furthermore, QI practitioners may not emphasize a criterion such as, response speed as much as say call center customer service people. Thus, these evaluations should not be made independent of the context. An important difference of this study from the others is that all evaluations concerning the classification and prediction methods are made within the context of the selected QI and control problems.

Another important feature of this study is the methodology used to aggregate significant aspects of the problem context to compare and rank the classification and prediction methods. The ranking of the selected classification and prediction methods are obtained by following a mixture of two multi-criteria decision making approaches: Analytical Network Process (ANP) and Preference Ranking Organisation MeTHod for Enrichment Evaluations (PROMETHEE). There is an attempt to combine ANP with PROMETHEE in the literature (Bozkurt, 2007). Both approaches require decision makers' or experts' input in comparing and evaluating the methods. For this reason, experts with prior experience and background with the application of these methods on relevant data have contributed to this study. Information gathered from literature has also been utilized to facilitate the comparison and ranking process. Furthermore, results of applications of the selected methods on different data sets have been analyzed to provide additional information about relationships among some of the comparison criteria.

ANP method is used to find out relative priorities of the decision criteria with respect to the goal of the problem. ANP handles both interdependencies and feedbacks between the decision criteria and the goal. Thus, resulting priorities (represented as weights) may be different than the expected ones, since human brain cannot handle that much of a complexity.

ANP application results in relative weights of the decision criteria and then these weights are used as inputs of the PROMETHEE method, which we used to model preference of the decision makers. This method compares each pair of alternative classification and prediction methods with respect to each decision criterion and determines the outranking character of these methods. These characters are interpreted and the ranking of the alternative classification and prediction methods are obtained.

Next, sensitivity of the results to changes in weights and thresholds of the decision criteria is analyzed. In the sensitivity analysis, changes in the ranking of the alternative methods are studied with changes in the weights and thresholds one at a time.

It is important to point out that the main aim of this study is to compare and rank commonly used classification and prediction methods according to important criteria of the QI and control problems and suggests the most favorable ones especially for practitioners. Conclusions of this study may not be valid beyond the scope of this problem. This study investigates which classification and prediction method should be preferred for specific QI and control problems. These specific problems are also described in Chapter 2.

In the following sections, a literature review and background is given in Chapter 2. Several classification and prediction methods, performance measures used to assess these methods and the most common multi-criteria decision making tools are introduced in this chapter. Moreover, some background information is provided for the methods used in this study. In Chapter 3, the ranking approach of this study, ANP and PROMETHEE applications for classification and prediction methods are presented. Interpretation of the results and sensitivity analysis are also presented in this chapter. Finally, concluding remarks are provided in Chapter 5.

# CHAPTER 2

# LITERATURE REVIEW AND THEORETICAL BACKGROUND

In the recent years, knowledge discovery from databases (KDD) and data mining (DM) has been widely applied in various fields. Data mining can be defined as extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data (Han and Kamber, 2001). *Classification* and *Prediction* are the most familiar and popular data mining functions (Han and Kamber, 2001; Rokach and Maimon, 2005).

Data mining approaches can be applied in different areas and *Quality Improvement (QI) and control* is one of them. The literature presents several studies that examine the data mining methods in QI. For instance, Huang (2005) uses decision tree method to identify important factors influencing the percentage of defectives. Kang (2000) suggests integrated machine learning approaches for solving certain quality problems. Fan et al. (2001) integrate the concepts of quality control, data mining, and process knowledge. Köksal et al. (2008) provide a review of DM applications on selected QI problems.

During this study, the literature is reviewed in mainly three different areas:

1. Classification and prediction methods in DM
2. Performance measures of classification and prediction methods
3. Multi-criteria decision making approaches for comparing alternatives

## 2.1 Classification and Prediction Methods in Data Mining

Köksal et al. (2008b) present a review of DM applications on selected QI problems. These selected QI problems are product/process quality description, predicting quality, classification of quality and parameter optimization. In this study, we also focus on the classification and prediction DM methods' performance on these selected QI problems. Classification and prediction methods widely used in literature are listed in Table 2.1.

Classification is used to forecast future values of categorical data. Major classification algorithms are Statistical based (S-based) algorithms, Decision Tree-based (DT-based) algorithms, Artificial Neural Network-based (ANN-based) systems and other classification algorithms (Köksal et al., 2008b).

S-based algorithms such as generalized linear models (GLZ) such as logistic regression (LR) and naïve bayesian classifier (NBC) use statistical techniques. DT-based algorithms generate rules by using 'if then' type structures. ID3, C4.5, C5.0, Chi-squared automatic interaction detection (CHAID) and classification and regression tree (CART) are widely used DT based algorithms. ANN-based systems are used to model complex relationships. ANN-based systems can represent both linear and non-linear relationships. They consist of input, hidden and output layers and between these layers there are weighted connections. These weights are updated during the learning phase. Widely used forms of ANN are perceptron, radial basis function (RBF), competitive ANN (CompetANN), Probabilistic NN (PNN), bayesian NN (BNN) and rectangular basis function network (RecBFN) (Köksal et al., 2008b).

Other classification algorithms consist of k-nearest-neighbours (KNN) which is a distance based algorithm, Genetic Algorithm (GA) combined with rough set theory (RST), fuzzy set theory (FST), support vector machines (SVM), entropy network (EN) and association rule-based methods (Köksal et al., 2008b).

**Table 2.1 Classification and prediction methods commonly used for quality problems**

| Methods | Algorithm | Some References |
|---|---|---|
| Neural Network (NN) based methods<br>(for both classification and prediction) | MLP with BP | Braha and Shmilovici (2002); Kim et.al. (2003); Han et. al (1999) |
| | SOM | Braha and Shmilovici (2002) |
| | Recurrent NN<br>Feedforward BB | Kim and Lee (1997) |
| Decision Tree (DT) based methods<br>(for both classification and prediction) | C4.5 | Braha and Shmilovici (2002) |
| | CHAID | Huang and Wu (2005) |
| | C5.0 | Bakır et al. (2007); Huang and Wu (2005) |
| | CART | Li et al. (2003); Chien et. al. (2006) |
| Statistical based methods<br>(for prediction ) | MLR | Kim and Lee (1997) |
| | NLR | Kim et al. (2003) |
| Multivariate Adaptive Regression Splines ( MARS)<br>(for both classification and prediction) | MARS | Yerlikaya (2008)<br>Uysal and Güvenir (1999) |
| Logistic Regression (LR)<br>(for classification) | LR | Yenidünya (2009) ; Grimm and Yarnold (1994); Agresti (1996) |
| The Mahalanobis-Taguchi System (MTS)<br>(for classification) | MTS | Taguchi et al. (2001); Ayhan (2009) |
| Fuzzy Classifier (FC) (for classification) | FC Functions | Özer (2009) |
| | Fuzzy Rule Based Methods (FRBM) | Meier et al.(2007) |
| Support Vector Machine (SVM)<br>(for classification) | SVM | Cristianini and Taylor (2000) |
| Fuzzy Regression (FR) (for prediction ) | Fuzzy Functions | Kılıç (2009); Ip et. al. (2003) |
| Robust Regression (RR) (for prediction ) | Huber-M | Avcı (2009) |
| | LMS | Ortiz et al. (2006) |

Prediction is performed to forecast future values of continuous type data. Köksal et al. (2008b) list the major prediction algorithms as follows: S-based methods, DT-

based methods, ANN-based methods and others. Multiple Linear Regression (MLR), nonlinear regression (NLR), time series analysis (TSA) and response surface methodology (RSM) are widely used statistical based methods. CART is the DT based method used in predicting quality. The most widely used ANN based methods for predictions are multi-layer perceptron (MLP), RBF and BNN. Case-based reasoning (CBR), fuzzy adaptive network (FAN), info-fuzzy network (IFN) and abductive network (AN) are among the other methods used for predicting quality.

In addition to these methods mentioned above there is a relatively new technique Multivariate Adaptive Regression Splines (MARS). MARS is a very powerful regression based method used to fit models especially to large and complex data sets (Uysal and Güvenir, 1999; Köksal et al., 2008b, Yerlikaya, 2008).

Compared classification and prediction methods in this study, accuracy data collected from their application on two data sets are available at Köksal et. al. (2009).

There are some studies comparing some classification and prediction methods, these studies and suggested methods are given in Table 2.2. These studies are applied in different areas and different data sets are used accordingly, such as diagnostics and ecological data. Suggested classification or prediction methods may differ according to problem area and the data.

Some DM tools and their evaluations performed by Dhar and Stein (1997) are illustrated in Table A.1 in APPENDIX A. Patel (2003) also presents a comparison study and findings of this study is illustrated in Table A.2 in APPENDIX A.

**Table 2.2 Studies comparing and suggesting the some classification and prediction methods**

| Reference | Data sets | Methods Compared | Criteria | Suggest |
|---|---|---|---|---|
| Moisen and Frescino (2002) | Forest inventory field data and ancillary satellite-based information | - LR<br>- Generalized Additive Models (GAMs) (including MLR)<br>- CART<br>- MARS<br>- NN | -PCC<br>-Kappa<br>-RMSE<br>-PWI<br>-Run time | MARS and ANN for prediction; MARS and GAMs |
| Razi and Athappilly (2005) | A set of data on the smoking habits of people | - MLR<br>- NN<br>- CART | -MAE<br>-MAPE<br>-MSE<br>-Large Prediction Error (LPE) | NN or CART rather than MLR |
| Muñoz and Felicísimo (2004) | Two typical ecological data sets representative of typical ecological data sets | - LR<br>- Principal Component Regression<br>- CART<br>- MARS | The area under Receiver Operating Characteristics curve (AUC) | MARS and CART |
| Zhu and Hastie (2004) | Three cancer diagnosis data sets | - Penalized logistic regression (PLR)<br>- SVM | -PCC<br>-Number of parameters<br>-Cross-validation error | PLR |
| Braha and Shmilovici (2002) | Yield data generated during daily semiconductor manufacturing | - DT<br>- NN<br>- Composite classifiers | PCC | Composite classifiers |
| Bradley (1997) | Six "real world" medical diagnostics data sets | - C4.5<br>- Multiscale Classifier<br>- Perceptron<br>- NN (MLP)<br>- KNN<br>- Quadratic Discriminant Function | AUROC | Bayes, MLP and KNN based methods rather than the DT |
| Huang et al. (2003) | Six medical diagnostics data sets | - NBC;<br>- DT;<br>- SVM | -PCC<br>-AUROC | NBC, C4.4, and SVM have similar scores and outperform C4.5 |
| Manel et al. (1999) | Ecological data | - Multiple discrimant analysis (MDA);<br>- LR;<br>- NN | -Specificity<br>-Sensitivity<br>-ROC plots | LR and MDA |
| Köksal et al. (2008a) | Customer satisfaction data | -LR<br>-NN<br>-MARS | - R Square<br>-PCC<br>-Log Odds Ratio<br>-Kappa<br>-AUC | MARS performs better and LR competes with it |

**Table 2.2 (Continued) Studies comparing and suggesting the some classification and prediction methods**

| Reference | Data sets | Methods Compared | Criteria | Suggest |
|---|---|---|---|---|
| Lim et. al.(2000) | Thirty-two different data sets | -DT<br>-NN<br>-Statistical classifiers | -Mean Error Rate<br>-Mean rank of error rate<br>-Training times | C4.5 , IND-CART and QUEST algorithms are suggested among twenty two different DT algoritms; POLYCLAS algorithm is suggested among nine Statistical classifiers |
| Uysal and Güvenir (1999) | Unspecified | -LWR<br>-Rule based regression<br>-Projection pursuit regression<br>-KNN<br>-MARS<br>-CART | -Mean Absolute Distance<br>-Adaptive<br>-Incremental<br>-Interpretable<br>-Memory based<br>-Partitioning | Rule based regression and MARS |
| West et al. (1997) | Simulation data (customer patronage behaviour) | -NN<br>-LR<br>-Discriminant Analysis | -within and out of sample predictive accuracy | NN |
| Stolzer and Halford (2007) | Fuel consumption data of an aircraft | -DT<br>-NN<br>-MLR | -MSE<br>-MAE<br>-Mean Relative Squared Error (MRSE)<br>-Mean Relative Absolute Error (MRAE)<br>-Correlation Coefficient | NN and MLR |
| Shang et al. (2000) | MRSA (resistant to penicillin) diagnosis data | -LR<br>-NN | -AUC<br>-Robustness (Cross Validation) | NN |

## 2.2 Performance Measures

Prediction accuracy is the most common performance measure used in the literature (Hyndman and Koehler, 2006; Braha and Shmilovici, 2002; Han and Kamber,2001; Dhar and Stein,1997; Patel, 2003). In general, for classification methods percent of correctly classified (PCC) or misclassification rate (MCR), and for prediction methods mean square error (MSE) are used to measure accuracy. Han and Kamber (2001) use different accuracy measures such as recall (sensitivity), precision and specificity. These measures are also used in several sources (Fielding and Bell, 1997; Fawcett, 2004).

Fawcett (2004) uses F measure which is a weighted combination of precision and recall that produces scores ranging from 0 to 1. Instead of using both precision and recall, using F measure is suggested by Billsus and Pazzani (1998).

Fielding and Bell (1997) suggest normalized mutual information (NMI), kappa and odds ratio in addition to recall (sensitivity), precision, specificity and other standard accuracy measures. Kappa, which is a measure of agreement, and odds ratio are also suggested by Agresti (1996).

Besides accuracy, Dhar and Stein (1997) list the important performance measures as compactness, computing resource, development speed/effort, ease of use, embeddability, explicability, flexibility, independence from expert, learning curve requirements, response speed, scalability, tolerance for complexity, tolerance for data sparseness and tolerance for noise in data. Han and Kamber (2001) state that in addition to accuracy, classifiers can be compared with respect to their speed, robustness, scalability, and interpretability.

Application time, cost of obtaining labeled data, expert evaluation and field testing are some other measures suggested by Weiss and Zhang (2003).

The percent of correctly classified, kappa statistics, root mean square error (RMSE), proportion of plots within some user-specified range (PWI), correlation coefficient

and also the amount of time it took to run each model are used as evaluation criteria by Moisen and Frescino (2002).

Mean square errors (MSE), R square and adjusted R square are widely used measures to evaluate adequacy of the prediction methods (Montgomery and Runger, 1996). Mean absolute percentage error (MAPE) is suggested by Kim and Lee (1997). The area under the curve (AUC) of a receiver operating characteristic (ROC) plot and kappa statistics are used to evaluate model performance in Virkkala et al. (2005). According to Bradley (1997) AUC is one of the best ways to evaluate a classifier's performance since it seems to be the only one that is independent of the decision threshold and not biased by prior probabilities. Mallow's Cp and prediction sum of squares (PRESS) are measures which are described in Mosteller and Tukey (1977).

Besides the measures mentioned above, Patel (2003) construct a measure list and evaluate the performance of Multiple Linear Regression, Logistic Regression, Discriminant Analysis, Naïve Bayes, Neural Networks, Trees and k-Nearest Neighbors methods according to these measures. Interpretability, speed-deployment and speed-training, robustness to outliers in independent variables, robustness to irrelevant variables, robustness to missing values and effort to tune performance parameters are some of the criteria used by Patel (2003).

Literature presents numerous performance measures and the whole list of performance measures found out during this study is given in APPENDIX A. In this section, only widely used measures are mentioned. Important measures in QI context and their definitions are given in APPENDIX B.

### 2.3 Ranking and Prioritization Methods

Multiple criteria decision making (MCDM) techniques support the decision makers to deal with decision problems that involve multiple and conflicting criteria. Since more than one criterion is evaluated, there is no optimal decision but a satisfactory one. MCDM methods use decision maker preferences to make recommendations (Guitouni and Martel, 1998).

Discrete multiple criteria problem is one where decision space consists of finite set of alternatives and also the criteria set is explicitly known. There are several MCDM methods. Type of data they use and the number of the decision makers are taken into consideration during the selection of the appropriate method (Triantaphyllou, 2000). MCDM methods for discrete problems can be assigned to one of the following categories: multi attribute utility theory (MAUT) methods, outranking methods and interactive methods and others (Pardalos et al. 1995).

MAUT is an extension of the classical utility theory. Its aim is to model the decision maker's preferences through a utility function *u* aggregating all the decision criteria. Analytical hierarchy process (AHP) can be included in MAUT (Pardalos et al. 1995; Belton and Stewart, 2002).

AHP is based on three main principles; hierarchy construction, priority setting and logical consistency checking. AHP is introduced by Saaty in 1980s and this method is widely used to solve complicated problems with multiple decision criteria.

Yoon and Hwang, (1995) states that there are two methods to deal with qualitative data: the median ranking method which use ordinal (rank) data, and AHP method which accept paiwise comparison data.

At the beginning of the AHP application, problem is analyzed to determine goal, decision criteria, their sub-criteria and alternatives. Next, a uni-directional hierarchy is constructed. Constructed hierarchy is a linear hierarchy, with a goal at the top level. Main criteria and their sub-criteria are placed into the second and third level of the hierarchy, respectively. At the bottom level, there are alternatives.

After constructing the hierarchy, comparisons are performed. Elements at each level are compared with each other with respect to the elements in the upper level. Upper level element is used as a decision criterion during the comparison of the elements in lower level. Saaty's nine point scale is used (Saaty, 2000) during these pairwise comparisons. Pairwise comparison matrices are developed for each decision criteria namely for each set of pairwise comparison. By using the eigenvector method

relative priorities of the compared elements are calculated. These priorities are used to determine overall priorities of the alternatives.

The most important drawback of the AHP is independency assumption of the elements in the hierarchy. AHP is not sufficient to model the problems with dependent elements.

The Analytical Network Process (ANP) is a generalization of the Analytic Hierarchy Process (AHP) and developed by Saaty (2000). ANP deals with problems with interdependent elements. Feedbacks and dependencies can be represented by networks. Since Saaty (1999) introduced ANP, this subject has been extensively studied and successfully applied to various fields such as undesirable facility location selection problems (Tuzkaya et al. 2008), solid waste management (Khan and Faisal, 2008), the identification of an organization's strategic management concepts (Asan and Soyer, 2008), Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis (Yüksel and Dağdeviren, 2007), evaluating Digital video recorder (DVR) systems (Chang et al., 2007), landslide hazard assessment (Neaupane and Piantanakulchai, 2006), evaluation of emergency bridge (Sun et al, 2007) and so on. Yet, performance evaluation of the Data Mining Tools is not one of these fields.

Main and the most important difference in these ANP applications is the constructed network structures. For instance, in Büyüközkan et al. (2004) ANP is used to prioritize design requirements (DRs) by taking into account the degree of the interdependence between the customer needs (CNs) and DRs and the inner dependence among them. Constructed network structure is a three level hierarchy (goal, criteria and alternatives) with inner dependence within components and no feedback.

The ANP model in Khan and Faisal (2008) is used to prioritize and select appropriate municipal solid waste disposal methods. The network consists of five levels. The first level is the decision problem of prioritizing the waste disposal methods namely goal. The second level is that of the actors that influence the prioritizing of the alternatives. The third level is the criteria or the determinants upon which the prioritization of the waste disposal method is broadly based. The next level consists

of the enablers or the sub-criteria that support the determinants and the last level is the alternatives. Feedback occurs between the actors and the criteria and there is interdependence among the sub-criteria in the fourth level of the network.

The proposed ANP network in Asan and Soyer (2008) consists five clusters and there is no hierarchical structure and there is only outer dependency between these clusters. Elements in the clusters are independent from each other.

ANP is applied to a supplier selection problem in Gencer and Gürpınar (2007). Resulting ANP network consists four clusters and one of them is alternatives cluster. Alternatives cluster has only outer dependencies but the other three clusters have both inner and outer dependencies. There are also feedbacks in this network structure.

ANP is used in a Strengths, Weaknesses, Opportunities and Threats (SWOT) analysis by Yüksel and Dağdeviren (2007). They propose a network structure having both hierarchy and network with four levels. In this network, the goal (best strategy) is indicated in the first level, the criteria (SWOT factors) and sub-criteria (SWOT sub-factors) are found in the second and third levels, respectively, and the last level is composed of the alternatives (alternative strategies). Indeed it is a case of a hierarchy with inner dependence within clusters but no feedback. There is inner dependence only in the second level SWOT factors.

Fuzzy ANP is also presented by literature (Dağdeviren et al., 2008; Promentilla et al. 2008; Yu and Cheng, 2007). According to Dağdeviren et. al. (2008) Fuzzy ANP is more appropriate to deal with measuring the qualitative factors since evaluating their performance with fuzzy numbers is easier and it gets more realistic results.

Büyüközkan et. al.(2004) states that Fuzzy ANP method offers more precise analysis but it requires more time and resource and indeed determining the correct fuzzy numbers in the ANP approach may need additional work.

The Decision Making Trial and Evaluation Laboratory (DEMATEL) is an alternative method modeling the cause effect relationships and also handling inner dependencies

within a set of criteria. DEMATEL method was conducted in 1973, by the Battelle Memorial Institute through its Geneva Research Centre and has been succesfully applied to various fields such as solid waste management (Tseng and Lin, 2009), knowledge management (Wu, 2008), competency development of global managers (Wu and Lee, 2007) and so on. DEMATEL constructs interrelations among criteria. In the DEMATEL process, an appropriate threshold value is important in order to obtain adequate information to define the impact-relations map for further analysis and decision-making. A theoretical method to aid in deciding the threshold value is necessary (Li and Tzeng, 2009). Indeed, DEMATEL is a useful method to define relationships between criteria and also influence directions, and it can be used to raise the accuracy of applying ANP (Tsai and Chou, 2007).

The second category of the MCDM methods is "the outranking methods". ELECTRE (Elimination et choix traduisant la realite) and PROMETHEE (Preference Ranking Organisation MeTHod for Enrichment Evaluations) are widely used outranking methods. ELECTRE was introduced by Benayoun et.al. (1966) and PROMETHEE was introduced by Brans (1982). In literature there are other outranking based approaches such as ORESTE but they are considered as ordinal (Guitouni and Martel, 1998).

In the ELECTRE methods, preference and indifference thresholds are required to aggregate criteria. There are several extensions of the ELECTRE used in the literature. In ELECTRE II, to develop two extreme outranking relations, which are strong and weak relations, multiple levels of concordance and discordance are considered. ELECTRE III is very similar to ELECTRE II, but fuzzy set theory is used to derive outranking relation. To handle the imprecision and uncertainty involved in the evaluation of alternatives ELECTRE IV is developed. These extensions are used in several studies in literature (Belton and Steward, 2002; Ertay and Kahraman, 2007; Pardalos et al. 1995).

PROMETHEE developed by Brans (1982) and further extended by Brans and Vincke (1985). Brans et al. (1986) states that PROMETHEE is superior to ELECTRE since it is more stable to the threshold deviations.

Halouani et al. (2009) suggests PROMETHEE method among the MCDM methods, to project selection problem with quantitative and qualitative data. Dulmin and Mininno, (2003) also applies PROMETHEE to qualitative and quantitative criteria.

Al-Kloub et. al. (1997) states that PROMETHEE outstands among other ranking methods, since it is software driven, user-friendly, provides direct interpretation of parameters, and a sensitivity analysis of results.

There are also studies which are combination of the two MCDM approaches. Macharis et al.(2004) combines AHP and PROMETHEE approaches to strengthen the PROMETHEE with ideas of AHP. Also, Dağdeviren (2008) proposes an AHP-PROMETHEE integrated approach for the equipment selection problem. Bozkurt (2007) combine ANP with PROMETHEE.

Lastly, interactive local judgment approach with trial and error iterations are categorized as interactive methods by Pardalos et al. (1995). Belton and Stewart (2002) categorize indirect and interactive value function methods and goal and reference point methods as interactive methods. In these methods, a complete preference model is not constructed; rather a sequence of real or hypothetical alternatives is presented to decision makers and they indicate their preferences between recently seen alternatives.

The main aim of this study is to determine the overall performance of the classification and prediction methods and suggest the most favorable ones. During their performance evaluation, several criteria, which are both qualitative and quantitative, are taken into account. Most of these criteria are correlated. Their relative importance according to the decision makers in Quality Improvement context is needed in order to prioritize the Classification and Prediction Methods.

Analytical Network Process (ANP) is the most suitable method to consider qualitative and quantitative criteria as well as the interdependencies and feedbacks. PROMETHEE is also selected as the most suitable method among other ranking methods. Thus, to prioritize the related Classification and Prediction Methods ANP and PROMETHEE will be used.

## 2.4 Analytical Network Process (ANP)

In this study main aim is to determine the overall performance of the classification and prediction methods with respect to decision criteria that are selected from the literature after discussion sessions of the decision makers and the help of the statistical analyses and then choose the most favorable methods. Since there are lots of interdependent decision criteria, and these criteria are also mixed type, in other words qualitative and quantitative.

Most of the possible decision criteria in literature are statistically or conceptually dependent to each other and there are also feedbacks. Ignoring these facts, and treating the problem as a simple decision problem, which can be structured hierarchically in a unidirectional relationship among decision levels, leads to misleading results.

Analytical Network Process (ANP) is suggested as the most suitable method to organize qualitative and quantitative criteria as well as the dependencies and feedbacks (Saaty, 1999). ANP is especially suited for modeling the interdependencies, feedbacks and hierarchies in the problem. It uses networks to identify relationships among the components of the problem. This network structure enables to represent and analyze relations and synthesize their mutual effects by a single method (Saaty, 2000). With the ANP, we gathered the weights of the criteria with respect to the goal and these weights are more than simple priorities. ANP evaluates the interdependencies, feedbacks and hierarchies, and then reaches the absolute priority of any criterion regardless of which criteria it influences. Resulting priorities are determined according to influences between criteria. For instance, a criterion influencing the most of the remaining criteria has higher priority than a criterion mostly influenced by other criteria. Thus resulting priorities should be evaluated accordingly since they represent more than straightforward rankings of the criteria.

Steps of the ANP are as follows:

**Step 1:** Main goal of the study is clearly stated and then, accordingly decision criteria are selected from the related problem context. This is the most important step, since decision makers should decide on the significant aspects of the problem. Number of decision criteria should be kept as low as possible to avoid asking too many pairwise comparisons.

**Step 2:** Important aspects, that a classification or prediction method is supposed to have, are determined and then decision criteria are labeled as sub criteria of these aspects and clusters are formed accordingly.

**Step 3:** Questions of the pairwise comparisons are formulated in terms of dominance or influence. There are two types of question:

Given a parent element and comparing elements 1 and 2 under it, which element has greater influence on the parent element?

Given a parent element and comparing elements 1 and 2, which element is influenced more by the parent element?

Make sure the questions on influencing or being influenced by are posed in a consistent way throughout the exercise. Whether the influence is flowing from the parent element to the elements being compared, or the other way around, must be kept in mind.

Then, a relation matrix is constructed. Dependencies are determined with respect to the main goal and only existence of the dependencies is illustrated in the relation matrix.

**Step 4:** Based on the constructed relation matrix and formed clusters, the network is structured. Dependencies between criteria are represented by the arrows. Links between criteria in the same cluster are called inner dependencies. Links between a criterion in one cluster and a criterion in another cluster are called outer dependencies.

Clusters are also connected by an arrow, if there is a link from at least one criterion of a cluster to at least one criterion of another cluster.

This network is used to determine pairwise comparisons that are needed to be performed by the decision makers.

**Step 5:** Pairwise comparisons are performed by the decision makers. If there is a goal node in the network, a pairwise comparison may be needed to see the contribution of the other elements (criteria or clusters) to the goal.

**Table 2. 3 Saaty's Nine Point Scale**

| | Intensity of Importance |
|---|---|
| Equally preferred | 1 |
| Moderately preferred | 3 |
| Strongly preferred | 5 |
| Very strongly preferred | 7 |
| Extremely preferred | 9 |
| 2,4,6,8 are intermediate values. | |
| **Reciprocals of above**<br>if activity *i* has one of the above nonzero numbers assigned to it when compared with activity *j*, then *j* has the reciprocal value when compared with *i.* | |

During the pairwise comparisons, Saaty's nine point scale is used (Saaty, 2000). This scale is represented in Table 2.3.

Pairwise comparison matrix is constructed and filled up by the decision makers for each parent element as in Figure 2. 1.

In Figure 2. 1., to fill the cell with the question mark in it, decision makers answer a question such as "Compare the element 2 and element 3 with respect to their influence on the element 1. Decide how much more the influence of the element 2 (element in the row) than the influence of the element 3 (element in the column). (Question type is determined in the Step 3 and used throughout the pairwise comparisons)

| **With respect to Element 1** | **Element 2 (E2)** | **Element 3 (E3)** | **…** |
|---|---|---|---|
| **Element 2 (E2)** | 1 | ? | … |
| **Element 3 (E3)** | 1/? | 1 | … |
| **…** | … | … | … |

**Figure 2. 1 A pairwise comparison matrix representation (element can be a criterion or a cluster)**

**Step 6:** The local priorities are calculated from each pairwise comparison matrix by using the eigenvector method.

The matrix of paired comparisons leads to the condition:

$$Aw = \lambda_{Max} w \qquad (2.\ 4.\ 1)$$

A non-zero solution implies that the determinant $|A - \lambda_{Max} I|$ is equal to zero. $\lambda_{Max}$ is a root of the equation obtained by setting the determinant to zero and called the eigenvector of A. Principle eigenvector *w* is a positive column vector.

Consistency of the comparisons should be checked. Consistency Ratio (C.R.) measures the consistency of the pairwise comparison matrix. C.R. is obtained by forming the ratio of Consistency Index (C.I) and Random Consistency Index (R.I). (Saaty, 2000)

$C.R. = \frac{C.I.}{R.I.}$ where

$C.I. = \frac{\lambda_{Max} - n}{n-1}$ and $R.I. = 1.98\frac{n-2}{n}$ ( $n$ is the size of the matrix )

If C.R. is greater than 0.10, an adjustment is needed to improve the consistency. (Saaty, 2000)

**Step 7:** Eigenvectors are calculated and placed into the Unweighted Supermatrix.

This Supermatrix is a partitioned matrix, columns of which contain the local priorities derived from the pairwise comparisons. This matrix shows all interactions between elements in the problem. If there is no interaction, corresponding entries in the matrix are zero. Elements can interact along more than a single path and the priorities of these elements are measured over all the paths and cycles which connect them.

A standard supermatrix structure is illustrated in Figure 2. 2. This structure may change according to the network structure.

Unweighted Supermatrix should be normalized to make the matrix column stochastic (each of its columns sums to unity). Resulting matrix is called as Weighted Supermatrix. This is required to ensure convergence of the matrix when the Weighted Supermatrix is raised to limiting powers.

| Cluster Node labels | | Goal Cluster | Criteria Cluster | | | | Alternatives Cluster | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Goal Node | Criterion 1 | Criterion 2 | Criterion 3 | Criterion ... | Alternative1 | Alternative2 | Alternative... |
| Goal Cluster | Goal Node | | | | | | | | |
| Criteria Cluster | Criterion 1 | | | | | | | | |
| | Criterion 2 | | | | | | | | |
| | Criterion 3 | | | | | | | | |
| | Criterion ... | | | | | | | | |
| Alternatives Cluster | Alternative1 | | | | | | | | |
| | Alternative2 | | | | | | | | |
| | Alternative... | | | | | | | | |

**Figure 2. 2 A standard supermatrix structure**

**Step 8:** The Weighted Supermatrix is raised to powers and the Limit Matrix is obtained. From the Limit Matrix the final priorities, which are steady state priorities, are extracted with respect to the goal.

## 2.5 PROMETHEE

PROMETHEE (Preference Ranking Organization METHod for Encrichment Evaluations) is a sequencing method which was introduced by Brans (1982). Priorities can be calculated with this method. Steps of the PROMETHEE method are, as follows (Brans and Vickle, 1986).

**Step 1:** Data matrix is constructed. Data matrix structure and notation are as follows:

Notation:

I: Set of alternatives, (1, 2… i)

A: Alternatives ($A_1$, $A_2$ …$A_i$),

K: Set of criteria (1, 2… k),

f: real valued criteria ($f_1$, $f_2$, …, $f_k$),

w: Weights of criteria ($w_1$, $w_2$,…,$w_k$) , namely relative importance of criterion $f_k$.

| Alternatives / Criteria | $A_1$ | $A_2$ | $A_3$ | ... | w |
|---|---|---|---|---|---|
| $f_1$ | $f_1(A_1)$ | $f_1(A_2)$ | $f_1(A_3)$ | ... | $w_1$ |
| $f_2$ | $f_2(A_1)$ | $f_2(A_2)$ | $f_2(A_3)$ | ... | $w_2$ |
| … | ... | ... | ... | ... | … |
| $f_k$ | $f_k(A_1)$ | $f_k(A_2)$ | $f_k(A_3)$ | ... | $w_k$ |

**Figure 2. 3 Data matrix structure**

**Step 2:** Preference functions are determined for each criterion by the decision makers according to properties of the criterion.

If two alternatives; $A_1$ and $A_2$ Є A are compared, the result of this comparison is represented as preference. Let a preference function is represented as P:

$P: A \times A \rightarrow (0, 1)$

Preference function represents the intensity of the preference of $A_1$ over $A_2$.

$P_j(A_1, A_2) = 0$ means an indifference of $A_1$ over $A_2$ for criterion j;

$P_j(A_1, A_2) \approx 0$ means weak preference of $A_1$ over $A_2$ for criterion j;

$P_j(A_1, A_2) \approx 1$ means strong preference of $A_1$ over $A_2$ for criterion j;

$P_j(A_1, A_2) = 1$ means strict preference of $A_1$ over $A_2$ for criterion j.

Preference function is a function of the difference between two alternatives. This can be represented as follows:

$P_1(A_1, A_2) = P(f_1(A_1) - f_1(A_2))$ for criterion $f_1$

$d = f_1(A_1) - f_1(A_2)$

Preference function has to be a non-decreasing function and equal to zero for negative values of *d*.

There are six different preference functions and they are illustrated in APPENDIX C.

After determining the preference function, related parameters of these selected functions should be set by the decision makers.

**Step 3:** The multi-criteria preference index $\prod$ is defined as the weighted average of the preference function P :

$$\Pi(A_1, A_2) = \frac{\sum_{j=1}^{k} w_j P_j(A_1, A_2)}{\sum_{j=1}^{k} w_j} \qquad (2.5.1)$$

$\Pi(A_1, A_2)$ represents the DM's preference intensity of alternative $A_1$ over $A_2$ by considering all criteria at the same time.

$\Pi(A_1, A_2) \approx 0$ means weak preference of $A_1$ over $A_2$ for all the criteria,

$\Pi(A_1, A_2) \approx 1$ means strong preference of $A_1$ over $A_2$ for all the criteria.

**Step 4:** For each alternative leaving and entering flows are defined. Entering flow measures the outranked character of the alternatives. Leaving flow measures the outranking character of the alternatives.

Let:

$\Phi^{-}(A_1)$: Entering flow of Alternative 1

$$\Phi^{-}(\mathbf{A_1}) = \sum_{i \in I} \Pi(A_i, A_1) \qquad (2.5.2)$$

$\Phi^{+}(A_1)$: Leaving flow of Alternative 1

$$\Phi^{+}(A_1) = \sum_{i \in I} \Pi(A_1, A_i) \qquad (2.5.3)$$

$\Phi\ (A_1)$: Net flow of Alternative 1

$$\Phi\ (A_1) = \Phi^{+}(A_1) - \Phi^{-}(A_1) \qquad (2.5.4)$$

Leaving and entering flows are represented graphically respectively in Figure 2. 4 and Figure 2. 5.

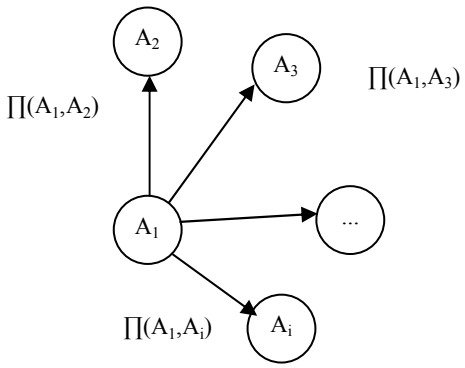



**Figure 2. 4 Leaving flow from Alternative 1**

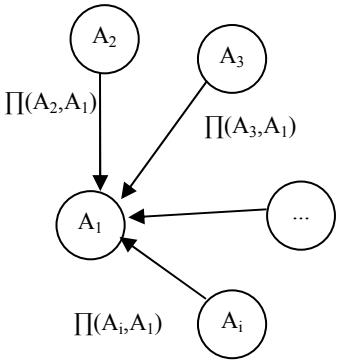



**Figure 2. 5 Entering flow to Alternative 1**

**Step 5:** Partial preorders or complete preorders are determined with PROMETHEE.

Partial preorders are determined with the PROMETHEE I. The higher the leaving flow and the lower the entering flow, the better that alternative.

If one of the following conditions is satisfied, $A_1$ is preferred to $A_2$:

$\Phi^+(A_1) > \Phi^+(A_2)$ and $\Phi^-(A_1) < \Phi^-(A_2)$

$\Phi^+(A_1) > \Phi^+(A_2)$ and $\Phi^-(A_1) = \Phi^-(A_2)$

$\Phi^+(A_1) = \Phi^+(A_2)$ and $\Phi^-(A_1) < \Phi^-(A_2)$

If following condition is satisfied, $A_1$ is indifferent to $A_2$:

$\Phi^+(A_1) = \Phi^+(A_2)$ and $\Phi^-(A_1) = \Phi^-(A_2)$

If one of the following conditions is satisfied, $A_1$ and $A_2$ are incomparable:

$\Phi^+(A_1) > \Phi^+(A_2)$ and $\Phi^-(A_1) > \Phi^-(A_2)$

$\Phi^+(A_1) < \Phi^+(A_2)$ and $\Phi^-(A_1) < \Phi^-(A_2)$

Complete preorders are determined with PROMETHEE II. Net flows are used as follows:

If $\Phi\ (A_1) > \Phi\ (A_2)$, $A_1$ outranks $A_2$,

If $\Phi\ (A_1) = \Phi\ (A_2)$, $A_1$ is indifferent to $A_2$.

Indeed, partial preorders are more useful then complete preorders since they can give more realistic information. Partial preorder gives the incomparability information that is not obtained by the complete preorders.

# CHAPTER 3

# RANKING OF DM METHODS

In order to prioritize the classification and prediction methods applied to quality data, several interdependent decision criteria are taken into consideration. Relative weights of these criteria are determined using the ANP method. These weights are used as input to the PROMETHEE method. PROMETHEE tries to model decision makers' preference function. Based on the some thresholds priorities of the alternative classification and prediction methods are determined by means of PROMETHEE. In this chapter, this mixture of two multi-criteria decision making approaches used for the raking is explained in detail as well as its application results. Furthemore results of a sensitivity analysis performed for the criteria weights and thresholds are presented.

## 3.1. The Approach

As a first step of the ANP application, main goal is clearly stated as "Determining the overall performance of the classification/prediction methods and suggesting the most favorable ones."

Both ANP and PROMETHEE require decision makers' input in comparing and evaluating the methods. Decision makers consulted with in this study are listed in APPENDIX D. They are all experienced in application of several classification and prediction data mining methods on quality related manufacturing data.

Several criteria are collected from literature to make multidimensional evaluation of the classification and prediction methods. These criteria are given in APPENDIX A. After a comprehensive literature survey, a preliminary list of potential criteria is

constructed. That list consists of both qualitative and quantitative criteria. Criteria which are method specific are eliminated in the first place. Then, these criteria are labeled according to their usage with prediction or classification methods. Most of the qualitative criteria are used with both classification and prediction methods. Two separate criteria lists are formed for classification and prediction methods.

Resulting criteria lists are evaluated and refined by the decision makers. Firstly, to refine the criteria lists, decision makers decide the most important aspects, a classification or prediction method is supposed to have. Secondly, criteria which represent same aspects of the classification or prediction methods are also determined. Among those criteria, the most comprehensive ones are selected to measure the related aspect. Decision criteria in these resulting lists and their definitions are listed APPENDIX B.

The number of the decision criteria affects the number of pairwise comparisons in the ANP application. In order to decrease the number of the criteria further first, the qualitative criteria are evaluated by the experts and based on their importance for quality problems and their similarities a shorter list is obtained. Then to decrease the number of quantitative criteria further, correlation and factor analyses of some accuracy data collected from DM method applications on two quality data sets are performed. The data is available at Köksal et al. (2009).

Results of these analyses are given in APPENDIX E. According to these analyses, highly correlated measures are reevaluated and the measures representing each correlated group (factor) is selected for use in this study.

Decision criteria obtained at the end are grouped and clusters are formed according to their common properties. These clusters are referred to as criteria, and elements in these clusters are referred to as sub-criteria hereafter.

Next, the direction of the influence flow is determined as “from the elements being compared to the parent element”. According to determined flow direction, following question type is used:

“Given a parent element and comparing elements 1 and 2 under it, which element has greater influence on the parent element?”

The relation matrices are constructed according to the selected flow direction. Relation matrices of classification and prediction methods are illustrated in APPENDIX F. Discussion sessions are arranged to determine relations between sub-criteria. The decision makers assess the influence of the sub-criteria in the rows of the relation matrix on the each sub-criterion in the column of the relation matrix. Statistical analyses given in APPENDIX E are also taken into consideration during these discussion sessions. Final decisions are reached with the consensus of the decision makers. Links between sub-criteria are formed according to relation matrices.

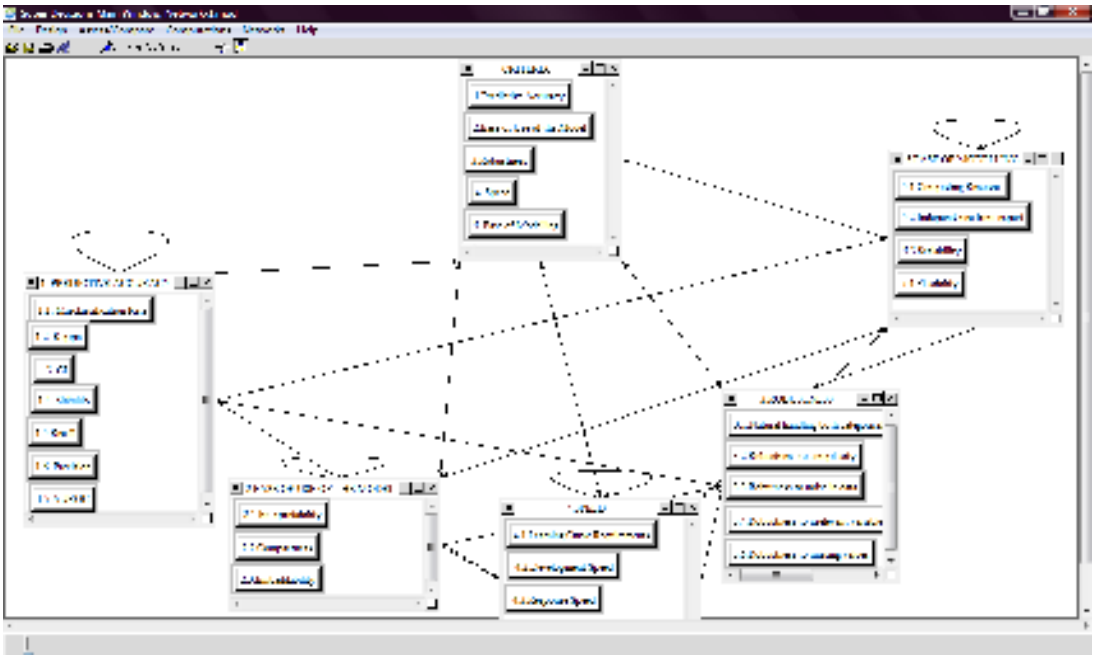


**Figure 3. 1 Initial network structure constructed according to Saaty (1999)**

Initial network is structured according to Saaty (1999). This preliminary network is represented in Figure 3. 1. In literature survey, several ANP applications and different network structures are examined. However, none of these networks suits well to represent this problem. The best way to describe the problem is to form a mixed type network which is both hierarchical and having dependencies and feedbacks. This mixed type network structure is also applied successfully to the performance evaluation of the Research and Development projects being executed in a Defense Research and Development Institute by Tohumcu and Karasakal (2008).

Second Network, which is constructed according to Tohumcu and Karasakal (2008), is illustrated in Figure 3. 2.

Necessity of mixed type network arises due to several drawbacks of the network defined by Saaty (1999). These drawbacks can be generalized as follows:

1. **Meaningless pairwise comparisons:**

   When elements in different clusters are linked to each other, the clusters which they belong to are automatically linked. These links cause pairwise cluster comparisons which are meaningless and cannot be answered by the decision makers.

2. **Insufficient pairwise comparisons:**

   With this network structure, comparing elements from different clusters under a parent element, which is influenced from these two elements, is impossible.

3. **Missing information:**

   The most important drawback of the standard ANP network is that clusters cannot be compared according to their contribution to goal. These comparisons may give invaluable information and should not be ignored.

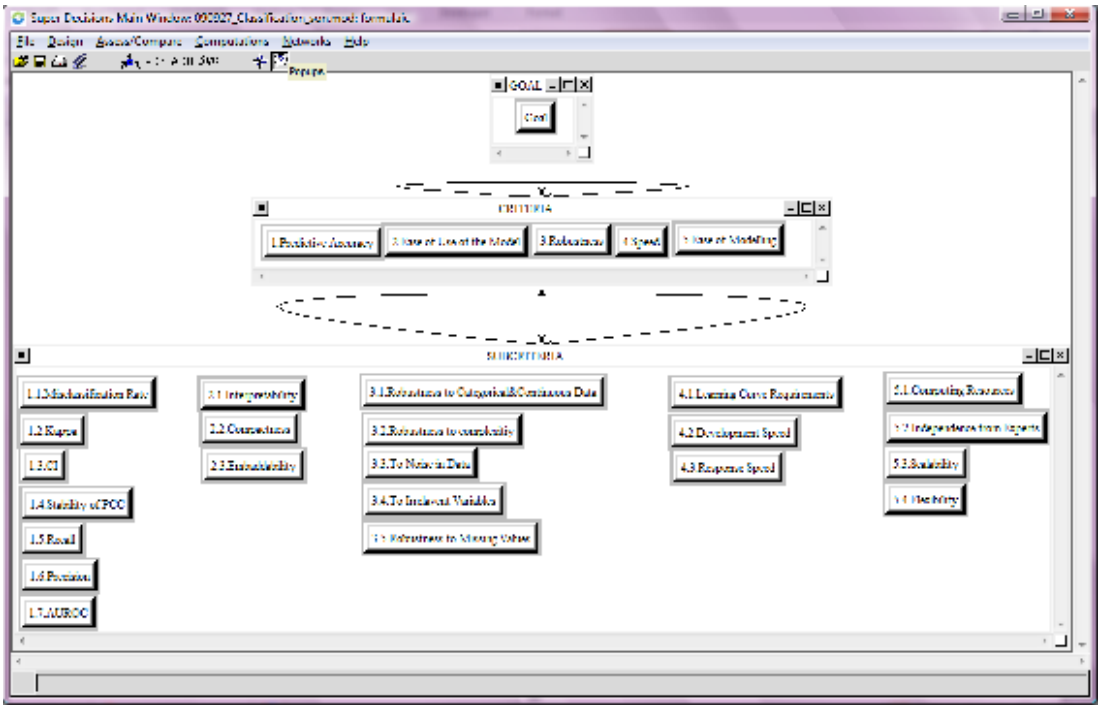


**Figure 3. 2 Network for the classification methods**

Network constructed for the prediction methods is illustrated Figure 3.8 in Section 3.3.2.

Then questions of the pairwise comparisons are generated according to the network in Figure 3.2. A drawback of the second network arises at this point. Pairwise comparison number increases considerably since while preliminary network ignores the pairwise comparisons of the sub-criteria from different criteria, the second one allows comparison of these sub-criteria.

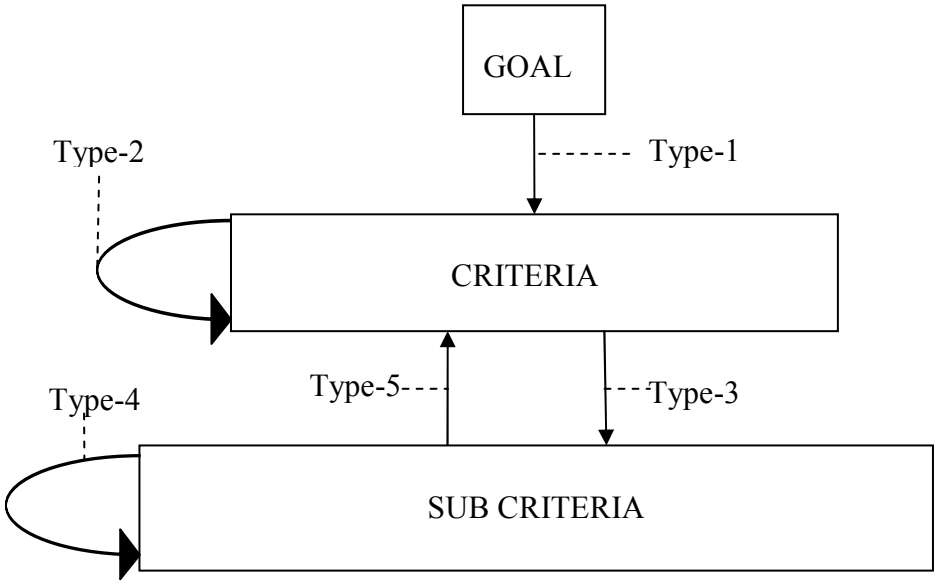



**Figure 3. 3 Network representation and resulting question types**

Prepared questionnaires given in APPENDIX G consist five types of questions:

1. **Pairwise comparison of criteria with respect to the goal:**

   Decision makers compare the criteria with respect the specified goal as stated before which is "Determining the overall performance of the classification/prediction methods and suggesting the most favorable ones."

   Decision makers answer the questions such as "Which criterion should be emphasized more for the evaluation of classification and prediction methods' performance? Predictive Accuracy or Ease of Use of the Model? And how much more? "

2. **Pairwise comparison of criteria with respect to other criteria:**

Decision makers compare the criteria with respect to a criterion that is influenced by them.

In this part of the questionnaire decision makers answer the questions such as "Which criterion influences criterion Predictive Accuracy more? Robustness or Speed? And how much more?"

3. **Pairwise comparison of sub-criteria with respect to criteria:**

Decision makers compare the sub-criteria with respect to criteria. Contribution of the sub-criteria to the criterion to which they belong is evaluated.

Decision makers answer the questions such as "Which sub-criterion should be emphasized more for criterion Predictive Accuracy? Misclassification Rate or Kappa? And how much more?"

4. **Pairwise comparison of sub-criteria with respect to sub-criteria:**

Decision makers compare the sub-criteria with respect to a sub-criterion which is influenced by them.

Decision makers answer the questions such as "Which sub-criterion influences sub-criterion Misclassification Rate more? Kappa or CI? And how much more?"

5. **Pairwise comparison of criteria with respect to sub-criteria (feedback):**

In the last part of the questionnaire decision makers answer the questions such as "Which criterion is influenced more from the sub-criteria Misclassification Rate? Predictive Accuracy or Speed? And how much more?"

It is obvious that style of asking question differs due to the hierarchical structure. If parent element is in upper level, questions are asked to find out the contribution of the compared elements to the parent element (in part 1 and 3). If all of the elements

are in same hierarchical level, questions are asked to find out level of the interdependencies (in part 2 and 4).

In part 5 of the questionnaire, feedbacks are evaluated by the decision makers. Feedbacks are defined as follows: If a sub-criterion say "subcriterion-1" has influence on a criterion and the same sub-criterion is influenced by another sub-criterion say "sub-criterion-2", one can conclude that "sub-criterion-2" indirectly has influence on that criterion which is influenced by "sub-criterion-1".

Comparisons are performed according to Saaty's nine-point scale in Table 2.3. During the evaluation, intermediate values of this scale are not considered in order to simplify to discriminate adjacent values. Intermediate values are used when compromise is needed between the adjacent values.

Consistency check for each pairwise comparison matrix is done and 0.10 limit is used as threshold value (Saaty, 2000). Adjustments are done to improve the consistency when C.R. is above 0.10.

**Table 3. 1 Numbers of pairwise comparisons for classification and prediction methods**

| Question type | Number of Questions | |
|---|---|---|
| | **Classification methods** | **Prediction methods** |
| Type-1 | 10 | 10 |
| Type-2 | 12 | 12 |
| Type-3 | 43 | 25 |
| Type-4 | 1432 | 551 |
| Type-5 | 74 | 70 |

There are totally 1571 pairwise comparisons for classification, 668 for prediction tools; their distribution is illustrated in Table 3. 1.

Total number of questions of pairwise comparisons is very high. To minimize the decision maker effort, discussion sessions are arranged. These four-hour sessions are organized three times with the participation of the decision makers. During these sessions, selected questions are discussed and answered by the decision makers (List of decision makers can be seen in APPENDIX D). About 100 questions can be answered during these sessions, since discussions take long time due to different views of the decision makers having different backgrounds and experiences. Group decision making approach suggested by literature (Saaty, 2000, 2001) is used during pairwise comparisons. Each decision maker expresses his/her own evaluations and their reasoning. Then one combines these evaluations into a group evaluation and questions are answered accordingly.

According to Saaty (2001) number of judgments made by decision makers and their validity are two constant concerns to users of the ANP. Saaty (2001) suggests several methods to expedite decision making without loss of validity. Most of these methods are based on simplifying the constructed network. Since, we have already simplified the network by decreasing the number of decision criteria as much as possible, the most applicable method to expedite decision making is using the completed comparisons to fill the missing judgments. Decision makers' judgments are assumed to be consistent and remaining questions are answered by the author, accordingly. After answering the remaining questions of pairwise comparisons consistently, results are confirmed by the decision makers. Resulting weights are reviewed and unexpected results and corresponding pairwise comparisons are reevaluated by the decision makers.

The local priorities are calculated from each pairwise comparison matrix by using the eigenvector method. Eigenvectors are calculated and placed into the Unweighted Supermatrix and then normalized to get the Weighted Supermatrix. Then, the Weighted Supermatrix is raised to limiting powers to obtain the steady state priorities of the elements in this matrix. The resulting matrix called as Limit Matrix. From the

Limit Matrix the final priorities are extracted with respect to the goal. Resulting Unweighted Supermatrix, Weighted Supermatrix and Limit Matrix for both classification and prediction methods are illustrated in Appendix H.

The resulting weights of the criteria and sub-criteria for classification and prediction methods are illustrated respectively in Table 3. 7 in Section 3.2.2 and Table 3.15 in Section 3.3.2. In the next part of this study, only sub-criteria weights are used as input to the PROMETHEE.

As a first step of the PROMETHEE, data matrices are constructed for both classification and prediction methods.

PROMETHEE needs real valued criteria. Classification methods are evaluated according to the 22 sub-criteria and prediction methods are evaluated according to the 18 sub-criteria. These sub-criteria and objectives of these functions are illustrated in Table 3.8 in Section 3.2.3 and Table 3.16 in Section 3.3.3 respectively.

First seven sub-criteria in Table 3.8 and first three sub-criteria in Table 3.16 are already real valued, but remaining 15 sub-criteria do not have general real valued representations. They are qualitative measures and commonly evaluated verbally such as “high”, “low”, “medium” etc. To convert them to real valued sub-criteria, their common verbal evaluations are matched with a scale. This scale is represented in Table 3. 2.

**Table 3. 2 Scale used to convert verbal evaluation of the subjective measures**

| High | High-Medium | Medium | Medium-Low | Low |
|---|---|---|---|---|
| 5 | 4 | 3 | 2 | 1 |

Decision makers apply different classification and prediction methods on the same quality related data. Then, their performance on these data is evaluated according to predefined sub-criteria in Table 3.6 and Table 3.14.

Quantitative sub-criteria scores of the applied methods are calculated and placed in to the data matrices. Qualitative sub-criteria scores of the applied methods are determined by the decision makers and with the help of literature (Dhar and Stein, 1997; Patel, 2003). Each decision maker evaluates the method, which he/she is experienced in. Scale in Table 3. 2 is used during the evaluation. There can be more than one evaluation for a specific method. Scores for these methods are reevaluated, and then final scores are determined which are confirmed by all decision makers who are experienced in that method. Consensus of the decision makers is needed to reach a final decision.

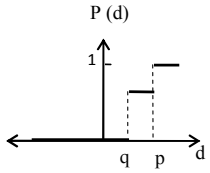
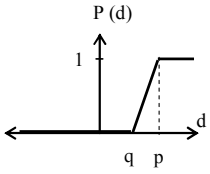
The next step is that determination of the preference function and related parameters. For each sub-criterion, decision makers choose a preference function (preference function types are illustrated in APPENDIX C.

For quantitative sub-criteria, continuous functions are more suitable. Since, small differences between two alternative methods are negligible up to a point and increase in the difference also increases the preference intensity, decision makers agree on the preference function “Criterion with Linear Preference and Indifference Area”.

Scores for qualitative sub-criteria are discrete; they can only get values in Table 3. 2 and there is no intermediate value. Thus, discrete preference functions are more suitable for qualitative sub-criteria. Decision makers agree on the preference function “Level Criterion”, since difference between two alternative methods also can take discrete values.

Chosen preference functions are illustrated in Table 3. 3.

**Table 3. 3 Selected preference functions**

| Type | Graph | Parameter | Function |
|---|---|---|---|
| Level Criterion |   | q, p | $P(d)=\begin{cases}0, & q \geq d \\ 1/2, & q < d \leq p \\ 1, & p < d\end{cases}$ |
| Criterion with Linear Preference and Indifference area |   | q, p | $P(d)=\begin{cases}0, & q \geq d \\ (d-q)/(p-q), & q < d \leq p \\ 1, & p < d\end{cases}$ |

After choosing the preference functions, the next step is determination of the corresponding parameters. Decision makers state their preference by determining thresholds. They answer two questions for each sub-criterion:

Let us compare two methods such as $A_1$ and $A_2$ with respect to sub criterion k. For that sub criterion, alternatives $A_1$ and $A_2$ have scores which are calculated with function $f_k(A_i)$. Difference of those scores is $\boldsymbol{d} = f_k(A_1) - f_k(A_2)$.

1. What is the smallest ***d*** value at which, your preference function, P(***d***), equals to 1? In another words, what is the minimum ***d*** value at which, you prefer $A_1$ to $A_2$ without hesitation.
2. What is the highest ***d*** value at which your preference function, P(***d***), equals to 0 ? In another words, what is the maximum ***d*** value at which you are indifferent between $A_1$ and $A_2$ without hesitation.

Decision makers' answer to the first question gives the ***p*** value of the preference function. And, their answer to the second question gives the ***q*** value of the preference function.

Literature is also used to determine these thresholds. As an example; for kappa, determined thresholds are as follows: $\boldsymbol{q}$ = 0.1 and $\boldsymbol{p}$= 0.2. Landis and Koch (1977) suggest a rough guide to assess kappa. According to this guide, strength of the agreement increase with each 0.2 increase in the kappa score. Thus, 0.2 is used as a threshold to define the smallest difference between kappa scores of the alternative methods to prefer one method over another without hesitation.

Moreover, for quantitative criteria, results of the RANOVA (Repeated-Measures Analysis of Variance) and the Fisher's least significant difference (LSD) test are also evaluated. Actually, these tests are conducted to compare the alternative classification/prediction methods' accuracy statistically. The test results are illustrated in APPENDIX I.

In order to decide whether there are statistically significant differences among the alternative classification/prediction methods, repeated analysis of variance (RANOVA) is performed by using the statistical software SPSS, and the following hypotheses are tested for each comparison criteria:

$H_0$: $\mu_{DT}= \mu_{NN}= \mu_{MARS}= \mu_{LR}= \mu_{SVM}= \mu_{MTS}= \mu_{FCF}$

$H_{1:}$ At least one of them is different

Fisher's LSD compares the alternative methods according to their mean scores of the selected comparison criteria. Before Fisher's LSD test, RANOVA is conducted to see whether there is enough evidence to reject the null hypothesis for each criterion. If the null hypothesis of RANOVA is rejected at $\alpha = 0.05$ significance level, Fisher's (LSD) test is performed at $\alpha = 0.05$ to identify the statistically different alternative methods.

The test results illustrated in APPENDIX I, support the selection of the $\boldsymbol{p}$ and $\boldsymbol{q}$ values as 0.1 and 0.05 respectively. Average mean differences are also around 0.1 for compared methods whose mean criteria scores are significantly different from each other. Average, median and minimum of the significant mean differences are calculated and compared to the selected thresholds. Test results do not much

contradict the selected thresholds. For instance, for PCC (1-MCR) average, median and minimum of the significant mean differences are 0.184, 0.58 and 0.08 respectively. Moreover, threshold selection for kappa is also supported with these test results. For kappa, average, median and minimum of the significant mean differences are 0.28, 0.29 and 0.136 respectively. These results support the selection of the ***p*** and ***q*** values as 0.2 and 0.1 respectively for kappa.

For qualitative sub criteria ***p*** and ***q*** values are determined as 2 and 1 respectively. In literature qualitative sub criteria are evaluated verbally such as high, high-medium, medium, medium-low and low. According to the scale used to convert these verbal evaluations to real valued scores, with every 2-point increase in the score, performance of the method raises one step to an upper level.

Final ***p*** and ***q*** values are reached with the consensus of the decision makers. The resulting preference functions and corresponding parameters are illustrated in Table 3.4.

The next step is calculating the decision makers' preference intensity and then multi-criteria preference index is calculated for each pair of sub-criteria.

The multi-criteria preference index, $\prod$, is defined as follows:

$$\Pi(A_1, A_2) = \frac{\sum_{j=1}^{k} w_j P_j(A_1, A_2)}{\sum_{j=1}^{k} w_j} \qquad (3.1.1)$$

$\Pi(A_1, A_2)$ represents the decision makers' preference intensity of alternative $A_1$ over $A_2$ by considering all sub criteria at the same time.

**Table 3. 4 Selected preference functions and related parameters for classification and prediction methods**

| Preference Function | Parameters | Related Sub-criteria |
|---|---|---|
| Criterion with Linear Preference and Indifference area | ***q***=0.05<br>***p***=0.10 | **For classification methods**<br>1.1.Misclassification Rate<br>1.3.CI<br>1.4.Stability of PCC<br>1.5.Recall<br>1.6.Precision<br>1.7.AUROC<br>**For prediction methods**<br>1.1. RMSE<br>1.2. Stability of RMSE<br>1.3. R Square |
| | ***q***=0.10<br>***p***=0.20 | **For classification methods**<br>1.2.Kappa |
| Level Criterion | ***q***=1<br>***p***=2 | **For classification and prediction methods**<br>2.1.Interpretability<br>2.2.Compactness<br>2.3.Embaddability<br>3.1.Robustness to Categorical &Continuous Data<br>3.2.Robustness to complexity<br>3.3.Robustness to Noise in Data<br>3.4.Robustness to Irrelevant Variables<br>3.5.Robustness to Missing Values<br>4.1.Learning Curve Requirements<br>4.2.Development Speed<br>4.3.Response Speed<br>5.1.Computing Resources<br>5.2.Independence from Experts<br>5.3.Scalability<br>5.4.Flexibility |

For each alternative, leaving and entering flows are calculated. In Figure 3.4, leaving flow of the Alternative 1 of the classification methods (DT) is illustrated.

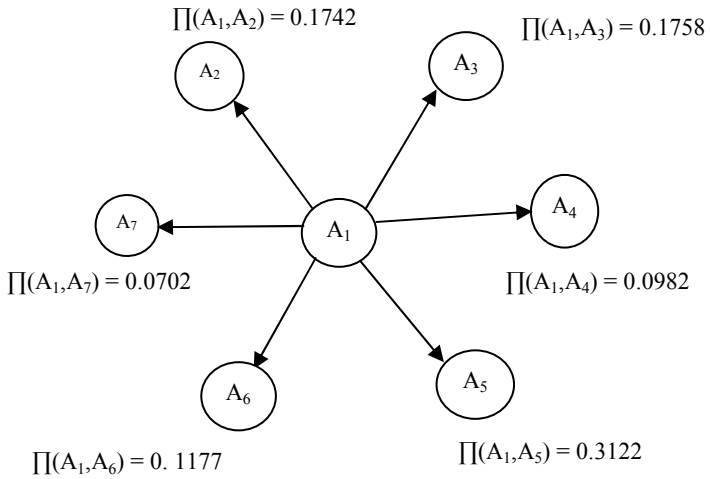



**Figure 3. 4 Leaving flow of Alternative 1 (DT) of the classification methods**

Leaving flow represented in Figure 3.4 measures the outranking character of the Alternative 1 (DT).

In Figure 3. 5, entering flow of the Alternative 1 of the classification methods is illustrated. Entering flow measures the outranked character of the Alternative 1 (DT).

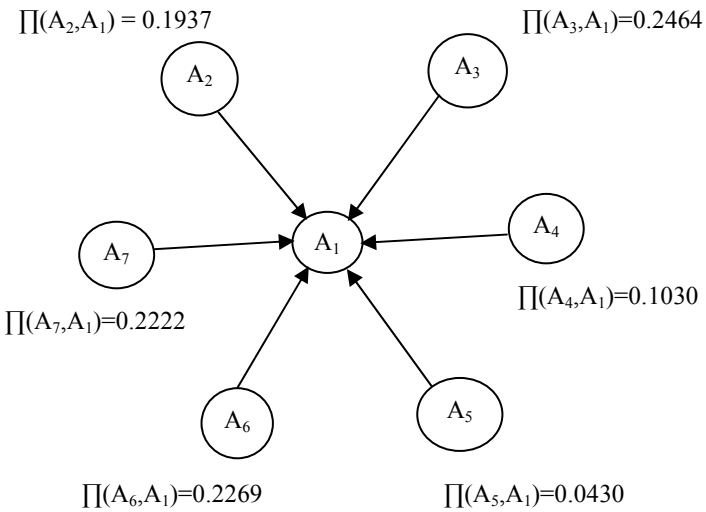



**Figure 3. 5 Entering flow representation of Alternative 1 of the classification methods**

For each alternative, leaving and entering flows are calculated with equations 2.5.2 and 2.5.3. Then, firstly partial preorders are determined with the PROMETHEE I. The higher the leaving flow and the lower the entering flow, the better that alternative. Since partial preorders do not provide much information for this study, PROMETHEE II is also used to prioritize the alternative classification methods.

Classification and prediction methods are prioritized in the section 3.2. and 3.3. respectively.

## 3.2. Ranking of the Classification Methods

### 3.2.1. Selection of the ranking criteria and DM methods

For the classification methods, reevaluated quantitative criteria are MCR (or PCC), recall, precision, kappa, F measure ($F_{0.5}$, $F_1$, and $F_2$), specificity and AUROC. Other quantitative criteria, CI and stability of PCC are evaluated independent from these criteria since they measure different properties of the classification methods. For instance, stability measures the difference between the model performance on test

and train data. Thus, stability and CI are included into the final criteria list without additional statistical analysis. Stability is also included to the statistical analysis to see whether it represents another quantitative criterion.

Moreover, in literature using F measure is suggested instead of using both precision and recall (Billsus and Pazzani, 1998). Likewise, Kappa can be thought as a different measure since it measures the proportion of correctly classified units after the probability of chance agreement has been removed. Nevertheless, precision, recall and kappa are included to the statistical analyses. Additionally, widely used F measures $F_{0.5}$, $F_1$ and $F_2$ are also included to these analyses to select the appropriate one.

Two data sets are used during the statistical analyses of MCR, recall, precision, kappa, $F_{0.5}$, $F_1$, $F_2$, specificity, stability and AUROC. According to correlation analysis in Figure E.1 and Correlation matrix in Figure E. 2 in APPENDIX E, almost all of these criteria are correlated.

Recall is highly correlated with $F_2$ and kappa is highly correlated with $F_1$ and $F_{0.5}$. These are the most remarkable correlations extracted from the correlation analysis. As expected, F Measures ($F_{0.5}$, $F_1$, $F_2$,) are also highly correlated with each other. Decision of using F measure instead of both recall and precision is not supported by this correlation analysis since precision is not highly correlated with any of the F Measures.

Besides correlation analysis, factor analysis is also conducted to support criteria selection. Factor analysis is a statistical method used to describe covariance relationships among many variables in terms of a few unobservable variables called factors. The observed variables are modeled as linear combinations of the factors, plus error terms. The information gained on the interdependencies can be used later to reduce the set of variables in a data set (Johnson and Wichern, 2002). Results and findings are given in Figure E. 6 in APPENDIX E.

According to factor analysis in Figure E. 6, recall, $F_1$ and $F_2$ form a group; MCR, specificity and stability of PCC form another group; and precision forms a group by

itself. Other criteria kappa, AUC and $F_{0.5}$ can be incorporated into one of these groups. Kappa is close to the "recall, $F_1$ and $F_2$" group. $F_{0.5}$ is very close to the "precision" group. AUC can be incorporated into "recall, $F_1$ and $F_2$" or "MCR, specificity and stability of PCC" group, since factor loadings of AUC are very close to each other (0.608 and 0.621) for corresponding factors of these groups. Indeed, these loadings are very low to include AUC in one of these groups. Although kappa and AUC can be incorporated into a group and correlated, they are included in the resulting criteria list since as stated before kappa measures different properties of the models, AUC is also suggested by Bradley (1997). According to Bradley (1997) AUC is one of the best ways to evaluate a classifier's performance since it seems to be the only one that is independent of the decision threshold and not biased by prior probabilities. MCR is also included in the final criteria list, since it is one of the widely used and well known criteria.

To sum up, recall is chosen to represent the first group; MCR is chosen to represent the second group. Precision is also selected since it behaves independent from the other criteria. F measure is not included in the final criteria list since its components recall and precision are already selected. Kappa, AUC, CI and stability of PCC are included in the final criteria list since they measure different properties of the classification methods as stated before.

Classification methods which are applied on the data are illustrated in Table 3. 5. Findings of these applications are used to fill in the data matrices of classification methods.

**Table 3. 5 Alternative methods of classification and prediction**

| Alternative | Classification methods |
|---|---|
| A1 | DT |
| A2 | NN |
| A3 | MARS |
| A4 | LR |
| A5 | MTS |
| A6 | FC |
| A7 | SVM |

### 3.2.2. Determination of criteria weights using ANP

Selected decision criteria for classification methods are grouped and clusters are formed according to their common properties. The clusters and their elements are illustrated in Table 3. 6. As stated before, these clusters are entitled as criteria, and elements in these clusters are entitled as sub-criteria hereafter.

According to relation matrix in APPENDIX F, links between sub-criteria are formed and accordingly criteria are also linked. The resulting network is illustrated in Figure 3. 2. in the Section 3.1. Then questions of the pairwise comparisons are generated according to the network in Figure 3.2. The pairwise comparisons are completed with the help of decision makers and the literature. Eigenvector method is used to calculate the local priorities from the pairwise comparison matrices. These calculations are done with "The Super Decisions" software implementing the Analytic Network Process developed by Thomas Saaty. Calculated eigenvectors are used to form the Unweighted Supermatrix.

**Table 3. 6 Clusters and their elements for classification methods**

| |
|---|
| **1.Predictive Accuracy** |
| 1.1.Misclassification Rate |
| 1.2.Kappa |
| 1.3.CI |
| 1.4.Stability of PCC |
| 1.5.Recall |
| 1.6.Precision |
| 1.7.AUROC |
| **2.Ease of Use of the Model** |
| 2.1.Interpretability |
| 2.2.Compactness |
| 2.3.Embaddability |
| **3.Robustness** |
| 3.1.Robustness to Categorical &Continuous Data |
| 3.2.Robustness to Complexity |
| 3.3.Robustness to Noise in Data |
| 3.4.Robustness to Irrelevant Variables |
| 3.5.Robustness to Missing Values |
| **4.Speed** |
| 4.1.Learning Curve Requirements |
| 4.2.Development Speed |
| 4.3.Response Speed |
| **5.Ease of Modeling** |
| 5.1.Computing Resources |
| 5.2.Independence from Experts |
| 5.3.Scalability |
| 5.4.Flexibility |

Unweighted Supermatrix is normalized to get Weighted Supermatrix which is column stochastic (each of its columns sums to unity). This is required to ensure convergence of the matrix when the Weighted Supermatrix is raised to limiting powers. Unweighted Supermatrix, Weighted Supermatrix and Limit Matrix for classification methods are illustrated in APPENDIX H. From the Limit Matrix, final priorities, which are steady state priorities of the criteria and sub-criteria with respect to the goal, are extracted. The resulting weights of the criteria and sub-criteria are illustrated in Table 3.7. In the next section, only sub-criteria weights are used as input.

**Table 3. 7 Criteria and sub-criteria weights for classification methods**

| | | |
|---|---|---|
| Criteria | 1.Predictive Accuracy | 0.11321 |
| | 2.Ease of Use of the Model | 0.06252 |
| | 3.Robustness | 0.36007 |
| | 4.Speed | 0.21539 |
| | 5.Ease of Modeling | 0.24882 |
| Sub-criteria | 1.1.Misclassification Rate | 0.02419 |
| | 1.2.Kappa | 0.01617 |
| | 1.3.CI | 0.00700 |
| | 1.4.Stability | 0.00546 |
| | 1.5.Recall | 0.03368 |
| | 1.6.Precision | 0.02093 |
| | 1.7.AUROC | 0.01283 |
| | 2.1.Interpretability | 0.02375 |
| | 2.2.Compactness | 0.01162 |
| | 2.3.Embaddability | 0.00788 |
| | 3.1.Robustness to Categorical &Continuous Data | 0.02702 |
| | 3.2.Robustness to Complexity | 0.22656 |
| | 3.3.Robustness to Noise in Data | 0.10012 |
| | 3.4.Robustness to Irrelevant Variables | 0.04470 |
| | 3.5.Robustness to Missing Values | 0.09105 |
| | 4.1.Learning Curve Requirements | 0.02235 |
| | 4.2.Development Speed | 0.05368 |
| | 4.3.Response Speed | 0.07631 |
| | 5.1.Computing Resources | 0.02886 |
| | 5.2.Independence from Experts | 0.04128 |
| | 5.3.Scalability | 0.06153 |
| | 5.4.Flexibility | 0.06303 |

### 3.2.3. Ranking of the classification methods using PROMETHEE

In PROMETHEE, to prioritize the alternative classification methods in Table 3.5 sub-criteria and determined weights in Table 3.7 are used as input. These sub-criteria and their objectives are illustrated in Table 3.8.

**Table 3. 8 Sub-criteria and objectives for classification methods**

| Sub-criteria (Classification) | Objective |
|---|---|
| 1.1.Misclassification Rate | Minimize |
| 1.2.Kappa | Maximize |
| 1.3.CI | Minimize |
| 1.4.Stability | Minimize |
| 1.5.Recall | Maximize |
| 1.6.Precision | Maximize |
| 1.7.AUROC | Maximize |
| 2.1.Interpretability | Maximize |
| 2.2.Compactness | Maximize |
| 2.3.Embaddability | Maximize |
| 3.1.Robustness to Categorical &Continuous Data | Maximize |
| 3.2.Robustness to complexity | Maximize |
| 3.3.Robustness to Noise in Data | Maximize |
| 3.4.Robustness to Irrelevant Variables | Maximize |
| 3.5.Robustness to Missing Values | Maximize |
| 4.1.Learning Curve Requirements | Maximize |
| 4.2.Development Speed | Maximize |
| 4.3.Response Speed | Maximize |
| 5.1.Computing Resources | Maximize |
| 5.2.Independence from Experts | Maximize |
| 5.3.Scalability | Maximize |
| 5.4.Flexibility | Maximize |

**Table 3. 9 Preference Index ∏ Table for Classification Methods**

| Preference Indices π | | | | | | |
|---|---|---|---|---|---|---|
| A1 | A2 | A3 | A4 | A5 | A6 | A7 |
| DT | NN | MARS | LR | SVM | MTS | FCF |

| | | | | | |
|---|---|---|---|---|---|
| A1-A2 | A1-A3 | A1-A4 | A1-A5 | A1-A6 | A1-A7 |
| 0.174 | 0.114 | 0.101 | 0.312 | 0.118 | 0.070 |
| A2-A1 | A3-A1 | A4-A1 | A5-A1 | A6-A1 | A7-A1 |
| 0.194 | 0.246 | 0.103 | 0.043 | 0.227 | 0.222 |
| | A2-A3 | A2-A4 | A2-A5 | A2-A6 | A2-A7 |
| | 0.058 | 0.201 | 0.430 | 0.248 | 0.247 |
| | A3-A2 | A4-A2 | A5-A2 | A6-A2 | A7-A2 |
| | 0.118 | 0.207 | 0.154 | 0.163 | 0.236 |
| | | A3-A4 | A3-A5 | A3-A6 | A3-A7 |
| | | 0.196 | 0.444 | 0.293 | 0.288 |
| | | A4-A3 | A5-A3 | A6-A3 | A7-A3 |
| | | 0.106 | 0.038 | 0.109 | 0.181 |
| | | | A4-A5 | A4-A6 | A4-A7 |
| | | | 0.329 | 0.162 | 0.107 |
| | | | A5-A4 | A6-A4 | A7-A4 |
| | | | 0.050 | 0.267 | 0.246 |
| | | | | A5-A6 | A5-A7 |
| | | | | 0.066 | 0.047 |
| | | | | A6-A5 | A7-A5 |
| | | | | 0.109 | 0.181 |
| | | | | | A6-A7 |
| | | | | | 0.053 |
| | | | | | A7-A6 |
| | | | | | 0.163 |

As a first step, data matrix is constructed for classification methods. Then, preference functions and their parameters are determined. Selected preference functions are illustrated in Table 3.4 in Section 3.1. Corresponding thresholds ***p*** and ***q*** are determined with the consensus of the decision makers and also with the help of the literature and statistical analyses. For all quantitative criteria, except Kappa, thresholds are ***q*** = 0.05 and ***p***= 0.1. For kappa, thresholds are determined as ***q*** = 0.1

and ***p***= 0.2. Lastly, for qualitative criteria thresholds are determined as ***q*** = 1 and ***p***= 2. (In section 3.1, threshold selection is explained in detail.)

For each pair of alternative methods, preference index ∏ is calculated and they are used to calculate entering and leaving flows of the alternative methods. Preference indices are illustrated in Table 3.9.

From the calculated leaving and entering flows of the alternative methods partial preorders are determined. Leaving and entering flows are illustrated in Table 3.10.

**Table 3. 10 Leaving and entering flows of classification methods**

| Classification | | φ+ | φ- |
|---|---|---|---|
| A1 | DT | 0.889669 | 1.034809 |
| A2 | NN | 1.377826 | 1.051224 |
| A3 | MARS | 1.585933 | 0.605642 |
| A4 | LR | 1.014313 | 1.061436 |
| A5 | MTS | 0.397704 | 0.605642 |
| A6 | FC | 1.191624 | 1.049469 |
| A7 | SVM | 1.466055 | 0.812653 |

Firstly, partial preorder is determined with the PROMETHEE I. Values represented in Table 3. 10 are used while determining the partial preorder. The partial preorder is represented in Figure 3.6.

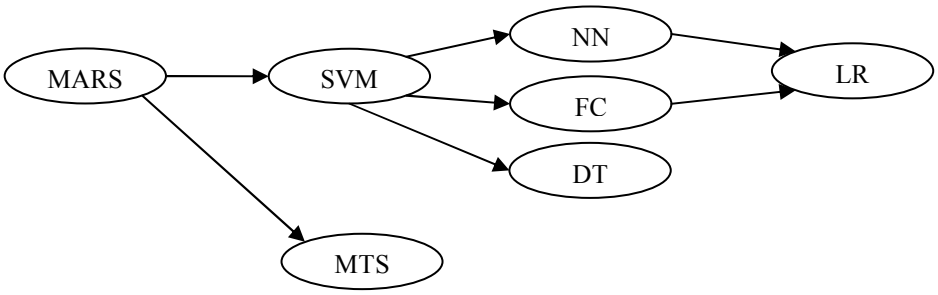



**Figure 3. 6 The partial preorder induced by Table 3.10**

From the partial preorder illustrated in Figure 3.6 following conclusions are obtained:

$A_3$ (MARS) is preferred to $A_1$ (DT) since $\Phi^+(A_3) > \Phi^+(A_1)$ and $\Phi^-(A_3) < \Phi^-(A_1)$

$A_3$ (MARS) is preferred to $A_2$ (NN) since $\Phi^+(A_3) > \Phi^+(A_2)$ and $\Phi^-(A_3) < \Phi^-(A_2)$

$A_3$ (MARS) is preferred to $A_4$ (LR) since $\Phi^+(A_3) > \Phi^+(A_4)$ and $\Phi^-(A_3) < \Phi^-(A_4)$

$A_3$ (MARS) is preferred to $A_5$ (MTS) since $\Phi^+(A_3) > \Phi^+(A_5)$ and $\Phi^-(A_3) = \Phi^-(A_5)$

$A_3$ (MARS) is preferred to $A_6$ (FC) since $\Phi^+(A_3) > \Phi^+(A_6)$ and $\Phi^-(A_3) < \Phi^-(A_6)$

$A_3$ (MARS) is preferred to $A_7$ (SVM) since $\Phi^+(A_3) > \Phi^+(A_7)$ and $\Phi^-(A_3) < \Phi^-(A_7)$

$A_2$ (NN) is preferred to $A_4$ (LR) since $\Phi^+(A_2) > \Phi^+(A_4)$ and $\Phi^-(A_2) < \Phi^-(A_4)$

$A_6$ (FC) is preferred to $A_4$ (LR) since $\Phi^+(A_6) > \Phi^+(A_4)$ and $\Phi^-(A_6) < \Phi^-(A_4)$

$A_7$ (SVM) is preferred to $A_1$ (DT) since $\Phi^+(A_7) > \Phi^+(A_1)$ and $\Phi^-(A_7) < \Phi^-(A_1)$

$A_7$ (SVM) is preferred to $A_2$ (NN) since $\Phi^+(A_7) > \Phi^+(A_2)$ and $\Phi^-(A_7) < \Phi^-(A_2)$

$A_7$ (SVM) is preferred to $A_4$ (LR) since $\Phi^+(A_7) > \Phi^+(A_4)$ and $\Phi^-(A_7) < \Phi^-(A_4)$

$A_7$ (SVM) is preferred to $A_6$ (FC) since $\Phi^+(A_7) > \Phi^+(A_6)$ and $\Phi^-(A_7) < \Phi^-(A_6)$

Remaining combinations of the alternatives are incomparable such as:

$A_1$ (DT) is incomparable to $A_2$ (NN) since $\Phi^+(A_2) > \Phi^+(A_1)$ and $\Phi^-(A_2) > \Phi^-(A_1)$

$A_1$ (DT) is incomparable $A_4$ (LR) to since $\Phi^+(A_4) > \Phi^+(A_1)$ and $\Phi^-(A_4) > \Phi^-(A_1)$

$A_1$ (DT) is incomparable to $A_5$ (MTS) since $\Phi^+(A_1) > \Phi^+(A_5)$ and $\Phi^-(A_1) > \Phi^-(A_5)$

$A_1$ (DT) is incomparable to $A_6$ (FC) since $\Phi^+(A_6) > \Phi^+(A_1)$ and $\Phi^-(A_6) > \Phi^-(A_1)$

$A_2$ (NN) is incomparable to $A_5$ (MTS) since $\Phi^+(A_2) > \Phi^+(A_5)$ and $\Phi^-(A_2) > \Phi^-(A_5)$

$A_2$ (NN) is incomparable to $A_6$ (FC) since $\Phi^+(A_2) > \Phi^+(A_6)$ and $\Phi^-(A_2) > \Phi^-(A_6)$

$A_4$(LR) is incomparable to $A_5$ (MTS) since $\Phi^+(A_4) > \Phi^+(A_5)$ and $\Phi^-(A_4) > \Phi^-(A_5)$

$A_5$ (MTS) is incomparable to $A_6$ (FC) since $\Phi^+(A_6) > \Phi^+(A_5)$ and $\Phi^-(A_6) > \Phi^-(A_5)$

$A_5$ (MTS) is incomparable to $A_7$ (SVM) since $\Phi^+(A_7) > \Phi^+(A_5)$ and $\Phi^-(A_7) > \Phi^-(A_5)$

Partial preorders do not provide much information; one can only conclude that MARS is superior to other six methods. PROMETHEE II is used to prioritize the alternative classification methods.

Complete preorders are determined with PROMETHEE II. Net flow values of the alternative classification methods are listed in Table 3. 11:

**Table 3. 11 Net Flows of the Alternative Classification Methods**

| Classification Methods | | φ |
|---|---|---|
| A1 | DT | -0.14514 |
| A2 | NN | 0.326602 |
| A3 | MARS | 0.980291 |
| A4 | LR | -0.04712 |
| A5 | MTS | -0.20794 |
| A6 | FC | 0.142155 |
| A7 | SVM | 0.653402 |

According to net flows of the alternative classification methods, resulting priorities are given in Figure 3.7:

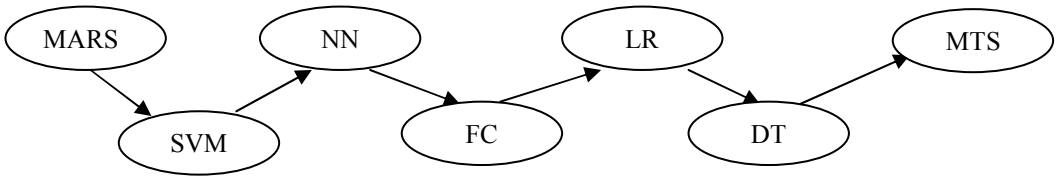



**Figure 3. 7 The complete preorder induced by Table 3.11**

### 3.2.4. Sensitivity analysis and discussions

Sensitivity analyses are conducted to see the effects of the criteria weights and thresholds on the resulting net flows, namely priorities.

In the first place, sensitivity analysis for the changing criteria weights is conducted. There are 22 sub-criteria for classification methods and they are dependent to each

other, since their weights have to sum up 1. Increasing one of the sub-criteria weights causes to decrease in remaining ones. Thus, we focused on the sub-criteria weights one by one. For instance, we changed the weight of the MCR in the range of [0, 1] and difference in the weight of the MCR is allocated to the other sub-criteria proportional to their own weights. Used formula during the weight generation and resulting graphs of the changing net flows of the alternative methods for each sub-criterion are illustrated in the APPENDIX J.

Graphs illustrated in APPENDIX J show that the net flows and also priorities of the alternative methods are not very sensitive to change in the sub-criteria weights. The main reason providing insensitivity to the sub-criteria weights is that there are 22 sub-criteria and their weights have to sum up 1. A small change in a sub-criterion weight has much smaller effect on other sub-criteria after allocating this change to remaining 21 sub-criteria. If the number of sub-criteria is lower, the effect of the change in weight will be more significant. Moreover, graphs show that after a point, priorities of the alternative methods remain same; this is because while the analyzed sub-criterion weight is increasing, other sub-criteria weights are decreasing accordingly and they become quite small. Then alternative methods are prioritized according to scores of the analyzed sub-criterion having very high weight relative to other sub-criteria.

Besides the number of the sub-criteria, sub-criteria scores of the alternative methods also affect the sensitivity of the net flows. If for a sub-criterion, alternative methods have close scores, net flows of these methods are not sensitive to change in that sub-criterion. For instance, priorities of the alternative methods do not change with the changing weight of the sub-criterion Response Speed. Sensitivity analysis graph of the Response Speed is illustrated in Figure J.18 in APPENDIX J. For Response Speed only one classification method, MTS has different score than other methods; and with the increase in the weight of the response speed other sub-criteria weights are losing their significance. Figure J.18 shows that when the response speed weight reaches to 1, other subcriteria weights will be equal to zero and all of the methods except MTS, have equal net flow values since there is only one sub-criterion,

response speed, having positive weight and all of these methods have same score for that sub-criterion.

Moreover, sub-criteria weights are relatively small numbers. According to the graphs in APPENDIX J, one can conclude that priorities of the alternative methods are very insensitive to sub-criteria weights. The reasoning behind this conclusion is that indeed to change the priorities of the alternative methods, percentage increase or decrease in sub-criteria weights should be very high. These percentage changes are given in Table 3.12. Additionally, these sub-criteria weights are generated from the ANP application and small changes in the pairwise comparisons hardly affect the resulting weights since number of the pairwise comparisons is very high and interdependencies and feedbacks are taken into consideration while the final priorities are calculated by the ANP. Moreover, it is important to note that the percentage changes in the Table 3.12 are calculated according to the observed first change in the priority of the alternative methods. The graphs in APPENDIX J show that these alternative methods are already lower ranked methods and these alterations do not change our suggested classification methods.

In addition to sub-criteria weights, thresholds are also analyzed to see their effects on the resulting priorities. We only focused on the thresholds of the quantitative sub-criteria, since qualitative sub-criteria can get only discrete scores which are illustrated in Table 3.2 and conducting a sensitivity analyze for these sub-criteria is unnecessary. Contrary to sub-criteria weights, thresholds are independent from each other and thus there are numerous threshold combinations can be used during the PROMETHEE application. Handling these combinations that is very hard and it requires remarkable effort. Thus, during the sensitivity analysis, we only change one threshold and keep other threshold as they are. The resulting graphs are illustrated in APPENDIX J. Net flows and priorities are very insensitive to the changes in thresholds. One of the reasons is again number of the sub-criteria since overall effects of the changes in the preferences are multiplied with the sub-criteria weights to obtain the preference intensity of the alternative methods. The differences become more insignificant after multiplication since sub-criteria weights are very small numbers and overall effects of the changes become quite insignificant. Another

reason of this insensitivity is used preference function which is “Criterion with Linear Preference and Indifference Area”. The selected thresholds’ effects on the preference intensity are decreased by this preference function since values below the preference threshold is also evaluated with a linear function having a slope determined by the determined preference threshold. This slope lessens the change in the the preference function result.

**Table 3. 12 Percentage increases and decreases in weights to change the priorities of the alternative classification methods**

| Sub-criteria (Classification) | Increase in weight | Decrease in weight |
|---|---|---|
| 1.1.Misclassification Rate | 148% | * |
| 1.2.Kappa | 147% | * |
| 1.3.CI | * | * |
| 1.4.Stability | 3563% | * |
| 1.5.Recall | 48% | * |
| 1.6.Precision | 139% | * |
| 1.7.AUROC | 212% | * |
| 2.1.Interpretability | 195% | * |
| 2.2.Compactness | 158% | * |
| 2.3.Embaddability | 661% | * |
| 3.1.Robustness to Categorical &Continuous Data | 307% | * |
| 3.2.Robustness to complexity | 6% | 29% |
| 3.3.Robustness to Noise in Data | 40% | 70% |
| 3.4.Robustness to Irrelevant Variables | 34% | 33% |
| 3.5.Robustness to Missing Values | 43% | 34% |
| 4.1.Learning Curve Requirements | 124% | * |
| 4.2.Development Speed | 86% | 81% |
| 4.3.Response Speed | * | * |
| 5.1.Computing Resources | 73% | * |
| 5.2.Independence from Experts | 166% | * |
| 5.3.Scalability | 30% | 19% |
| 5.4.Flexibility | * | * |

* There is not such a value that may change the ranking.

Resulting priorities of the alternative methods are also compared with the findings of the other studies in literature. These studies and their suggestions are illustrated in Table 2.2. in Chapter 2. These findings do not contradict with our findings. Of course only some of the alternative methods used in this study, are presented in the literature. In some of these studies MARS is applied to the different problem contexts and these available comparison studies in literature also presents MARS as a superior method to other methods as well. Moreover, in general NN is also superior to DT according to these comparison studies and these findings support our resulting priorities as well.

## 3.3. Ranking of the Prediction Methods

### 3.3.1. Selection of the ranking criteria and DM methods

For the prediction methods, reevaluated quantitative criteria are MSE, MAE, RMSE, R, R2, Adjusted R2, PWI1, PWI2, Stability of MSE and Stability of RMSE. Correlation analysis and factor analysis are given in APPENDIX E. According to correlation analysis and correlation matrix in APPENDIX E, as expected MSE and RMSE; R and R Square; PWI1 and PWI2; Stability of MSE and Stability of RMSE are highly correlated.

According to factor analysis in Figure E.8, there are four different groups. MSE, RMSE and Adjusted R square form a group. R and R Square form another group. Stability of MSE and Stability of RMSE form the next group. Finally, PWI1 and PWI2 form the last group. MAE is very close the first group "MSE, RMSE, Adjusted R square" and it is incorporated into this group.

In literature, using Adjusted R Square is suggested instead of using R or R square. Unlike R square, adjusted R square allows for the degrees of freedom associated with the sums of the squares. Therefore, even though the residual sum of squares decreases or remains the same as new explanatory variables are added, the residual variance does not. Thus, adjusted R square is generally considered to be a more accurate goodness-of-fit measure than R square (Montgomery and Runger, 1996). Although literature supports the usage of Adjusted R square, statistical analyses

show that Adjusted R Square highly correlated with MSE and RMSE. Finally, RMSE is chosen to represent the first group (MSE, RMSE, Adjusted R square, MAE) by the decision makers.

R Square is chosen to represent the second group and Stability of RMSE is chosen to represent the third group. PWI1 and PWI2 are not used in this study; "1-PWI2" measuring the outliers that fall outside the range of $2\sigma$ is already represented by the predetermined qualitative sub-criteria "Robustness to noise in data". Final decision criteria lists are formed accordingly. Prediction methods which are applied on the data are illustrated in Table 3.13. Findings of these applications are used to fill in the data matrices of prediction methods.

**Table 3. 13 Alternative methods of prediction**

| Alternative | Prediction methods |
|---|---|
| A1 | DT |
| A2 | NN |
| A3 | MARS |
| A4 | MLR |
| A5 | Fuzzy Regression |
| A6 | Robust Regression |

### 3.3.2. Determination of criteria weights using ANP

Same as the classification methods, selected decision criteria for prediction methods are grouped and clusters are formed according to their common properties. The clusters and their elements are illustrated in Table 3.14. These clusters are identical with the clusters of the classification methods. Only elements under the cluster "Predictive Accuracy" are different. As stated before, these clusters are entitled as criteria, and elements in these clusters are entitled as sub-criteria hereafter.

**Table 3. 14 Clusters and their elements for prediction methods**

| **1.Predictive Accuracy** |
|---|
| 1.1.RMSE |
| 1.2. Stability of RMSE |
| 1.3. R Square |
| **2.Ease of Use of the Model** |
| 2.1.Interpretability |
| 2.2.Compactness |
| 2.3.Embaddability |
| **3.Robustness** |
| 3.1.Robustness to Categorical &Continuous Data |
| 3.2.Robustness to Complexity |
| 3.3.Robustness to Noise in Data |
| 3.4.Robustness to Irrelevant Variables |
| 3.5.Robustness to Missing Values |
| **4.Speed** |
| 4.1.Learning Curve Requirements |
| 4.2.Development Speed |
| 4.3.Response Speed |
| **5.Ease of Modeling** |
| 5.1.Computing Resources |
| 5.2.Independence from Experts |
| 5.3.Scalability |
| 5.4.Flexibility |

According to relation matrix in APPENDIX F, links between sub-criteria are formed and accordingly criteria are also linked. Then questions of the pairwise comparisons are generated according to the resulting network illustrated in Figure 3.7. The pairwise comparisons are completed with the help of decision makers and the literature. Moreover, comparisons which are performed for classification methods are also utilized. Eigenvector method is used to calculate the local priorities from the pairwise comparison matrices. The Unweighted Supermatrix is formed with the eigenvectors generated from the pairwise comparison matrices. Unweighted Supermatrix is normalized and Weighted Supermatrix is generated and then Weighted Supermatrix is raised to limiting powers to get Limit Matrix. From the Limit Matrix, final priorities, which are steady state priorities of the criteria and sub-criteria with respect to the goal, are extracted. Unweighted Supermatrix, Weighted Supermatrix and Limit Matrix for prediction methods are illustrated in APPENDIX H. The resulting weights of the criteria and sub-criteria are illustrated in Table 3.15. In the next section only sub-criteria weights are used as input.

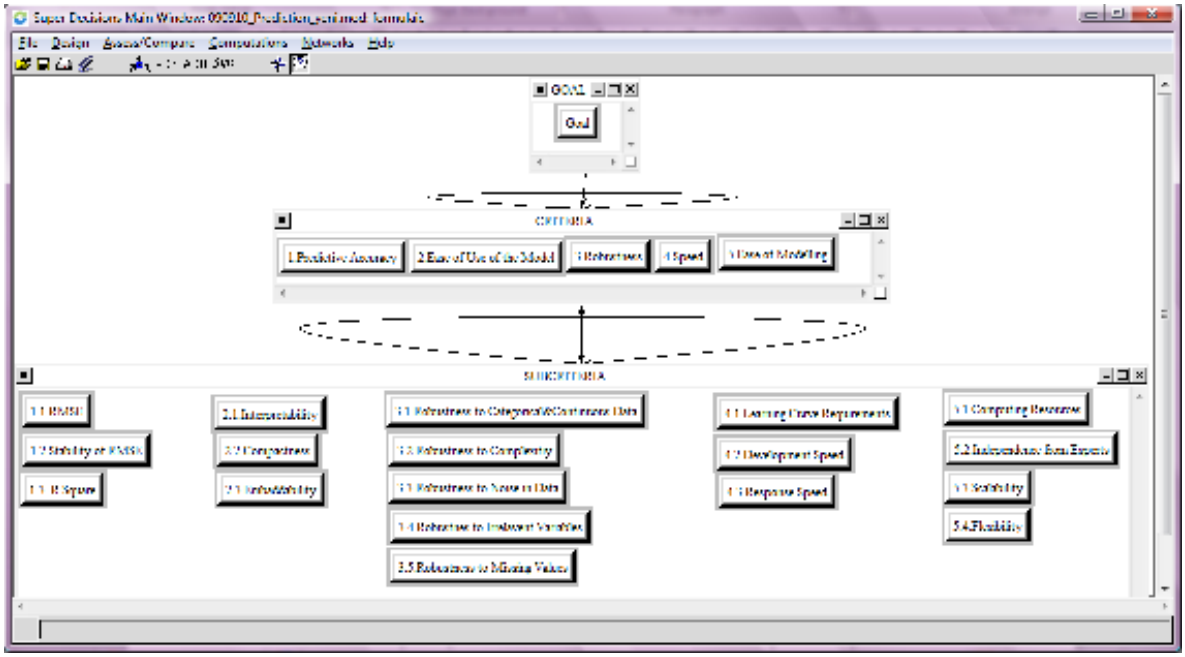



**Figure 3. 8 Network for the prediction methods**

**Table 3. 15 Criteria and subcriteria weights for prediction methods**

| | | |
|---|---|---|
| Criteria | 1.Predictive Accuracy | 0.09673 |
| | 2.Ease of Use of the Model | 0.06504 |
| | 3.Robustness | 0.36389 |
| | 4.Speed | 0.21960 |
| | 5.Ease of Modeling | 0.25473 |
| Sub-criteria | 1.1.RMSE | 0.03496 |
| | 1.2.Stability of RMSE | 0.00880 |
| | 1.3.R Square | 0.03115 |
| | 2.1.Interpretability | 0.02510 |
| | 2.2.Compactness | 0.01228 |
| | 2.3.Embaddability | 0.00829 |
| | 3.1.Robustness to Categorical &Continuous Data | 0.02828 |
| | 3.2.Robustness to Complexity | 0.23597 |
| | 3.3.Robustness to Noise in Data | 0.10346 |
| | 3.4.Robustness to Irrelevant Variables | 0.04637 |
| | 3.5.Robustness to Missing Values | 0.09523 |
| | 4.1.Learning Curve Requirements | 0.02363 |
| | 4.2.Development Speed | 0.05717 |
| | 4.3.Response Speed | 0.07865 |
| | 5.1.Computing Resources | 0.03161 |
| | 5.2.Independence from Experts | 0.04593 |
| | 5.3.Scalability | 0.06456 |
| | 5.4.Flexibility | 0.06857 |

### 3.3.3. Ranking of the classification methods using PROMETHEE

In PROMETHEE, to prioritize the alternative prediction methods in Table 3.13 sub-criteria and determined weights in Table 3.15 are used as input. These sub-criteria and their objectives are illustrated in Table 3.16.

**Table 3. 16 Sub-criteria functions and objectives for prediction methods**

| Sub-criteria (Prediction) | Objective |
|---|---|
| 1.1.RMSE (Root Mean Square Error) | Minimize |
| 1.2.Stability of RMSE | Minimize |
| 1.3. R Square | Maximize |
| 2.1.Interpretability | Maximize |
| 2.2.Compactness | Maximize |
| 2.3.Embaddability | Maximize |
| 3.1.Robustness to Categorical &Continuous Data | Maximize |
| 3.2.Robustness to complexity | Maximize |
| 3.3.Robustness to Noise in Data | Maximize |
| 3.4.Robustness to Irrelevant Variables | Maximize |
| 3.5.Robustness to Missing Values | Maximize |
| 4.1.Learning Curve Requirements | Maximize |
| 4.2.Development Speed | Maximize |
| 4.3.Response Speed | Maximize |
| 5.1.Computing Resources | Maximize |
| 5.2.Independence from Experts | Maximize |
| 5.3.Scalability | Maximize |
| 5.4.Flexibility | Maximize |

As a first step, data matrix is constructed for prediction methods. Then, preference functions and their parameters are determined. Selected preference functions are illustrated in Table 3.4 in Section 3.1. Corresponding thresholds $p$ and $q$ are determined with the consensus of the decision makers and also with the help of the literature and statistical analyses. For all quantitative criteria, thresholds are determined as $q$ = 0.05 and $p$= 0.1 and for qualitative criteria thresholds are determined as $q$ = 1 and $p$= 2.

For each pair of alternative methods, preference index $\prod$ is calculated and they are used to calculate entering and leaving flows of the alternative prediction methods. Preference indices are illustrated in Table 3.17.

From the calculated leaving and entering flows of the alternative methods partial preorders are determined. Leaving and entering flows are illustrated in Table 3.18.

**Table 3. 17 Preference Index ∏ Table for Prediction Methods**

| Preference Indices π | | | | | |
|---|---|---|---|---|---|
| A1 | A2 | A3 | A4 | A5 | A6 |
| DT | NN | MARS | MLR | FR | RR |
| A1-A2 | A1-A3 | A1-A4 | A1-A5 | A1-A6 | |
| 0.149 | 0.048 | 0.084 | 0.137 | 0.112 | |
| A2-A1 | A3-A1 | A4-A1 | A5-A1 | A6-A1 | |
| 0.219 | 0.267 | 0.161 | 0.179 | 0.199 | |
| | A2-A3 | A2-A4 | A2-A5 | A2-A6 | |
| | 0.036 | 0.197 | 0.290 | 0.151 | |
| | A3-A2 | A4-A2 | A5-A2 | A6-A2 | |
| | 0.124 | 0.215 | 0.100 | 0.119 | |
| | | A3-A4 | A3-A5 | A3-A6 | |
| | | 0.193 | 0.288 | 0.178 | |
| | | A4-A3 | A5-A3 | A6-A3 | |
| | | 0.112 | 0.009 | 0.052 | |
| | | | A4-A5 | A4-A6 | |
| | | | 0.211 | 0.156 | |
| | | | A5-A4 | A6-A4 | |
| | | | 0.193 | 0.194 | |
| | | | | A5-A6 | |
| | | | | 0.051 | |
| | | | | A6-A5 | |
| | | | | 0.052 | |

**Table 3. 18 Leaving and entering flows of prediciton methods**

| Prediction Methods | | φ+ | φ- |
|---|---|---|---|
| A1 | DT | 0.529906 | 1.024830 |
| A2 | NN | 0.892457 | 0.706893 |
| A3 | MARS | 1.049250 | 0.256517 |
| A4 | MLR | 0.855377 | 0.860088 |
| A5 | RR | 0.532370 | 0.256517 |
| A6 | FR | 0.736704 | 0.649254 |

Firstly, partial preorders are determined with the PROMETHEE I. Values represented in Table 3.18 are used while determining the partial preorders. The partial preorder is represented in Figure 3.9.

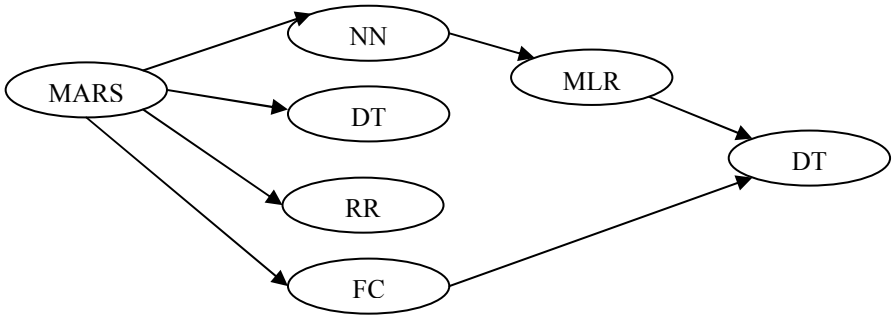



**Figure 3. 9 The complete preorder induced by Table 3.18**

From the partial preorder illustrated in Figure 3.9, following conclusions are obtained:

$A_2$ (NN) is preferred to $A_1$ (DT) since $\Phi^+(A_2) > \Phi^+(A_1)$ and $\Phi^-(A_2) < \Phi^-(A_1)$

$A_2$ (NN) is preferred to $A_4$ (MLR) since $\Phi^+(A_2) > \Phi^+(A_4)$ and $\Phi^-(A_2) < \Phi^-(A_4)$

$A_3$ (MARS) is preferred to $A_1$ (DT) since $\Phi^+(A_3) > \Phi^+(A_1)$ and $\Phi^-(A_3) < \Phi^-(A_1)$

$A_3$ (MARS) is preferred to $A_2$ (NN) since $\Phi^+(A_3) > \Phi^+(A_2)$ and $\Phi^-(A_3) < \Phi^-(A_2)$

$A_3$ (MARS) is preferred to $A_4$ (MLR) since $\Phi^+(A_3) > \Phi^+(A_4)$ and $\Phi^-(A_3) < \Phi^-(A_4)$

$A_3$ (MARS) is preferred to $A_5$ (RR) since $\Phi^+(A_3) > \Phi^+(A_5)$ and $\Phi^-(A_3) = \Phi^-(A_5)$

$A_3$ (MARS) is preferred to $A_6$ (FR) since $\Phi^+(A_3) > \Phi^+(A_6)$ and $\Phi^-(A_3) < \Phi^-(A_6)$

$A_4$ (MLR) is preferred to $A_1$ (DT) since $\Phi^+(A_4) > \Phi^+(A_1)$ and $\Phi^-(A_4) < \Phi^-(A_1)$

$A_5$ (RR) is preferred to $A_1$ (DT) since $\Phi^+(A_5) > \Phi^+(A_1)$ and $\Phi^-(A_5) < \Phi^-(A_1)$

$A_6$ (FR) is preferred to $A_1$ (DT) since $\Phi^+(A_6) > \Phi^+(A_1)$ and $\Phi^-(A_6) < \Phi^-(A_1)$

Remaining combinations of the alternatives are incomparable such as:

$A_2$ (NN) is incomparable to $A_5$ (RR) since $\Phi^+(A_2) > \Phi^+(A_5)$ and $\Phi^-(A_2) > \Phi^-(A_5)$

$A_2$ (NN) is incomparable to $A_5$ (FR) since $\Phi^+(A_2) > \Phi^+(A_6)$ and $\Phi^-(A_2) > \Phi^-(A_6)$

$A_4$ (MLR) is incomparable to $A_5$ (RR) since $\Phi^+(A_4) > \Phi^+(A_5)$ and $\Phi^-(A_4) > \Phi^-(A_5)$

$A_4$ (MLR) is incomparable to $A_6$ (FR) since $\Phi^+(A_4) > \Phi^+(A_6)$ and $\Phi^-(A_4) > \Phi^-(A_6)$

$A_5$ (RR) is incomparable to $A_6$ (FR) since $\Phi^+(A_6) > \Phi^+(A_5)$ and $\Phi^-(A_6) > \Phi^-(A_5)$

According to partial preorders, two conclusions can be reached; MARS is superior to the other five methods and DT is the worse method among the six alternative methods. PROMETHEE II is used to prioritize the all of the alternative prediction methods.

Complete preorders are determined with PROMETHEE II. Net flow values of the alternative prediction methods are listed in Table 3.19.

**Table 3. 19 Net Flows of the Alternative Prediction Methods**

| Prediction Methods | | φ |
|---|---|---|
| A1 | DT | -0.494924 |
| A2 | NN | 0.185564 |
| A3 | MARS | 0.792732 |
| A4 | MLR | -0.004711 |
| A5 | RR | 0.275852 |
| A6 | FR | 0.087450 |

According to net flows of the alternative prediction methods, resulting priorities are given in Figure 3.10:

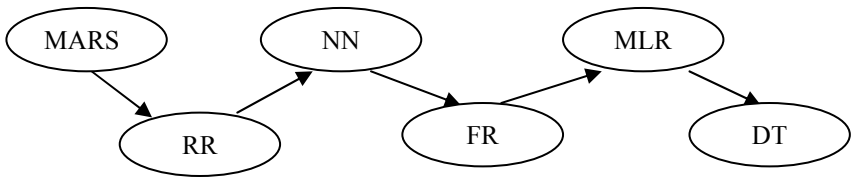



**Figure 3. 10 The complete preorder induced by Table 3.19**

### 3.3.4. Sensitivity analysis and discussions

Sensitivity analyses are conducted for prediction methods as well. Sensitivity analysis for the changing criteria weights is conducted for 18 sub-criteria of the predicition methods. Resulting graphs are illustrated in APPENDIX J.

According to graphs illustrated in APPENDIX J, the net flows and priorities of the alternative predicition methods are not sensitive to change in the sub-criteria weights. The reasons of this insensitivity are same as the reasons stated for classification methods. For prediction methods, there are 18 sub-criteria and this is also a high number and decreases the sensitivity of the net flows to the sub-criteria weights. Percentage changes in criteria weights to change the priorities of the alternative prediction methods are given in Table 3.20.

For some of the sub-criteria, such as flexibility, alternative methods have close even identical scores and thus net flows and priorities of the alternative methods are insensitive to the change in weight of this sub-criterion. The sensitivity graph of the Flexibility is illustrated in Figure J. 40. When the weight of the flexibility is equal to 1, other subcriteria weights will be equal to zero and all of the methods have the same net flow values since all of them have the same score for that sub-criterion.

For prediction methods, thresholds of the quantitative sub-criteria are also analyzed to see their effects on the resulting priorities. According to the resulting graphs illustrated in APPENDIX J net flows of the prediction methods are also very insensitive to the changes in thresholds. One of the reasons is again number of the sub-criteria that is 18 for predicition methods. As stated before, overall effects of the changes in the preferences are multiplied with the sub-criteria weights and increasing number of the sub-criteria lowers these weights. For prediction methods used preference function is same as the classification methods' which is "Criterion with Linear Preference and Indifference Area". As stated before in Section 3.2.4 effects of the selected thresholds on the preference intensity are decreased by this preference function.

**Table 3. 20 Percentage increases and decreases in weights to change the priorities of the alternative prediction methods**

| Sub-criteria (Prediction) | Increase in weight | Decrease in weight |
|---|---|---|
| 1.1.RMSE | 987% | * |
| 1.2.Stability of RMSE | 127% | |
| 1.3.R Square | 61% | |
| 2.1.Interpretability | 99% | 60% |
| 2.2.Compactness | * | * |
| 2.3.Embaddability | 383% | * |
| 3.1.Robustness to Categorical &Continuous Data | * | * |
| 3.2.Robustness to Complexity | 53% | 11% |
| 3.3.Robustness to Noise in Data | 55% | 42% |
| 3.4.Robustness to Irrelevant Variables | 51% | 35% |
| 3.5.Robustness to Missing Values | 26% | 37% |
| 4.1.Learning Curve Requirements | 69% | * |
| 4.2.Development Speed | 40% | 48% |
| 4.3.Response Speed | * | * |
| 5.1.Computing Resources | 58% | * |
| 5.2.Independence from Experts | 118% | 35% |
| 5.3.Scalability | 210% | * |
| 5.4.Flexibility | * | * |

* There is not such a value that may change the ranking.

For prediction methods, MARS is superior to other methods as well and this finding does not contradict with the studies in literature summarized in Table 2.2.

# CHAPTER 4

# CONCLUSIONS AND FUTURE WORK

This study investigates which classification and prediction method should be preferred for specific QI and control problems. The aim of this study is to comprehensively evaluate performance of the selected classification and prediction methods applied on the selected QI applications.

This is a discrete multiple criteria decision problem since decision space consists of finite set of alternatives and the criterion set is explicitly known. ANP and PROMETHEE are selected to apply to this problem and used to prioritize the alternative classification and prediction methods. Experts with prior experience and background with the application of these methods on relevant data have contributed to this study since both of these methods require decision makers' or experts' input in comparing and evaluating the methods.

An important point, the practitioners should be careful about is that during the ANP application, they should be consistent in determining the directions of the influences between the elements of the network. Otherwise resulting comparisons and their interpretations do not represent the interdependencies and feedbacks of the problem. Of course final priorities do not represent the real priorities of the criteria either. This kind of mistakes can be easily noticed by the careful practitioners since the resulting pairwise comparisons will be meaningless and probably unexpected.

The weights extracted from the ANP, are used in the PROMETHEE and alternative classification and prediction methods are prioritized according to these weights and determined preference functions and related thresholds. According to these priorities

MARS is superior to the other classification methods. Among the other classification methods the second best method is SVM. For prediction methods, MARS is the suggested method among the others, as well. Besides, its performance is by far the best. The second best prediction method is RR and its performance is very close to following method NN. It is important to note that these priorities are not valid for all problem contexts. The resulting priorities are determined according to selected decision criteria, their relative importance with respect to goal of the problem and performance of the alternative methods.

Used data structure also changes with the problem context. In this study, quality data is used and the most significant characteristics of the quality related data are imbalanced classes (such as defective, non defective), curse of dimensionality (small sized of data, large number of variables) and mixed type of data. "Imbalanced classes" requires careful structuring of the decision criteria. For instance, especially for criteria, recall and precision stating the class of interest properly is very important. For QI problems, the class representing the defectives is determined as class of interest since cost of the misclassifying the defectives is very high. At the beginning of the study, class of interest should be stated clearly.

In this study, several parameters are determined by the decision makers/experts. Thus sensitivity analyses are needed to see the affect of these parameters such that criteria weights and thresholds. The analyses show that resulting priorities are not very sensitive to the change in these parameters. There are many reasons ensuring the insensitivity to the criteria weights and thresholds. The most important one is the number of the assessed criteria, since increase in the number of the criteria decreases their effects on the resulting priorities. Number of the decision criteria affects the pairwise comparison number in ANP application. The decision criteria number and the relations between components of the network determine the resulting number of the pairwise comparisons. Increase in the decision criteria number also increases the number of the pairwise comparisons. High pairwise comparison number decreases comparisons' individual effects on the resulting weights. Small changes in the pairwise comparisons have ignorable effects on the weights. To change the priorities

of the methods by changing the criteria weights, we need to change great amount of comparisons.

Sensitivity analysis to examine the effects of the thresholds is also conducted. Only thresholds of the quantitative criteria are analyzed since qualitative criteria can get discrete scores which are 1, 2, 3, 4 and 5 and difference between two alternative method can get only following scores: 0, 1, 2, 3, 4. Thus, using a threshold different than these scores is meaningless. However, selecting a threshold for quantitative criteria is very hard because there is unlimited alternative threshold choice. Since these thresholds are independent from each other, there are numerous possible threshold combinations. Thus we try to analyze their effect one by one. We change one threshold while others remain unchanged. Conducting a comprehensive sensitivity analysis consisting these combinations and possible alternative conditions needs great effort and can be evaluated as a future study.

Insensitivity to the thresholds can be explained with the preference function used during the PROMETHEE application. For quantitative criteria, we select the preference function "Criterion with Linear Preference and Indifference Area". This function uses a linear function when the difference between scores of the two compared alternative methods is below the preference threshold. Changing this threshold affects the slope of this linear function and effects of this change are lessened. In literature, PROMETHEE is also suggested since it is more stable to the threshold deviations (Brans et al. 1986).

Although this method was applied for prioritization of the classification and prediction methods applied to QI problems it is a generalized method that can be adapted or extended for any problem context. The criteria/sub-criteria determined in this study is specific to QI problem context, factors for evaluating method performance and their priorities will vary in each different context. The approach can be implemented in any discrete problem by making the necessary changes in the criteria/sub-criteria and the pairwise comparison judgments and thresholds of the PROMETHEE. And as a last word, for the problems having high number of decision criteria which are also interdependent, ANP produces considerable number of

pairwise comparisons. For this kind of problems, ANP should be improved otherwise application becomes really difficult for both experts and practitioners.

ANP evaluates the interdependencies, feedbacks and hierarchies, and then reaches the absolute priority of any criterion regardless of which criteria it influences. Resulting priorities are determined according to influences between criteria. For instance, a criterion influencing the most of the remaining criteria has higher priority than a criterion mostly influenced by other criteria. Thus, to analyze the relationships and their directions, supporting methods could be used to improve performance of the ANP. For this purpose, DEMATEL method can be used to specify interdependencies and determine direction of the influences. For future studies, DEMATEL may be used to improve the performance of the ANP but it should be noted that this method also requires expert contribution to determine its parameters (Li and Tzeng, 2009).

# REFERENCES


1. Agresti, A. *An Introduction to Categorical Data Analysis*; Wiley Series in Probability and Statistics: USA, 103-246, (1996).
2. Al-Kloub, B., Al-Shemmeri, T., Pearman, A. The role of weights in multi-criteria decision aid, and the ranking of water projects in Jordan. *European Journal of Operational Research.* 99, 278-288, (1997).
3. Asan, U., Ayberk Soyer, A., Identifying strategic management concepts: An analytic network process approach. *Computers & Industrial Engineering Article in Press.* (2008).
4. Avcı, Ezgi. *A Comparison of Robust Regression Methods for Outliers.* MSc. Thesis. Middle East Technical University, Industrial Engineering Department, Ankara, (2009).
5. Ayhan, D. *Multi-Class Classification Methods Utilizing Mahalanobis Taguchi System and a Resampling Approach for Imbalanced Data Sets.* MSc. Thesis. Middle East Technical University, Industrial Engineering Department, Ankara, (2009).
6. Bakır, B. *Defect Cause Modeling with Decision Tree and Regression Analysis: A Case Study in Casting Industry.* MSc. Thesis. Middle East Technical University, Informatics Institute, Ankara, (2007).
7. Belton, V., Stewart, T. *Multiple Criteria Decision Analysis: An Integrated Approach.* Kluwer Academic Publishers: USA, (2002).
8. Billsus, D. and Pazzani, M. Learning Collaborative Filters, *In: Proceedings of ICML'98, 46-53.* Morgan Kaufman Eds. (1998).
9. Bradley A.E. The Use of the Area under the Roc Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 30, 7, 1145-1159, (1997).

10. Bozkurt, A. Multi Criteria Decision Making with Interdependent Criteria Using Prospect Theory. MSc. Thesis. Middle East Technical University, Industrial Engineering Department, Ankara, (2007).

11. Brazdil, P.B. and Soares, C. A Comparison of Ranking Methods for Classification Algorithm Selection. *Lecture Notes in Computer Science*, Volume 1810/2000, 63-75, (2000).

12. Braha, D., Shmilovici, A. Data Mining for Improving a Cleaning Process in the Semiconductor Industry. *IEEE Transactions on Semiconductor Manufacturing*, 15, 1, (February 2002)

13. Brandimarte, P., Zotteri, G. (2007). *Introduction to distribution logistics*. Retrieved 08 10, 2009, from Google Books: http://books.google.com/books?id=G72OWrypw34C

14. Brans, J. P. and Vincke Ph. A Preference Ranking Organisation Method: (The PROMETHEE Method for Multiple Criteria Decision-Making). *Management Science,* 31, 6, 647-656, (June 1985).

15. Brans, J.P., Mareschal, B., Vincke, P. How to Select and How to Rank Projects: The PROMETHEE Method for MCDM. *European Journal of Operational Research,* 24, 228-238, (1986).

16. Brumen, B., Golob, I., Jaakkola, H., Welzer, T. and Rozman, I. Early Assessment of Classification Performance. *Australasian CS Week Frontiers,* 91–96, (2004)

17. Büyüközkan, G., Ertay, T., Kahraman, C., Ruan, D. Determining the Importance Weights for the Design Requirements in the House of Quality Using the Fuzzy Analytical Network Approach. *International Journal of Intelligent Systems,* 19, 443-461, (2004).

18. Cristianini, N., Taylor, J.S. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, (2000).

19. Chang and Ayyub, Fuzzy regression methods - a comparative assessment. *Fuzzy Sets and Systems,* 119, 187-203, (2001).

20. Chang C.W., Wu, C.R., Lin, C.T., Lin, H.L. Evaluating digital video recorder systems using analytic hierarchy and analytic network processes. *Information Sciences,* 177, 3383–3396, (2007).

21. Cheng, E. W. L., Li, H. Analytic hierarchy process an approach to determine measures for business performance. *Measuring Business Excellence*, 5, 30-36, (2001).

22. Chien, C., Li H., Jeang, A. Data mining for improving the solder bumping process in the semiconductor packaging industry. *Intell. Sys. Acc. Fin. Mgmt.* 14, 43–57, (2006).

23. Classification and Regression Trees (C&RT). (2009, 07 26). Retrieved 08 10, 2009, from www.statsoft.com: http://www.statsoft.com/TEXTBOOK/stcart.html

24. Cohen J. A coefficient of agreement for nominal scales. *Educ. Psychol Meas* 20, 1, pp 37–46, (1960).

25. Cohen's kappa. (2009, 07 27). Retrieved 08 10, 2009, from Wikipedia The Free Encyclopedia: http://en.wikipedia.org/wiki/Cohen%27s_kappa

26. Dağdeviren, M. Decision making in equipment selection: an integrated approach with AHP and PROMETHEE. *J Intell Manuf,* 19, 397–406, (2008).

27. Dağdeviren, M., Eraslan, E. Promethee Sıralama Yöntemi ile Tedarikçi Seçimi. *J. Fac. Eng. Arch. Gazi Univ.*, 23, 1, 69-75, (2008).

28. Dağdeviren, M., Yüksel, İ. Using the analytic network process (ANP) in a SWOT analysis – A case study for a textile firm. *Information Sciences,* 177, 3364–3382, (2007).

29. Dağdeviren M., Yüksel, I., Kurt, M. A fuzzy analytic network process (ANP) model to identify faulty behavior risk (FBR) in work system. *Safety Science,* 46, 771–783, (2008).

30. De Leeneer I., Pastijn, H. Selecting land mine detection strategies by means of outranking MCDM techniques. *European Journal of Operational Research,* 139, 327–338, (2002).

31. Dulmin, R., Mininno, V. Supplier selection using a multi-criteria decision aid method. *Journal of Purchasing & Supply Management,* 9, 177–187, (2003).

32. Dhar, V and Stein, R. *Seven Methods for Transforming Corporate Data into Business Inteligence.* Prentice Hall, 1-29, (1997).

33. Ertay, T., Kahraman, C. Evaluation of Desigh requirements using fuzzy Outranking Methods. *International Journal of Intelligent Systems,* 22, 1229-1220, (2007).

34. Fan, C., Guo, R., Chen, A., Hsu, K., Wei, C. Data Mining and Fault Diagnosis based on Wafer Acceptance Test Data and In-line Manufacturing Data. *IEEE*, 171-174, (2001)

35. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Researchers. Tech Report HPL-2004-3, HP Laboratories. , (2004). Retrieved 08 10, 2009, from http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf

36. Fielding, A. H., Bell J. F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24, 1, 38–49, (1997).

37. Figueira, J., Greco, S., Ehrgott, M. *Multiple criteria decision analysis: state of the art surveys.Volume 57*. (2005). Retrieved 08 10, 2009, from Google Books: http://www.google.com/books?lr=&hl=tr&id=YqmvlTiMNqYC&dq

38. Gencer C, Gürpınar, D. Analytic network process in supplier selection: A case study in an electronic firm. *Applied Mathematical Modelling*, 31, 2475–2486, (2007).

39. Grimm, L.G., Yarnold, P.R. *Reading and Understanding Multivariate Statistics*. American Psychological Association, USA, (1994) (Chapter 7 by Raymond E. Wright)

40. Guitouni, A., Martel, J.M. Tentative guidelines to help choosing an appropriate MCDA method. *European Journal of Operational Research,* 109, 501-521, (1998).

41. Halouani, N., Chabchoub, H., Martel, J.M. PROMETHEE-MD-2T method for project selection. *European Journal of Operational Research,* 195, 841–849, (2009).

42. Han, J. and Kamber, M, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, (2001).

43. Han, L., Han, L. and Liu, C. Neural network applied to prediction of the failure stress for a pressurized cylinder containing defects. *International Journal of Pressure Vessels and Piping,* 76, 4, 215–219, (1999).

44. Huang, H., Wu, D. *Product Quality Improvement Analysis Using Data Mining: A Case Study in Ultra-Precision Manufacturing Industry.* L. Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3614, 577 – 580, (2005).

45. Huang, J., Lu, J., Ling, C.X. Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy. *In: Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),* (2003).

46. Hyndman, R. J., Koehler, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting,* 22, 679– 688, (2006).

47. Jharkhariaa, S., Shankarb, R. Selection of logistics service provider: An analytic network process. *(ANP) approach Omega*, 35, 274 – 289, (2007).

48. Johnson, R., Wichern, D. *Applied Multivariate Statistical Anaylsis*. New Jersey: Prentice Hall, (2002).

49. Ip, K.W., Kwong, C.K. and Wong, Y.W. Fuzzy regression approach to modeling transfer moulding for microchip encapsulation. *Journal of Materials Processing Technology,* 140 (1-3), 147-151, (2003).

50. Kang, B., Park, S. Integrated machine learning approaches for complementing statistical process control procedures. *Decision Support Systems,* 29, 59–72, (2000).

51. Kangas, A., Kangas, J. and Pykäläinen, J. Outranking methods as tools in strategic natural resources planning. *Silva Fennica,* 35(2): 215–227, (2001).

52. Khan S., Faisal, M.N. An analytic network process model for municipal solid waste disposal options. *Waste Management,* 28, 1500–1508, (2008).

53. Kim, I., Son, J., Yarlagadda, P.K.D.V. A study on the quality improvement of robotic GMA welding process. *Robotics and Computer Integrated Manufacturing,* 19 (6), 567–572, (2003).

54. Kim, S., Lee, C.M. Nonlinear prediction of manufacturing systems through explicit and implicit data mining. *Computer and Industrial Engineering*, 33 (3-4), 461-464, (1997).

55. Köksal, G., Batmaz, I., Kartal, E. Developing a Classification Model for Customer Satisfaction with a Driver's Seat: A comparative case study. *Proceedings of 6$^{th}$ International Symposium on Intelligent and manufacturing Systems,* Sakarya, 520-530, (2008a).

56. Köksal, G., Batmaz, I., Testik, M.C. Data mining processes and a review of their applications for product and process quality improvement in manufacturing industry. Technical Report No: 08-03, Middle East Technical University, Industrial Engineering Department, Ankara, (2008b).

57. Köksal, G., Batmaz, I., Karasözen, B., Kayalıgil, S., Testik, M.C., Özdemirel, N.E., Weber, G.W., Bakır, B., Öztürk, B. Kalite İyileştirmede Veri Madenciliği Kullanımı ve Geliştirilmesi. Final Report, TÜBİTAK Project No:105M138, Ankara, (2009).

58. Landis, J. R., Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics,* 33(1), 159–174, (March 1977).

59. Lawrence, K.D., Arthur J.L. *Robust Regression: Analysis and Applications Statistics*, Textbooks and Monographs, Vol 108, USA, (1990).

60. Li, M., Feng, S., Sethi, L., Luciow, J., Wagner, K. Mining Production Data with Neural Network & CART. *Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03),* 0-7695-1978-4/03 (2003).

61. Li, C.W., Tzeng, G.H. Identification of a threshold value for the DEMATEL method using the maximum mean de-entropy algorithm to find critical services provided by a semiconductor intellectual property mall. *Expert Systems with Applications*, 36, 9891–9898, (2009).

62. Lian, J., Lai, X.M., Lin, Z.Q., Yao, F.S. Application of data mining and process knowledge discovery in sheet metal assembly dimensional variation diagnosis. *Journal of Materials Processing Technology*, 129, 315-320, (2002).

63. Lim, T.S., Loh, W.Y., Shih, Y.S. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. *Machine Learning,* 40, 203-229, (2000).

64. Macharis, C., Springael, J., De Brucker, K., Verbeke, A. PROMETHEE and AHP: The design of operational synergies in multicriteria analysis. Strengthening PROMETHEE with ideas of AHP. *European Journal of Operational Research*, 153, 307–317, (2004).

65. Manel, S., Dias, J.M,, Ormerod, S.J. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecological Modelling*, 120, 337–347, (1999).

66. Martel, J.M. Multicriterion Decision Aid: Methods and Applications. CORS - SCRO 1999 Annual Conference June 7-9, (1999) WINDSOR, ONTARIO Retrieved 08 10, 2009, from http://www.cors.ca/bulletin/v33n1_1e.pdf

67. Meier, A., Werro, N. A Fuzzy Classification Model for Online Customers. *Informatica*, 31, 175–182, (2007).

68. Moisen, G.G. and Frescino, T.S. Comparing five modeling techniques for predicting forest characteristics. *Ecol. Model.* 157, 209–225, (2002).

69. Montgomery, D., Peck E.A., Vining, G.G. *Introduction to Linear Regression Analysis.* Wiley and Sons: NY, (2006).

70. Mosteller, F., Tukey, J.W. *Data Analysis and Regression.* Addison-Wesley Publishing Company: Canada, (1977).

71. Muata, K., Bryson, O. Evaluation of decision trees: a multi-criteria approach. *Computers & Operations Research*, 31, 1933–1945, (2004).

72. Muñoz J., Felicísimo Á. Comparison of statistical methods commonly used in predictive modeling. *Journal of Vegetation Science,* 15, 285-292, (2004).

73. Neaupane K.M., Piantanakulchai, M. Analytic network process model for landslide hazard zonation. *Engineering Geology*, 85, 281–294, (2006).

74. Niemira, M.P., Saaty T.L., An Analytic Network Process model for financial-crisis forecasting. *International Journal of Forecasting*, 20, 573– 587, (2004).

75. Ortiz, M.C., Sarabia, L.A., Herrero, A. Robust regression techniques: A useful alternative for the detection of outlier data in chemical analysis. *Talanta*, 70, 499–512, (2006).

76. Özer, Gizem. *Fuzzy Classification Models Based on Tanaka's Fuzzy Linear Regression Approach and Nonparametric Improved Fuzzy Classifier Functions.* M.Sc. Thesis, Middle East Technical University, Industrial Engineering Department, Ankara, (2009).

77. Patel, N.R., 15.062 Data Mining, (Spring 2003), Retrieved 08 10, 2009, from http://ocw.mit.edu/NR/rdonlyres/Sloan-School-of-Management/15-062Data-MiningSpring2003/21399C69-6DB8-4F42-9A2B-4D1C1FAE0E78/0/comparison.pdf

78. Pardalos, P. M., Siskos Y., Zopounidis, C. *Advances in multicriteria analysis.* Kluwer Academic Publishers, (1995).

79. Promentilla, M. A. B., Furuichi, T., Ishii, K., Tanikawa, N. A fuzzy analytic network process for multi-criteria evaluation of contaminated site remedial countermeasures. *Journal of Environmental Management,* 88, 479–495, (2008)

80. Razi, M.A., Athappilly, K. A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications,* 29, 65–74, (2005).

81. Rokach, L and Maimon, O. Data mining for improving the quality of manufacturing: a feature set decomposion approach. *J Intell Manuf*, 17, 285-299, (2006).

82. Saaty, T.L. Time dependent decision-making; dynamic priorities in the AHP/ANP: Generalizing from points to functions and from real to complex variables. *Mathematical and Computer Modelling,* 46, 860–891, (2007).

83. Saaty, T.L. Decision-making with the AHP: Why is the principal eigenvector necessary. *European Journal of Operational Research*, 145, 85–91, (2003).

84. Saaty, T. L. *Decision Making with Dependence and Feedback: The Analytic Network Process,* RWS Publications, (2001).

85. Saaty, T. *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process,* RWS Publications, Pittsburgh, PA, (2000).

86. Saaty, T. L. Fundamentals of the Analytic Network Process, *ISAHP 1999*, Kobe, Japan, 12-14, (1999).

87. Saaty, T. L. *Fundamentals of Decision Making and Priority Theory with AHP.* Vol.VI, RWS Publications, 224-291, (2000).

88. Saaty, T.L. That Is Not theAnalytic Hierarchy Process: What theAHP Is and What It Is Not. *J. Multi-Criteria Decision Analysis,* 6, 320-339 (1997).

89. Shang, J.S., Lin, Y.E., Goetz, A.M. Diagnosis of MRSA with Neural Networks and Logistic Regression Approach. *Health Care Management Science,* 3, 287-297, (2000).

90. Shi, X., Schillings, P., Boyd, D. Applying artificial neural networks and virtual experimental design to quality improvement of two industrial processes. *Int. J. Prod. Res.,* 42, 1, 101–118, (2004).

91. Manel, S., Dias, J.M., Ormerod, S.J. Comparing Discriminant Analysis, Neural Networks And Logistic Regression For Predicting Species Distributions: A Case Study With A Himalayan River Bird. *ELSEVIER Ecological Modelling,* 120, 337–347, (1999).

92. Stolzer, A.J., Halford, C. Data Mining Methods Applied to Flight Operations Quality Assurance Data: A Comparison to Standard Statistical Methods. *Journal of Air transportation,* 12, 1, (2007).

93. Sun, H., Xu, G., Tian, P. Evaluation of the Design Alternatives of Emergency Bridge by Applying Analytic Network Process (ANP). *SETP*, 27(3), 63–70 (2007).

94. Taguchi, G., Chowdhury, S., Wu, Y. *The Mahalanobis-Taguchi System* , Technology & Engineering (Google Books) ( 2001).

*95.* Tohumcu, Z., Karasakal, E. R&D Project Performance Evaluation With Multiple And Interdependent Criteria. *To appear in IEEE Transactions on Engineering Management*

96. Triantaphyllou, E. *Multi-criteria decision making methods: a comparative study*. (2000). Retrieved 08 28, 2009, from Google Books: http://books.google.com/books?id=tuPGe_ur-TYC

97. Tsai, W.H., Chou, W.C. Selecting management systems for sustainable development in SMEs: A novel hybrid model based on DEMATEL, ANP, and ZOGP. *Expert Systems with Applications,* 36, 1444–1458, (2009).

98. Tseng, M.L., Lin, Y.H. Application of fuzzy DEMATEL to develop a cause and effect model of municipal solid waste management in Metro Manila. *Environ Monit Assess,* 158, 519–533, (2009).

99. Tuzkaya, G., Önüt, S., Tuzkaya U.R., Gülsün, B. An analytic network process approach for locating undesirable facilities: An example from Istanbul, Turkey. *Journal of Environmental Management,* 88, 970–983, (2008).

100. Uysal, I., Güvenir, H. A. An overview of regression techniques for knowledge discovery. *The Knowledge Engineering Review,* 14, 4, 319-340, (1999).

101. Yao, C.C., Yu, P. Fuzzy regression based on asymmetric support vector machines. *Applied Mathematics and Computation,* 182, 175–193, (2006).

102. Ye N. *The Handbook of Data Mining.* Lawrence Erlbaum; 426-440, (April 1, 2003).

103. Yenidünya, B. *Robust Design with Binary Response Using Mahalanobis Taguchi System*. M.Sc. Thesis, Middle East Technical University, Industrial Engineering Department, Ankara, (2009).

104. Yerlikaya, F. *A New Contribution to Nonlinear Robust Regression and Classification with MARS and Its Applications to Data Mining for Quality Control in Manufacturing*. M.Sc. Thesis, Middle East Technical University, Applied Mathematics Department, Ankara, (2008).

105. Yoon, K., Hwang, C.L. *Multiple attribute decision making: an introduction*. (1995). Retrieved 08 20, 2009, from Google Books: http://books.google.com/books?id=Fo47SWBuEyMC

106. Yu, J.R., Cheng, S. J. Short Communication, An integrated approach for deriving priorities in analytic network process. *European Journal of Operational Research,* 180, 1427–1432, (2007).

107. Yurdakul, M. AHP as a strategic decision-making tool to justify machine tool selection. *Journal of Materials Processing Technology,* 146, 365–376, (2004).

108. Yurdakul, M., İç, Y.T. AHP approach in the credit evaluation of the manufacturing firms in Turkey. *Int. J. Production Economics,* 88, 269–289, (2004).

109. Yurdakul, M., İç, Y.T. Development of a performance measurement model for manufacturing companies using the AHP and TOPSIS approaches. *International Journal of Production Research,* 43, 21, 4609–4641 (1 November 2005).

110. Yurdakul, M. Measuring long-term performance of a manufacturing firm using the Analytic Network Process (ANP) approach. Int. J. Prod. Res., 41, 11, 2501-2529, (2003).

111. Yüksel İ., Dağdeviren M. Using the analytic network process (ANP) in a SWOT analysis – A case study for a textile firm. *Information Sciences,* 177, 3364–3382, (2007).

112. Virkkala, R., Luoto, M., Heikkinen R.K., Leikola N. Distribution patterns of boreal marshland birds: modeling the relationships to land cover and climate. *Journal of Biogeography,* 32, 1957–1970, (2005).

113. West, P.M., Brockett, P. L. and Golden L. L. A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice. *Marketing Science*, 16, 4, 370-391, (1997).

114. Wu, W. W. Choosing knowledge management strategies by using a combined ANP and DEMATEL approach. *Expert Systems with Applications*, 35, 828–835, (2008).

115. Wu, W. W., Lee, Y. T. Developing global manager's competencies using the fuzzy DEMATEL method. *Expert Systems with Applications*, 32, 499–507, (2007).

116. Zopounidis, C., Doumpos, M.. Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research,* 138, 229–246, (2002).

117. Zhue, J., Hastie, T. Classification of gene microarrays by penalized logistic regression. *Biostatistics,* 5, 3, 427–443, (2004).

# APPENDIX A

# PERFORMANCE MEASURES USED IN THE LITERATURE

**Table A. 1 Evaluations of some DM methods performed by Dhar and Stein (1997)**

| | **DM Methods** | | |
|---|---|---|---|
| **Criteria** | **Neural Network** | **Rule Based Systems** | **Machine Learning Algorithms (Decision Trees)** |
| Accuracy | High | High | Moderate to High |
| Explainability | Low | Moderate | Moderate to High |
| Response Speed | High | High | High |
| Scalability | Moderate | Moderate | Moderate to High |
| Compactness | High | High | Moderate |
| Flexibility | High | High | High |
| Embeddability | High | Moderate | Moderate to High |
| Ease of use | Moderate | Moderate | - |
| Tolerance for complexity | High | High | Moderate |
| Tolerance for noise in data | Moderate-High | - | Moderate |
| Tolerance for sparse data | Low | - | Low |
| Independence from experts | High | Moderate | Moderate |
| Development speed | Moderate | Moderate | Moderate |
| Computing resources | Low to Moderate | Low | Moderate |

**Table A. 2 Evaluations of some DM methods performed by Patel (2003)**

| | DM Methods | | | |
|---|---|---|---|---|
| **Criteria** | **MLR** | **LR** | **Neural Nets** | **Trees** |
| Accuracy | M | M | H | M |
| Interpretability | H | H | L | H |
| Speed-Training | H | H | L | HM |
| Speed- Deployment | H | H | H | HM |
| Effort in choice and transformation of indep.Vars. | HM | HM | L | L |
| Effort to tune performance parameters | L | L | H | ML |
| Robustness to Outliers in indep vars | ML | ML | HM | H |
| Robustness to irrelevant variables | H | H | L | ML |
| Ease of handling of missing values | M | M | M | H |
| Natural handling both categorical and continuous variables | H | H | H | H |

H: high, M:medium, L:low.

**Table A. 3 Reference List of the all Performance Measures encountered in the literature**

| # | Performance Measure | Reference |
|---|---|---|
| 1 | Accuracy | Dhar and Stein (1997)<br>Chien (2006)<br>Weiss and Zhang (2003) |
| 2 | Adjusted R Square | Montgomery and Runger (1996) |
| 3 | Application Time | Weiss and Zhang (2003) |
| 4 | Area Under Curve (AUC) / AUROC | Virkkala et al. (2005)<br>Fielding and Bell (1997)<br>Fawcett (2004)<br>Bradley (1997) |
| 5 | Confidence Interval (CI) | Fielding and Bell (1997) |
| 6 | Compactness | Dhar and Stein (1997) |
| 7 | Computational Efficiency | Moisen and Frescino (2002) |
| 8 | Computing Resource | Dhar and Stein (1997)<br>Chien (2006) |
| 9 | Contribution of Predictors | Moisen and Frescino (2002) |
| 10 | Correlation Coefficient | Moisen and Frescino (2002) |
| 11 | Cost of Obtaining Labeled Data | Weiss and Zhang (2003) |
| 12 | Development Speed/Effort | Dhar and Stein (1997)<br>Chien (2006) |
| 13 | Ease of Handling Missing Values | Patel (2003) |
| 14 | Ease of Use | Dhar and Stein (1997)<br>Chien (2006) |
| 15 | Effort in choice and transformation of independent variables | Patel (2003) |
| 16 | Effort to tune performance parameters | Patel (2003) |
| 17 | Embeddability | Dhar and Stein (1997)<br>Chien (2006) |
| 18 | Expert evaluation | Weiss and Zhang (2003) |
| 19 | Explicability | Dhar and Stein (1997)<br>Chien (2006) |
| 20 | F Measure | Fawcett (2004) |
| 21 | Field testing | Weiss and Zhang (2003) |
| 22 | Flexibility | Dhar and Stein (1997)<br>Chien (2006) |
| 23 | Independence from expert | Dhar and Stein (1997)<br>Chien (2006) |
| 24 | Interpretability | Patel (2003)<br>Weiss and Zhang (2003) |

**Table A. 3 (Continued) Reference List of the all Performance Measures encountered in the literature**

| # | Performance Measure | Reference |
|---|---|---|
| 25 | Kappa Statistics | Fielding and Bell (1997)<br>Moisen and Frescino (2002)<br>Cohen (1960)<br>Landis and Koch (1977)<br>Virkkala et al. (2005)<br>Cohen's kappa (2009) |
| 26 | Learning curve requirements | Dhar and Stein (1997) |
| 27 | Log-odds Ratio | Agresti (1996)<br>Fielding and Bell (1997) |
| 28 | Mallows' Cp | Mosteller and Tukey (1977) |
| 29 | Mean Absolute Error (MAE) | Brandimarte and Zotteri (2007) |
| 30 | Mean Absolute Percentage Error (MAPE) | Kim and Lee (1997) |
| 31 | Mean Error (ME) | Brandimarte and Zotteri (2007) |
| 32 | Mean Squared Error (MSE) | Hyndman and Koehler (2006)<br>Montgomery and Runger (1996) |
| 33 | Misclassification (error) rate | Classification and Regression Trees (C&RT) (2009) |
| 34 | Natural handling both categorical and continuous variables | Patel (2003) |
| 35 | Normalized mutual information (NMI) | Fielding and Bell (1997) |
| 36 | Operating Characteristics (OC) or Receiver Operating Characteristics (ROC) curve | Montgomery and Runger (1996) |
| 37 | Percent of Correctly Classified (PCC) | Moisen and Frescino (2002) |
| 38 | Precision | Han and Kamber (2001) |
| 39 | Predicted R2 | Montgomery and Runger (1996) |
| 40 | Prediction error sum of squares (PRESS) | Mosteller and Tukey (1977) |
| 41 | PWI (proportion of plots within some user-specified range) | Moisen and Frescino (2002) |
| 42 | R2 | Montgomery and Runger (1996) |

**Table A. 3 (Continued) Reference List of the all Performance Measures encountered in the literature**

| # | Performance Measure | Reference |
|---|---|---|
| 43 | Recall | Han and Kamber (2001) |
| 44 | Response speed | Dhar and Stein (1997) |
| 45 | Robustness to irrelevant variables | Patel (2003) |
| 46 | Robustness to outliers in independent variables | Patel (2003) |
| 47 | Root Mean Squared Error (RMSE) | Moisen and Frescino (2002)<br>Brandimarte and Zotteri (2007) |
| 48 | Scalability | Dhar and Stein (1997) |
| 49 | Sensitivity | Fielding and Bell (1997)<br>Han and Kamber (2001)<br>Patel (2003)<br>Weiss and Zhang (2003) |
| 50 | Specificity | Fielding and Bell (1997)<br>Han and Kamber (2001) |
| 51 | Speed-deployment | Patel (2003) |
| 52 | Speed-training | Patel (2003)<br>Weiss and Zhang (2003) |
| 53 | Stability | Bryson (2007)<br>Muata and Bryson (2004) |
| 54 | Tolerance for complexity | Dhar and Stein (1997)<br>Chien (2006) |
| 55 | Tolerance for data sparseness | Dhar and Stein (1997)<br>Chien (2007) |
| 56 | Tolerance for noise in data | Dhar and Stein (1997)<br>Chien (2006) |

# APPENDIX B

# DEFINITIONS OF THE SELECTED PERFORMANCE MEASURES

**Table B. 1 Initial decision criteria list for classification methods**

| # | **Criteria** |
|---|---|
| **1** | Misclassification (error) rate |
| **2** | Kappa |
| **3** | Precision |
| **4** | Recall (Sensitivity) |
| **5** | CI |
| **6** | Stability |
| **7** | F measure |
| **8** | AUROC |
| **9** | Scalability |
| **10** | Flexibility |
| **11** | Interpretability (explanatory capability) |
| **12** | Compactness |
| **13** | Embaddability |
| **14** | Natural handling both categorical and continuous variables |
| **15** | Robustness to complexity |
| **16** | Robustness to noise in data |
| **17** | Robustness to irrelevant variables |
| **18** | Robustness to missing values |
| **19** | Development speed/effort |
| **20** | Response Speed |

**Table B. 1 (Continued) Initial decision criteria list for classification methods**

| # | **Criteria** |
|---|---|
| 21 | Computing resource |
| 22 | Learning curve requirements |
| 23 | Independence from expert |

**Table B. 2 Initial decision criteria list for prediction methods**

| # | Criteria |
|---|---|
| 1 | Adjusted R2 (R-sq adj) |
| 2 | R2 (R-sq) |
| 3 | Mean Absolute Error (MAE) |
| 4 | Mean Square Error (MSE) |
| 5 | Root Mean Square Error (RMSE) |
| 6 | Stability of MSE |
| 7 | Stability of RMSE |
| 8 | Scalability |
| 9 | Flexibility |
| 10 | Interpretability (explanatory capability) |
| 11 | Compactness |
| 12 | Embaddability |
| 13 | Natural handling both categorical and continuous variables |
| 14 | Robustness to complexity |
| 15 | Robustness to noise in data |
| 16 | Robustness to irrelevant variables |
| 17 | Robustness to missing values |
| 18 | Development speed/effort |
| 19 | Response Speed |
| 20 | Computing resource |
| 21 | Learning curve requirements |
| 22 | Independence from expert |

In the following section important measures for Quality Improvement context and their definitions are listed.

**Notation used in the following section:**

*For Measures of Classification Methods*

A confusion matrix illustrates the accuracy of the solution to a classification problem. Obviously, the best results will have only zero values outside the diagonal.

**Table B. 3 Confusion Matrix (Contingency Table)**

| | | Predicted class | |
|---|---|---|---|
| | | 1 | 2 |
| Actual class | 1 | a | b |
| | 2 | c | d |

**Table B. 4 Confusion Matrix (where class of interest is 1)**

| | | Predicted class | |
|---|---|---|---|
| | | 1 | 2 |
| Actual class | 1 | TP=a | FN=b |
| | 2 | FP=c | TN=d |

$N$ : Total number of observations

$N = a + b + c + d$

***Given a specific class $i$ (class of interest)***

True Positive (TP) : predicted to be in class $i$ and is actually in it

False Positives (FP) : predicted to be in class $i$ but is not actually in it

True Negative (TN) : not predicted to be in class $i$ and is not actually in it

False Negative (FN) : not predicted to be in class $i$ but is actually in it

*For Measures of Prediction Methods*

$y_i$ = i $^{th}$ observed response value

$\hat{y}_i$ =i $^{th}$ fitted response

$\bar{y}$ =mean response

n = number of observation

p = number of terms in the model

$\bar{\hat{y}}$ =mean fitted response

n = number of observation

$s(y)^2$ = sample variance for observed response

$s(\hat{y})^2$ = sample variance for fitted response

$e_i = y_i - \hat{y}_i$ >> i th ordinary residual

$h_i$ >> leverage value for the i th observation

$h_i$ is the leverage of i th observation, which is the i th diagonal element of the hat matrix, H.

H = X (X'X)$^{-1}$ X' Where X is the design matrix.

***Adjusted R2 (R-sq adj)***

Accounts for the number of predictors in your model and is useful for comparing models with different numbers of predictors. The higher the R2, the better the model fits your data. The formula is:

$$R^2_{Adj} = 1 - \frac{MSE}{MSTotal} = 1 - \left( \frac{\sum (y_{i-} \hat{y}_i)^2}{\sum (y_{i-} \bar{y}_i)^2} \right) \left( \frac{n-1}{n-p-1} \right)$$

*In this study, $R^2_{adj}$ is not used since MSE is used and it is highly correlated with the $R^2_{adj}$.*

***R2 (R-sq)***

Coefficient of determination; indicates how much variation in response is explained by the model. The higher the R2, the better the model fits your data. The formula is:

$$R^2 = 1 - \frac{SSError}{SS\ Total} = 1 - \left( \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \right)$$

***Mean Absolute Error (MAE)***

MAE gives the average magnitude of error. Smaller is the better. The formula is:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

*In this study, MAE is not used since RMSE is used and it is highly correlated with the MAE.*

***Mean Absolute Percentage Error (MAPE)***

Gives scale independent (relative) error. Smaller is the better. The formula is:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right|$$

*In this study, MAPE is not used since RMSE is used and it is highly correlated with the MAPE.*

***Mean Square Error (MSE)***

MSE emphasizes grossly inaccurate estimates. Smaller is the better. The formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

*In this study, MSE is not used since RMSE is used and it is highly correlated with the MSE.*

***Root Mean Square Error (RMSE)***

RMSE gives magnitude with more weight on grossly inaccurate estimates. Smaller is the better. Model independent formula is:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

**Model dependent formula is**

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

***Correlation Coefficient***

It is a measure of linear association between actual and predicted response values. The formula is:

$$r = \frac{\sum_{i=1}^{n}(y - \bar{y})(\hat{y} - \bar{\hat{y}})/(n-1)}{\sqrt{s(y)^2 s(\hat{y})^2}}$$

*In this study measuring the association is not needed since correlation can be positive or negative and interpretation of this measure may be difficult. Thus, correlation coefficient is not used as a decision criterion and eliminated.*

***Prediction error sum of squares (PRESS)***

PRESS is an assessment of your model's predictive ability. PRESS, similar to the residual sum of squares, is the sum of squares of the prediction error. In general, the smaller the PRESS value, the better the model's predictive ability. In least squares regression, PRESS is calculated with the following formula:

$$\sum_{i=1}^{n}\left(\frac{e_i}{1-h_i}\right)^2$$

*In this study, PRESS is not used.*

***Predicted R2***

Indicate how well the model predicts responses for new observations. Larger values of predicted R2 suggest models of greater predictive ability. The higher is the better. The formula is

$$R^2(pred) = 1 - \frac{PRESS}{SS\ Total} = 1 - \frac{\sum_{i=1}^{n} (\frac{e_i}{1-h_i})^2}{1 - \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

*In this study, $R^2_{(pred)}$ is not used.*

***Misclassification rate (MCR) and Percentage of correctly classified (PCC)***

It is simply the number of misclassified observations divided by the total number of observations in the test set. If necessary, *b* and *c* (misclassified observations) can be weighted with cost. Suitable only if frequencies of different levels of discrete variables are similar. (Otherwise a biased measure)

$$\text{Misclassification (error) rate} = \frac{b+c}{N}$$

$$\text{Percentage of correctly classified PCC} = 1 - \frac{b+c}{N}$$

***Kappa***

It is proportion of correctly classified units after the probability of chance agreement has been removed. (Unbiased measure)

Kappa Statistics is an index which compares the agreement against that which might be expected by chance. Kappa can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement).

Kappa = (Observed agreement - Chance agreement)/(1 - Chance agreement)

$$Kappa = (\theta_1 - \theta_2)/(1 - \theta_2),$$

*where*

$$\theta_1 = \frac{a+d}{N} \quad observed\ agreement$$

$$\theta_2 = \frac{(a+b)}{N} * \frac{(a+c)}{N} + \frac{(b+d)}{N} * \frac{(c+d)}{N} \quad \text{chance agreement}$$

Kappa is always less than or equal to 1. A value of 1 implies perfect agreement and values less than 1 imply less than perfect agreement.

One drawback of the Kappa statistic is that this measure may be sensitive to the sample size and may fail when the size of one class exceeds the other. (Fielding and Bell, 1997)

In different sources, a rough guide is proposed to assess the Kappa (Landis and Koch, 1977) (Cohen's kappa, 2009)

**Table B. 5 A rough guide to assess the Kappa statistic (not universally accepted)**

| Kappa | Strength of agreement |
|---|---|
| 0.00 | Poor |
| 0.01-0.20 | Slight |
| 0.21-0.40 | Fair |
| 0.41-0.60 | Moderate |
| 0.61-0.80 | Substantial |
| 0.81-1.00 | Almost perfect |

***Precision***

An indicator of sharpness in identifying class of interest

$$precision = \frac{TP}{TP + FP}$$

***Recall (Sensitivity)***

An indicator of hitting all cases of interest

$$recall = \frac{TP}{TP + FN}$$

***F measure***

There is trade-off between precision and recall. For high precision, hit rate is bound to drop. However to hit all the positives, the rule set has to shoot many false negatives as well. F Measure combines these to see the joint effect.

It is the weighted harmonic mean of precision and recall and tries to see how the tradeoff between precision and recall, is resolved.

$$F = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

This is also known as the $F_1$ measure, because recall and precision are evenly weighted.

The general formula for non-negative real ß is:

$$F_\beta = \left(1 + \beta^2\right) \cdot \left(precision \cdot recall\right) / (\beta^2 \cdot precision + recall)$$

Two other commonly used F measures are the $F_2$ measure, which weights recall twice as much as precision, and the $F_{0,5}$ measure, which weights precision twice as

much as recall. As ß increases the weight of recall increases in the measure. (F1 Score, 2009)

*F measure, which is a weighted combination of precision and recall, is not used in this study since precision and recall are both used in this study.*

***Proportion of plots within some user-specified range (PWI)***

PWI is the sum of indicator variables over all observations. The indicator variables take the value of one if the absolute value of the difference between actual and predicted response is within some user-specified thresholds. "1- PWI2" is also used to measure outliers of the observations. PWI2 uses 2σ as user specified range R.

$$PWI = \frac{1}{n}\sum_{i=1}^{n} I\left\{ \left| \hat{y}_i - y_i \right| < R \right\}$$

*PWI is not used in this study, since "robustness to noise in data" sub-criteria has a similar interpretation with 1-PWI2 measure.*

***Confidence Interval (CI)***

When the data partitioning methods such as bootstrapping, randomization, k-fold partitioning etc. are used accuracy is usually reported as a mean and confidence limits.

***Stability***

A classification/prediction model is stable when it performs just as well on both seen (training) and unseen (test) data sets. The stability can be measured as a number between 0 and 1, where 0 means completely stable and 1 means completely unstable.

$$Stability = \frac{CC_{Train} - CC_{Test}}{CC_{Train} + CC_{Test}}$$ where

$CC_{Train}$: Correct classification rate of the training set. $CC_{Train} = \frac{a+d}{N}$

$CC_{Test}$: Correct classification rate of the testing set. $CC_{Test} = \frac{a+d}{N}$

***Area Under Curve (AUC)***

AUC shows the area under the "Receiver Operating Characteristics Curve" (ROC). ROC is a curve of sensitivity versus (1-specificity) over a range of cutoff points. When the cutoff points are very high (i.e. 1.0), all claims are classified as legitimate. The baseline ROC curve (where no model is used) can be thought of as a straight line from the origin with a 45-degree angle. If the model's sensitivity increases faster than the specificity decreases, the curve "lifts" or rises above a 45-degree line quickly. The higher the "lift", the more accurate the model.

A statistic that summarizes the predictive accuracy of a model as measured by an ROC curve is the area under the ROC curve (AUROC). A curve that rises quickly has more area under the ROC curve.

Operating Characteristics (OC) or Receiver Operating Characteristics (ROC) curve:

A ROC plot is obtained by plotting all true positive fractions on the y-axis against their equivalent false positive fraction for all available thresholds on the x-axis.

$$\frac{a}{a+b} \quad vs. \quad \frac{c}{c+d}$$

AUC shows the area under the "Receiver Operating Characteristics Curve" (ROC). ROC is a curve of sensitivity versus (1-specificity) over a range of cutoff points. When dealing with highly skewed datasets, it gives overly optimistic view.

***Log-odds Ratio***

It measures the association between two binary variables. High association does not guarantee the model accuracy.

If we have two binary variables A and B, to look for a measure of association between the components the most useful general measure is the log-odds ratio, defined as follows.

Given A=$i$ , the odds for B=1 versus B=0 are

P(B=1/A= $i$ ) /P(B=0/A=$i$ )= $\pi_{i1}^{AB} / \pi_{i0}^{AB}$

If A and B are independent this ratio is the same at both levels of A. Hence the ratio of the separate odds ratios is a measure of association taking the value 1 for independent components. The log of the ratio is more convenient for many purposes, being zero when independence holds, so that we are led to define

$$\psi_{AB} = \log\left\{\left(\pi_{11}^{AB}\pi_{00}^{AB}\right)/\left(\pi_{10}^{AB}\pi_{01}^{AB}\right)\right\}$$

According to confusion matrix entries it is identical to the $\psi_{AB} = \log\left(\frac{ad}{bc}\right)$

*This study concerns with quality data, measuring the association is meaningless since high association may imply high misclassification or high correct classification rate. Thus, log-odds ratio is not used as a decision criterion in this study. It is eliminated by the decision makers.*

***Scalability:***

It refers to how well the system works when new variables are added or range of the values that variables can take is increased.

***Flexibility:***

It is the ease with which the relationships among the variables or their domains can be changed, or the goals of the system modified. Robustness to perform well as additional functionality added over time.

***Ease of use of the model:***

It describes how complicated the system is to use for the business people who will be using it on a daily basis.

This criterion consists of several decision criteria such as interpretability, compactness and embeddability. Thus, it is used as a cluster caption in this study.

***Interpretability***

Interpretability of a method can be defined as ability of extracting information that can be verified by experts. All the recursive partitioning algorithms have the interpretability property.

***Compactness***

It refers to how small the system can be made. Compactness deals with the ease with which the system can be encoded into a compact portable format.

***Embeddability***

It refers to the ease with which a system can be coupled with or incorporated into the infrastructure of an organization.

***Natural handling both categorical and continuous variables (Robustness to categorical and continuous variables)***

This is ability of the method to handle both categorical and continuous variables.

*This criterion is renamed as "Robustness to categorical and continuous variables" in this study.*

***Tolerance for complexity (Robustness to complexity)***

It refers to the degree to which the quality of a system is affected by interactions among the various components of the process being modeled or in the knowledge used to model a process. This criterion also covers the ability to detect interactions.

*This criterion is renamed as "Robustness to complexity" in this study.*

***Tolerance for noise in data (Robustness to noise in data)***

It is the degree to which the accuracy of a system is affected by noise in the data.

*This criterion is renamed as "Robustness to noise in data" in this study.*

***Effort in choice and transformation of independent variables***

This is required effort to choose the relevant attributes and transform the data into appropriate format by using data transformation techniques such as smoothing, normalization etc.

*Since almost all of the prediction and classification methods need same effort to choice and transformation of independent variables, it is not a discriminatory decision criterion and excluded from the initial decision criteria list by the decision makers.*

***Tolerance for data sparseness***

It is the degree to which the quality of a system is affected by incompleteness or lack of data.

The availability and level of detail of data and the accuracy are central issues in choosing among different techniques.

Sparse data occurs when many data cells in a data item contain NA values. For example, if a financial data item contains information that is dimensioned by Product and Market, it is likely that the data will be sparse because not all products are sold in all markets.

*Data sparseness is not a common problem of quality data. Thus, "Robustness to data sparseness" is not used as a decision criterion in this study. It is eliminated by the decision makers.*

***Ease of handling of missing values (Robustness to Missing Values)***

It is the degree to which the quality of a system is affected by missing values of data.

*This criterion is renamed as "Robustness to Missing Values" in this study.*

***Development speed/effort:***

The time that the organization can afford to develop a system or, conversely, the time a modeling technology would require to develop a system. (Dhar and Stein, 1997)

*In this study Development speed is selected as a decision criterion since it is the most comprehensive one. It consists following measures:*

*Computational time : This is the computation time required for an algorithm to generate a model for a given dataset.*

*Speed Training,*

*Speed deployment,*

*Effort in choice and transformation of independent variables,*

*Effort to tune performance parameters.*

***Response Speed***

It is the time it takes for a system to complete analysis at the desired level of accuracy. The flip side to this dimension is confidence in the sense that you can ask how confident you are that a certain period of time, within which the system must provide an answer, will be sufficient to perform the analysis. In applications that require that results be produced within a specified time frame missing that time frame means that no matter how accurate and otherwise desirable the results are, they will be useless in practice.

***Computing resource (computational ease)***

It is the degree to which a system can be implemented without requiring special-purpose hardware and software.

***Learning curve requirements***

These requirements indicate the degree to which the organization needs to experiment in order to become sufficiently competent at solving a problem or using a technique.

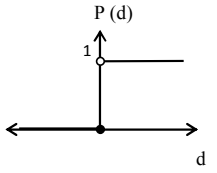
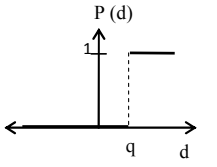
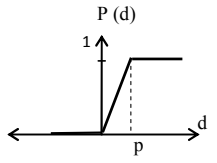
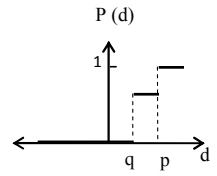
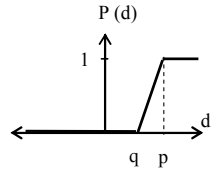
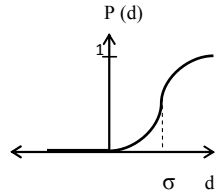
***Independence from expert***

It is the degree to which the system can be designed, built and tested without experts.

# APPENDIX C

# PROMETHEE PREFERENCE FUNCTIONS

**Table C. 1 PROMETHEE Preference Functions**

| Type | Graph | Parameter | Function |
|---|---|---|---|
| I. Usual Criterion |  | - | $P(d)=\begin{cases}0, & d \le 0\\ 1, & d>0\end{cases}$ |
| II. Quasi Criterion |  | q | $P(d)=\begin{cases}0, & q \ge d\\ 1, & d>q\end{cases}$ |
| III. Criterion with Linear Preference |  | p | $P(d)=\begin{cases}d/p, & p \ge d \ge 0\\ 1, & d>p\end{cases}$ |
| IV. Level Criterion |  | q, p | $P(d)=\begin{cases}0, & q \ge d\\ 1/2, & q<d \le p\\ 1, & p<d\end{cases}$ |
| V. Criterion with Linear Preference and Indifference area |  | q, p | $P(d)=\begin{cases}0, & q \ge d\\ (d-q)/(p-q), & q<d \le p\\ 1, & p<d\end{cases}$ |
| VI. Gaussian Criterion |  | σ | $P(d)=\begin{cases}0, & d \le 0\\ 1-\exp\left(-d^2/2\sigma^2\right), & d>0\end{cases}$ |

# APPENDIX D

# DECISION MAKER / EXPERT LIST

**Table D. 1 Experts who contributed to ANP and PROMETHEE evaluations**

| Name | Organization | Position Title | E-mail |
|---|---|---|---|
| Prof.Dr. Gülser KÖKSAL | METU Industrial Engineering Department | Faculty member | koksal@ie.metu.edu.tr |
| Prof. Dr. Sinan KAYALIGİL | METU Industrial Engineering Department | Faculty member | skayali@ie.metu.edu.tr |
| Assoc. Prof. Dr. İnci BATMAZ | METU Department of Statistics | Faculty member | ibatmaz@metu.edu.tr |
| Berna BAKIR | METU Informatics Institute | Research Assistant, Ph.D. Candidate | berna@ii.metu.edu.tr |
| Elçin KARTAL | METU Department of Statistics | Research Assistant, Ph.D. Candidate | kartalelcin@gmail.com |
| Fatma YERLİKAYA ÖZKURT | METU Institute of Applied Mathematics | Research Assistant, Ph.D. Student | fatmayerlikaya@gmail.com |

**Table D. 2 Experts who contributed to PROMETHEE evaluations**

| Name | Organization | Position Title | E-mail |
|---|---|---|---|
| Assist. Prof. Dr. Süreyya ÖZÖĞÜR AKYÜZ | Bahçeşehir University Mathematics and Computer Sciences Department | Faculty member | sureyya.akyuz @bahcesehir.edu.tr |
| Barış YENİDÜNYA | METU Industrial Engineering Department | M.S. Student | - |
| Dilber AYHAN | METU Industrial Engineering Department | M.S. Student | - |
| Gizem ÖZER | METU Industrial Engineering Department | M.S. Student | - |
| Ezgi AVCI | METU Industrial Engineering Department | M.S. Student | - |
| Tuna KILIÇ | Çankaya University Industrial Engineering Department | M.S. Student | |

# APPENDIX E

# STATISTICAL ANALYSIS OF MEAN ACCURACY MEASURES

**Correlations: MCR; Precision; Recall; F0.5; F1; F2; Kappa; Specificity; Stability of PCC**

| | MCR | Precision | Recall | F0.5 | F1 | F2 | Kappa | Specific | Stab_PCC |
|---|---|---|---|---|---|---|---|---|---|
| Precision | -0,710 | | | | | | | | |
| | 0,000 | | | | | | | | |
| Recall | -0,360 | 0,398 | | | | | | | |
| | 0,000 | 0,000 | | | | | | | |
| F0.5 | -0,815 | 0,898 | 0,564 | | | | | | |
| | 0,000 | 0,000 | 0,000 | | | | | | |
| F1 | -0,759 | 0,705 | 0,797 | 0,934 | | | | | |
| | 0,000 | 0,000 | 0,000 | 0,000 | | | | | |
| F2 | -0,595 | 0,480 | 0,947 | 0,774 | 0,944 | | | | |
| | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | | | | |
| Kappa | -0,787 | 0,814 | 0,808 | 0,950 | 0,975 | 0,890 | | | |
| | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | | | |
| Specific | -0,723 | 0,654 | -0,308 | 0,605 | 0,376 | 0,095 | 0,291 | | |
| | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,328 | 0,001 | | |
| Stab_PCC | 0,797 | -0,633 | -0,022 | -0,734 | -0,647 | -0,456 | -0,490 | -0,819 | |
| | 0,000 | 0,000 | 0,809 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | |
| AUC | -0,778 | 0,542 | 0,565 | 0,761 | 0,806 | 0,735 | 0,769 | 0,356 | -0,575 |
| | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 |

Cell Contents: Pearson correlation

P-Value

**Figure E.1 Correlation coefficients and p values of the accuracy measures for classification methods**

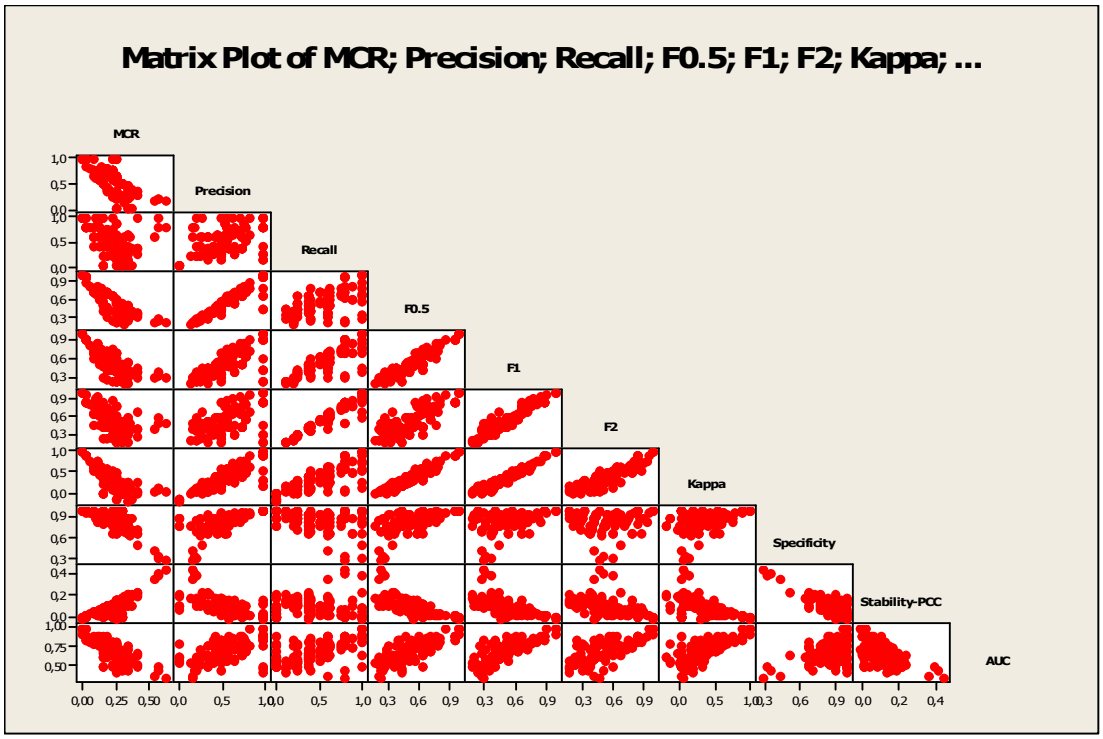



**Figure E. 2 Correlation matrix plot of the accuracy performance measures for classification methods**

**Correlations: MAE; MSE; RMSE; R; R2; Adj-R2; PWI1; PWI2; Stability_MSE; Stability_RMSE**

| | MAE | MSE | RMSE | R | R2 | Adj-R2 | PWI1 | PWI2 | Stab_MSE |
|---|---|---|---|---|---|---|---|---|---|
| MSE | 0,821 | | | | | | | | |
| | 0,000 | | | | | | | | |
| RMSE | 0,910 | 0,950 | | | | | | | |
| | 0,000 | 0,000 | | | | | | | |
| R | -0,409 | -0,324 | -0,499 | | | | | | |
| | 0,001 | 0,010 | 0,000 | | | | | | |
| R2 | -0,396 | -0,280 | -0,437 | 0,946 | | | | | |
| | 0,001 | 0,026 | 0,000 | 0,000 | | | | | |
| Adj-R2 | -0,555 | -0,889 | -0,802 | 0,298 | 0,239 | | | | |
| | 0,000 | 0,000 | 0,000 | 0,018 | 0,059 | | | | |
| PWI1 | 0,231 | 0,082 | 0,163 | -0,241 | -0,399 | -0,071 | | | |
| | 0,069 | 0,523 | 0,202 | 0,057 | 0,001 | 0,578 | | | |
| PWI2 | 0,207 | 0,034 | 0,101 | -0,168 | -0,345 | -0,006 | 0,960 | | |
| | 0,104 | 0,791 | 0,432 | 0,188 | 0,006 | 0,963 | 0,000 | | |
| Stab_MSE | -0,534 | -0,323 | -0,447 | 0,237 | 0,290 | 0,212 | -0,376 | -0,301 | |
| | 0,000 | 0,010 | 0,000 | 0,061 | 0,021 | 0,095 | 0,002 | 0,017 | |
| Sta_RMSE | -0,620 | -0,425 | -0,530 | 0,265 | 0,346 | 0,300 | -0,461 | -0,406 | 0,973 |
| | 0,000 | 0,001 | 0,000 | 0,036 | 0,005 | 0,017 | 0,000 | 0,001 | 0,000 |

Cell Contents: Pearson correlation
P-Value

**Figure E. 3 Correlation coefficients and p values of the accuracy measures for prediction methods**

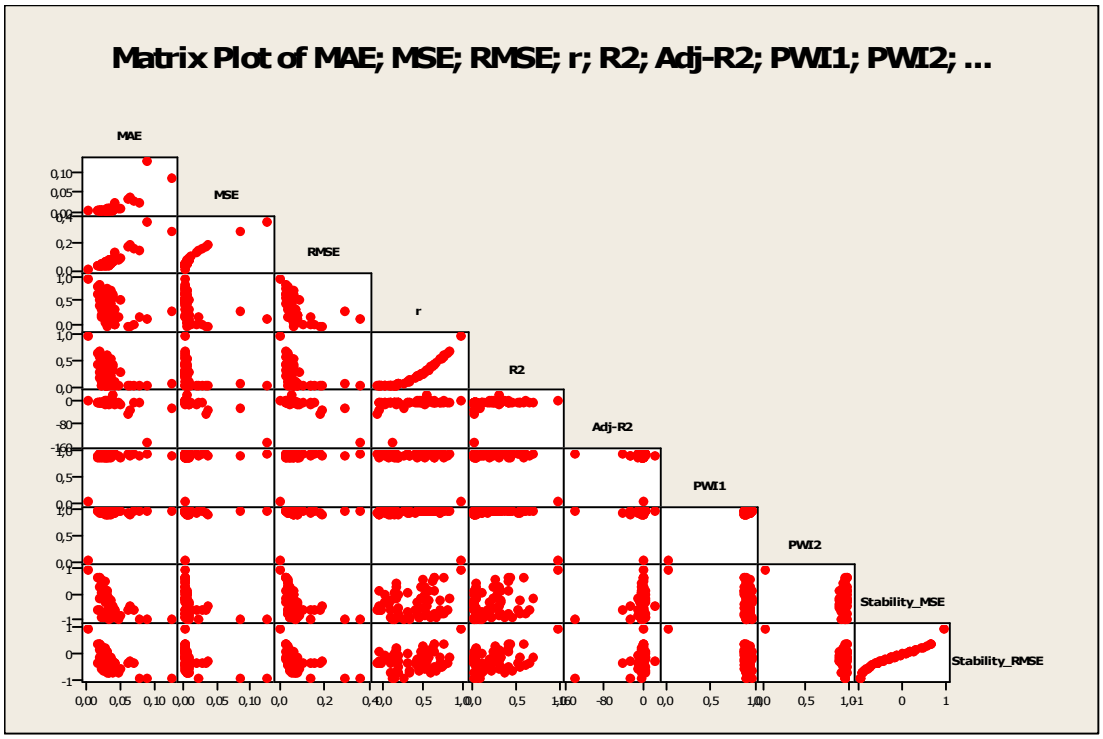



**Figure E. 4 Correlation matrix plot of the accuracy performance measures for prediction methods**

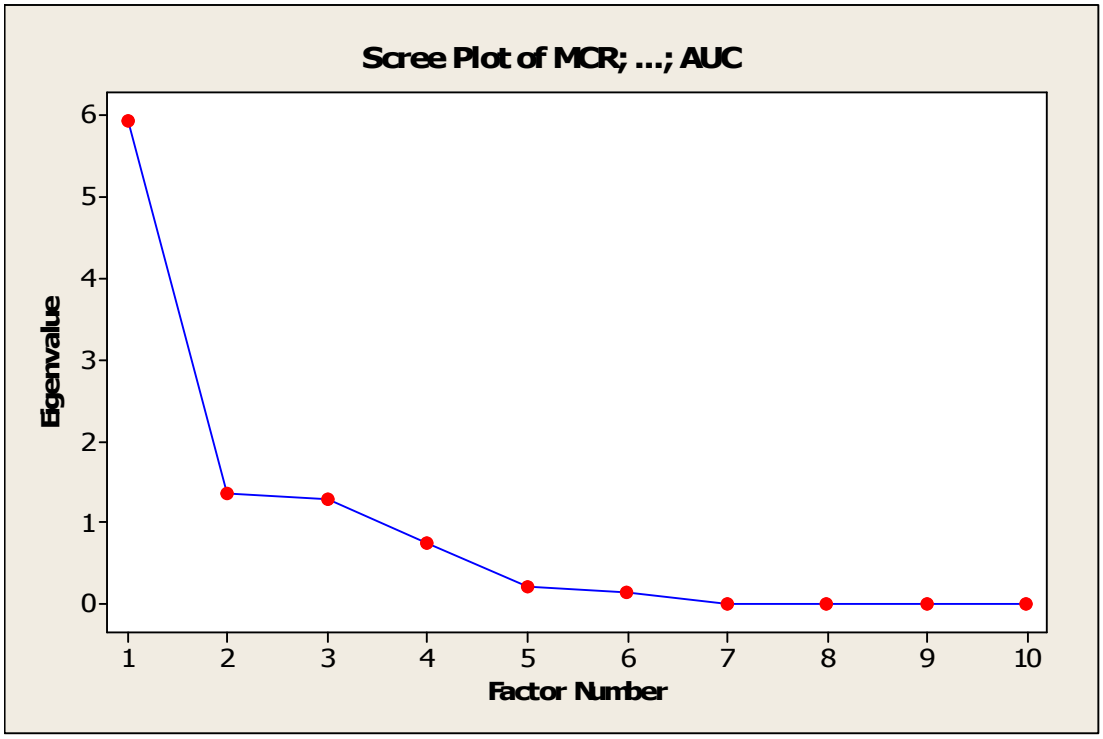



**Figure E. 5 Scree Plot of the factor analysis for MCR, precision, recall, F 0.5, F1, F2, kappa, specificity, stability of PCC, AUC**

According to Figure E. 5, number of factors is determined as 6. It appears that three-factor solution effectively summarizes the total variance (0.951). The factor analysis is conducted for 6 factors and the results are given in Figure E. 6. Recall, $F_1$, and $F_2$ form a group; MCR, specificity and stability form another group. Kappa also belongs to the "Recall, $F_1$, $F_2$" group and AUC can be associated with the "MCR, specificity, stability" group. Precision forms another group and $F_{0.5}$ is closer to the "Precision" group than the other groups. As a result, representing these measures only MCR, precision, recall, kappa, stability and AUC are selected. Here, even though kappa AUC and stabilitycould be eliminated (due to the fact that they highly correlate with recall and MCR respectively) our experts have found it useful for them to be explicitely in the analysis.

**Factor Analysis: MCR; Precision; Recall; F0.5; F1; F2; Kappa; Specificity; Stability of PCC**

Maximum Likelihood Factor Analysis of the Correlation Matrix

Rotated Factor Loadings and Communalities
Varimax Rotation

| Variable | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 | Factor6 | Communality |
|---|---|---|---|---|---|---|---|
| MCR | -0,395 | → **-0,864** | -0,272 | -0,024 | 0,073 | 0,090 | 0,991 |
| Precision | 0,244 | 0,465 | → **0,844** | 0,031 | -0,002 | 0,006 | 0,989 |
| Recall | → **0,993** | -0,063 | 0,051 | 0,080 | 0,014 | -0,028 | 1,000 |
| F0.5 | 0,578 | 0,486 | **0,635** | -0,155 | 0,047 | -0,009 | 1,000 |
| F1 | **0,820** | 0,389 | 0,389 | -0,154 | 0,013 | 0,027 | 1,000 |
| F2 | **0,962** | 0,195 | 0,184 | -0,043 | -0,005 | 0,040 | 1,000 |
| Kappa | → **0,742** | 0,513 | 0,420 | -0,037 | -0,032 | -0,082 | 1,000 |
| Specificity | -0,161 | **0,877** | 0,427 | -0,019 | -0,058 | -0,033 | 0,981 |
| Stability | -0,245 | → **-0,858** | -0,256 | 0,066 | -0,084 | -0,092 | 0,881 |
| AUC | 0,608 | → **0,621** | 0,152 | -0,029 | 0,147 | 0,017 | 0,801 |
| Variance | 4,1398 | 3,5469 | 1,8233 | 0,0647 | 0,0411 | 0,0278 | 9,6436 |
| % Var | 0,414 | 0,355 | 0,182 | 0,006 | 0,004 | 0,003 | 0,964 |

**Figure E. 6 Rotated factor loadings and communalities of MCR, precision, recall, $F_{0.5}$, $F_1$, $F_2$, kappa, specificity, stability of PCC, AUC**

According to Figure E. 7 the number of factors is chosen as 6. It appears that four-factor solution effectively summarizes the total variance (0.924). The factor analysis is conducted for six factors and the results are given in Figure E. 8. MSE, Adj R2 and RMSE form a group; R and R2 form another group. Stability of MSE and that of stability of RMSE also form a group as well as PWI1 and PWI2. From these groups only RMSE, R2 and stability of RMSE are selected as representatives. PWI measures are eliminated completely thinking that they have a similar meaning with "robustness to noise in data"

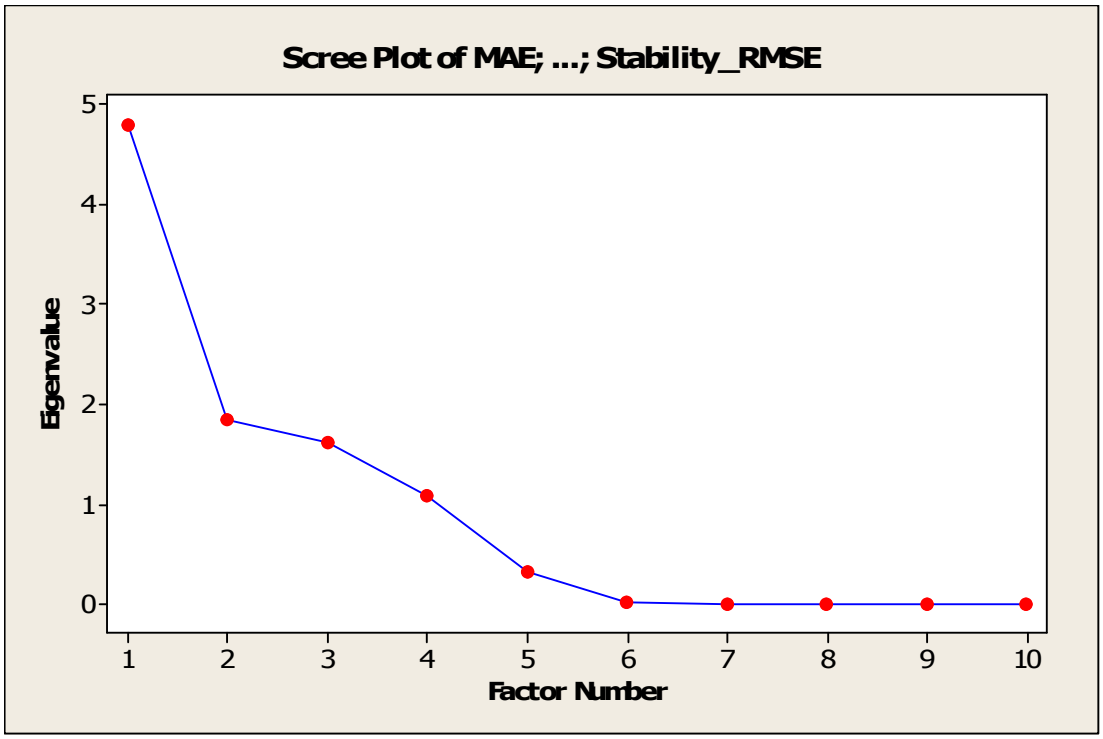



**Figure E. 7 Scree Plot of the factor analysis for MAE, MSE, RMSE, R, R2, Adj R2, PWI1, PWI2, Stability of MSE and Stability of RMSE measures**

**Factor Analysis: MAE; MSE; RMSE; R; R2; Adj-R2; PWI1; PWI2; Stability_MSE; Stability_RMSE**

Maximum Likelihood Factor Analysis of the Correlation Matrix

Rotated Factor Loadings and Communalities
Varimax Rotation

| Variable | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 | Factor6 | Communality |
|---|---|---|---|---|---|---|---|
| MAE | **-0,699** | 0,377 | 0,210 | -0,108 | -0,523 | 0,004 | 0,960 |
| MSE | **-0,964** | 0,161 | 0,110 | 0,001 | -0,122 | 0,043 | 0,983 |
| RMSE | → **-0,879** | 0,267 | 0,280 | -0,030 | -0,247 | -0,066 | 0,989 |
| R | 0,215 | -0,087 | **-0,952** | 0,061 | 0,042 | 0,073 | 0,971 |
| R2 | 0,151 | -0,137→ | **-0,928** | 0,232 | 0,032 | -0,077 | 0,963 |
| Adj-R2 | **0,925** | -0,064 | -0,107 | 0,004 | -0,243 | 0,001 | 0,931 |
| PWI1 | -0,037 | 0,211 | 0,162 | **-0,940** | 0,013 | -0,042 | 0,957 |
| PWI2 | 0,008 | 0,150 | 0,105 | **-0,961** | -0,051 | 0,042 | 0,961 |
| Stability_MSE | 0,161 | **-0,958** | -0,109 | 0,156 | 0,042 | 0,040 | 0,984 |
| Stability_RMSE | 0,263 | → **-0,909** | -0,122 | 0,261 | 0,071 | -0,047 | 0,986 |
| Variance | 3,2110 | 2,0820 | 1,9775 | 1,9680 | 0,4208 | 0,0248 | 9,6842 |
| % Var | 0,321 | 0,208 | 0,198 | 0,197 | 0,042 | 0,002 | 0,968 |

**Figure E. 8 Rotated factor loadings and communalities of MAE, MSE, RMSE, R, R2, Adj R2, PWI1, PWI2, Stability of MSE and Stability of RMSE**

# APPENDIX F

# RELATION MATRICES

Interpretation of the relation matrices is that if an element in the row has effect on the element in column, corresponding cell is marked with √ .

| | | CLASSIFICATION | 1. Prediction Accuracy | | | | | | | 2. Ease of Use of the Model | | | 3. Robustness | | | | | 4. Speed | | | 5. Ease of Modelling | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1.1. | 1.2. | 1.3. | 1.4. | 1.5. | 1.6. | 1.7. | 2.1. | 2.2. | 2.3. | 3.1. | 3.2. | 3.3. | 3.4. | 3.5. | 4.1. | 4.2. | 4.3. | 5.1. | 5.2. | 5.3. | 5.4. |
| | | | Misclassification Rate | Kappa | CI | Stability of PCC | Recall | Precision | AUROC | Interpretability | Compactness | Embaddability | To categorical and continuous variables | To complexity | To Noise in Data | To Irrelevant Variables | To Missing Values | Learning Curve Requirements | Development speed/effort | Response Speed | Computing Resource | Independence From Expert | Scalability | Flexibility |
| 1. Prediction Accuracy | 1.1. | Misclassification Rate | | √ | √ | √ | √ | √ | √ | | | | | | | | | √ | √ | √ | | | | |
| | 1.2. | Kappa | √ | | √ | √ | √ | √ | √ | | | | | | | | | √ | √ | √ | | | | |
| | 1.3. | CI | √ | √ | | √ | √ | √ | √ | | | | | | | | | √ | √ | √ | | | | |
| | 1.4. | Stability of PCC | √ | √ | √ | | √ | √ | √ | | | | | | | | | √ | √ | √ | | | | |
| | 1.5. | Recall | √ | √ | √ | √ | | √ | √ | | | | | | | | | √ | √ | √ | | | | |
| | 1.6. | Precision | √ | √ | √ | √ | √ | | √ | | | | | | | | | √ | √ | √ | | | | |
| | 1.7. | AUROC | √ | √ | √ | √ | √ | √ | | | | | | | | | | √ | √ | √ | | | | |
| 2. Ease of Use of the Model | 2.1. | Interpretability | | | | | | | | | √ | √ | | | | | | | | | | √ | | |
| | 2.2. | Compactness | | | | | | | | √ | | √ | | | | | | | | | √ | | | |
| | 2.3. | Embaddability | | | | | | | | √ | √ | | | | | | | | | | | | | |
| 3. Robustness | 3.1. | To categorical and continuous | √ | √ | √ | √ | √ | √ | √ | | | | | | | | | √ | √ | √ | √ | √ | | √ |
| | 3.2. | To complexity | √ | √ | √ | √ | √ | √ | √ | | √ | | | | | | | √ | √ | √ | √ | √ | √ | √ |
| | 3.3. | To Noise in Data | √ | √ | √ | √ | √ | √ | √ | | | | | | | | | √ | √ | √ | √ | √ | √ | √ |
| | 3.4. | To Irrelevant Variables | √ | √ | √ | √ | √ | √ | √ | | | | | | | | | √ | √ | √ | √ | √ | √ | √ |
| | 3.5. | To Missing Values | √ | √ | √ | √ | √ | √ | √ | | | | | | | | | √ | √ | √ | √ | √ | √ | √ |
| 4. Speed | 4.1. | Learning Curve Requirements | √ | √ | √ | √ | √ | √ | √ | | | | | | | | | | √ | | | √ | | |
| | 4.2. | Development speed/effort | √ | √ | √ | √ | √ | √ | √ | | | | | | | | | √ | | √ | | | √ | √ |
| | 4.3. | Response Speed | | | | | | | | | | √ | | | | | | | | | | | | |
| 5. Ease of Modelling | 5.1. | Computing Resource | √ | √ | √ | √ | √ | √ | √ | √ | | | | | | | | | √ | √ | | √ | √ | √ |
| | 5.2. | Independence From Expert | √ | √ | √ | √ | √ | √ | √ | √ | | √ | | | | | | √ | √ | | √ | | | |
| | 5.3. | Scalability | √ | √ | √ | √ | √ | √ | √ | √ | √ | √ | | | | | | | √ | √ | | | | √ |
| | 5.4. | Flexibility | √ | √ | √ | √ | √ | √ | √ | | √ | √ | | | | | | | √ | √ | | | √ | |

**Figure F. 1 Relation Matrix of decision criteria (for classification methods)**

| PREDICTION | | | 1. Prediction Accuracy | | | 2. Ease of Use of the Model | | | 3. Robustness | | | | | 4. Speed | | | 5. Ease of Modelling | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1.1. | 1.2. | 1.3. | 2.1. | 2.2. | 2.3. | 3.1. | 3.2. | 3.3. | 3.4. | 3.5. | 4.1. | 4.2. | 4.3. | 5.1. | 5.2. | 5.3. | 5.4. |
| | | | RMSE | Stability of RMSE | R Square | Interpretability | Compactness | Embaddability | To categorical and continuous variables | To complexity | To Noise in Data | To Irrelevant Variables | To Missing Values | Learning Curve Requirements | Development speed/effort | Response Speed | Computing Resource | Independence From Expert | Scalability | Flexibility |
| 1. Prediction Accuracy | 1.1. | RMSE | | √ | √ | | | | | | | | | √ | √ | √ | | | | |
| | 1.2. | Stability of RMSE | √ | | √ | | | | | | | | | √ | √ | √ | | | | |
| | 1.3. | R Square | √ | √ | | | | | | | | | | √ | √ | √ | | | | |
| 2. Ease of Use of the Model | 2.1. | Interpretability | | | | | √ | √ | | | | | | | | | | √ | | |
| | 2.2. | Compactness | | | | √ | | √ | | | | | | | | | √ | | | |
| | 2.3. | Embaddability | | | | √ | √ | | | | | | | | | | | | | |
| 3. Robustness | 3.1. | To categorical and continuous | √ | √ | √ | | | | | | | | | √ | √ | √ | √ | √ | | √ |
| | 3.2. | To complexity | √ | √ | √ | | √ | | | | | | | √ | √ | √ | √ | √ | √ | √ |
| | 3.3. | To Noise in Data | √ | √ | √ | | | | | | | | | √ | √ | √ | √ | √ | √ | √ |
| | 3.4. | To Irrelevant Variables | √ | √ | √ | | | | | | | | | √ | √ | √ | √ | √ | √ | √ |
| | 3.5. | To Missing Values | √ | √ | √ | | | | | | | | | √ | √ | √ | √ | √ | √ | √ |
| 4. Speed | 4.1. | Learning Curve Requirements | √ | √ | √ | | | | | | | | | | √ | | | √ | | |
| | 4.2. | Development speed/effort | √ | √ | √ | | | | | | | | | √ | | √ | | | √ | √ |
| | 4.3. | Response Speed | | | | | | √ | | | | | | | | | | | | |
| 5. Ease of Modelling | 5.1. | Computing Resource | √ | √ | √ | √ | | | | | | | | | √ | √ | | √ | √ | √ |
| | 5.2. | Independence From Expert | √ | √ | √ | √ | | √ | | | | | | √ | √ | | √ | | | |
| | 5.3. | Scalability | √ | √ | √ | √ | √ | √ | | | | | | | √ | √ | | | | √ |
| | 5.4. | Flexibility | √ | √ | √ | | √ | √ | | | | | | | √ | √ | | | √ | |

**Figure F. 2 Relation Matrix of decision criteria (for prediction methods)**

# APPENDIX G

# THE QUESTIONNAIRE

The questionnaire consists of five types of questions:

**Type-1 Pairwise comparison of criteria with respect to goal:** Type-1 Questions are same for prediction and classification methods.

**Type-2 Pairwise comparison of criteria with respect to criteria:** Type-2 Questions are same for prediction and classification methods.

**Type-3 Pairwise comparison of sub-criteria with respect to criteria:** Almost all of the Type-3 Questions are same for prediction and classification methods. Only questions related with Predictive accuracy and its sub-criteria are different.

**Type-4 Pairwise comparison of sub-criteria with respect to sub-criteria:** Most of the Type-4 Questions are different for prediction and classification methods.

**Type-5 Pairwise comparison of criteria with respect to sub-criteria (feedback):** Only questions related with Predictive accuracy and its sub-criteria are different for prediction and classification methods.

Since there are 1571 pairwise comparisons for classification and 668 for prediction methods, in this section only sample questions are illustrated for each question type.

**Table G. 1 Type-1 Pairwise comparison of criteria with respect to goal ( For both prediction and classification methods)**

| | "Which criterion should be emphasized more for evaluation of method performance? | | | How much more? 1 3 5 7 9 |
|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Ease of Use of the Model ☐ | ☐ ☐ ☐ ☐ ☐ |
| 2 | Predictive Accuracy | ☐ | Robustness ☐ | ☐ ☐ ☐ ☐ ☐ |
| 3 | Predictive Accuracy | ☐ | Speed ☐ | ☐ ☐ ☐ ☐ ☐ |
| 4 | Predictive Accuracy | ☐ | Ease of Modelling ☐ | ☐ ☐ ☐ ☐ ☐ |
| 5 | Ease of Use of the Model | ☐ | Robustness ☐ | ☐ ☐ ☐ ☐ ☐ |
| 6 | Ease of Use of the Model | ☐ | Speed ☐ | ☐ ☐ ☐ ☐ ☐ |
| 7 | Ease of Use of the Model | ☐ | Ease of Modelling ☐ | ☐ ☐ ☐ ☐ ☐ |
| 8 | Robustness | ☐ | Speed ☐ | ☐ ☐ ☐ ☐ ☐ |
| 9 | Robustness | ☐ | Ease of Modelling ☐ | ☐ ☐ ☐ ☐ ☐ |
| 10 | Speed | ☐ | Ease of Modelling ☐ | ☐ ☐ ☐ ☐ ☐ |

**Table G. 2 Type-2 Pairwise comparison of criteria with respect to criteria( For both prediction and classification methods)**

| | "Which criterion influences criterion **Predictive Accuracy** more? | | | How much more? 1 3 5 7 9 |
|---|---|---|---|---|
| 1 | Robustness | ☐ | Speed ☐ | ☐ ☐ ☐ ☐ ☐ |
| 2 | Robustness | ☐ | Ease of Modelling ☐ | ☐ ☐ ☐ ☐ ☐ |
| 3 | Speed | ☐ | Ease of Modelling ☐ | ☐ ☐ ☐ ☐ ☐ |

| | "Which criterion influences criterion **Ease of Use of the Model** more? | | | How much more? 1 3 5 7 9 |
|---|---|---|---|---|
| 1 | Robustness | ☐ | Speed ☐ | ☐ ☐ ☐ ☐ ☐ |
| 2 | Robustness | ☐ | Ease of Modelling ☐ | ☐ ☐ ☐ ☐ ☐ |
| 3 | Speed | ☐ | Ease of Modelling ☐ | ☐ ☐ ☐ ☐ ☐ |

**Table G. 2 (Continued) Type-2 Pairwise comparison of criteria with respect to criteria( For both prediction and classification methods)**

| "Which criterion influences criterion **Speed** more?" | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | 1 | 3 | 5 | 7 | 9 |
| 1 Predictive Accuracy | ☐ | Robustness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 Robustness | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| "Which criterion influences criterion **Ease of Modelling** more?" | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | 1 | 3 | 5 | 7 | 9 |
| 1 Ease of Use of the Model | ☐ | Robustness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 Ease of Use of the Model | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 Robustness | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G. 3 Type-3 Pairwise comparison of sub-criteria with respect to criteria (For classification methods)**

| # | Which subcriterion should be emphasized more for criterion **Predictive Accuracy**? | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Misclassification Rate | ☐ | Kappa | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Misclassification Rate | ☐ | CI | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Misclassification Rate | ☐ | Stability of PCC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Misclassification Rate | ☐ | Recall | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Misclassification Rate | ☐ | Precision | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Misclassification Rate | ☐ | AUROC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | Kappa | ☐ | CI | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | Kappa | ☐ | Stability of PCC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | Kappa | ☐ | Recall | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | Kappa | ☐ | Precision | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 11 | Kappa | ☐ | AUROC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 12 | CI | ☐ | Stability of PCC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 13 | CI | ☐ | Recall | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 14 | CI | ☐ | Precision | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 15 | CI | ☐ | AUROC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 16 | Stability of PCC | ☐ | Recall | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 17 | Stability of PCC | ☐ | Precision | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 18 | Stability of PCC | ☐ | AUROC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 19 | Recall | ☐ | Precision | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 20 | Recall | ☐ | AUROC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 21 | Precision | ☐ | AUROC | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G. 3 (Continued) Type-3 Pairwise comparison of sub-criteria with respect to criteria (For prediction and classification methods)**

| # | Which subcriterion should be emphasized more for criterion **Ease of Use of the Model** ? | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Interpretability | ☐ | Compactness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Interpretability | ☐ | Embaddability | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Compactness | ☐ | Embaddability | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| # | Which subcriterion should be emphasized more for criterion **Robustness**? | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | To categorical and continuous variables | ☐ | To complexity | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | To categorical and continuous variables | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | To categorical and continuous variables | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | To categorical and continuous variables | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | To complexity | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | To complexity | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | To complexity | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | To Noise in Data | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | To Noise in Data | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | To Irrelevant Variables | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| # | Which subcriterion should be emphasized more for criterion **Speed** ? | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Learning Curve Requirements | ☐ | Development speed/effort | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Learning Curve Requirements | ☐ | Response Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Development speed/effort | ☐ | Response Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| # | Which subcriterion should be emphasized more for criterion **Ease of Modelling**? | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Computing Resource | ☐ | Independence From Expert | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Computing Resource | ☐ | Scalability | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Computing Resource | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Independence From Expert | ☐ | Scalability | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Independence From Expert | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Scalability | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G. 4 Type-4 Pairwise comparison of sub-criteria with respect to sub-criteria criteria (For classification and prediction methods)**

| # | Which subcriterion influences subcriterion Computing Resources more? | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Compactness | ☐ | To categorical and continuous variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Compactness | ☐ | To complexity | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Compactness | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Compactness | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Compactness | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Compactness | ☐ | Independence From Expert | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | To categorical and continuous variables | ☐ | To complexity | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | To categorical and continuous variables | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | To categorical and continuous variables | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | To categorical and continuous variables | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 11 | To categorical and continuous variables | ☐ | Independence From Expert | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 12 | To complexity | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 13 | To complexity | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 14 | To complexity | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 15 | To complexity | ☐ | Independence From Expert | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 16 | To Noise in Data | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 17 | To Noise in Data | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 18 | To Noise in Data | ☐ | Independence From Expert | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 19 | To Irrelevant Variables | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 20 | To Irrelevant Variables | ☐ | Independence From Expert | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 21 | To Missing Values | ☐ | Independence From Expert | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G. 4 (Continued) Type-4 Pairwise comparison of sub-criteria with respect to sub-criteria criteria (For classification and prediction methods)**

| # | Which subcriterion influences subcriterion Independence from Experts more? | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Interpretability | ☐ | To categorical and continuous variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Interpretability | ☐ | To complexity | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Interpretability | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Interpretability | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Interpretability | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Interpretability | ☐ | Learning Curve Requirements | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | Interpretability | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | To categorical and continuous variables | ☐ | To complexity | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | To categorical and continuous variables | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | To categorical and continuous variables | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 11 | To categorical and continuous variables | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 12 | To categorical and continuous variables | ☐ | Learning Curve Requirements | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 13 | To categorical and continuous variables | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 14 | To complexity | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 15 | To complexity | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 16 | To complexity | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 17 | To complexity | ☐ | Learning Curve Requirements | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 18 | To complexity | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 19 | To Noise in Data | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 20 | To Noise in Data | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 21 | To Noise in Data | ☐ | Learning Curve Requirements | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 22 | To Noise in Data | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 23 | To Irrelevant Variables | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 24 | To Irrelevant Variables | ☐ | Learning Curve Requirements | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 25 | To Irrelevant Variables | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 26 | To Missing Values | ☐ | Learning Curve Requirements | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 27 | To Missing Values | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 28 | Learning Curve Requirements | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G. 4 (Continued) Type-4 Pairwise comparison of sub-criteria with respect to sub-criteria criteria (For classification and prediction methods)**

| # | Which subcriterion influences subcriterion Scalability more? | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | To complexity | ☐ | To Noise in Data | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | To complexity | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | To complexity | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | To complexity | ☐ | Development speed/effort | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | To complexity | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | To complexity | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | To Noise in Data | ☐ | To Irrelevant Variables | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | To Noise in Data | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | To Noise in Data | ☐ | Development speed/effort | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | To Noise in Data | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 11 | To Noise in Data | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 12 | To Irrelevant Variables | ☐ | To Missing Values | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 13 | To Irrelevant Variables | ☐ | Development speed/effort | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 14 | To Irrelevant Variables | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 15 | To Irrelevant Variables | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 16 | To Missing Values | ☐ | Development speed/effort | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 17 | To Missing Values | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 18 | To Missing Values | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 19 | Development speed/effort | ☐ | Computing Resource | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 20 | Development speed/effort | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 21 | Computing Resource | ☐ | Flexibility | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G. 4 (Continued) Type-4 Pairwise comparison of sub-criteria with respect to sub-criteria criteria (For classification and prediction methods)**

| # | Which subcriterion influences subcriterion Flexibility more? | | | | How much more? 1 3 5 7 9 |
|---|---|---|---|---|---|
| 1 | To categorical and continuous variables | ☐ | To complexity | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 2 | To categorical and continuous variables | ☐ | To Noise in Data | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 3 | To categorical and continuous variables | ☐ | To Irrelevant Variables | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 4 | To categorical and continuous variables | ☐ | To Missing Values | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 5 | To categorical and continuous variables | ☐ | Development speed/effort | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 6 | To categorical and continuous variables | ☐ | Computing Resource | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 7 | To categorical and continuous variables | ☐ | Scalability | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 8 | To complexity | ☐ | To Noise in Data | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 9 | To complexity | ☐ | To Irrelevant Variables | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 10 | To complexity | ☐ | To Missing Values | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 11 | To complexity | ☐ | Development speed/effort | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 12 | To complexity | ☐ | Computing Resource | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 13 | To complexity | ☐ | Scalability | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 14 | To Noise in Data | ☐ | To Irrelevant Variables | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 15 | To Noise in Data | ☐ | To Missing Values | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 16 | To Noise in Data | ☐ | Development speed/effort | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 17 | To Noise in Data | ☐ | Computing Resource | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 18 | To Noise in Data | ☐ | Scalability | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 19 | To Irrelevant Variables | ☐ | To Missing Values | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 20 | To Irrelevant Variables | ☐ | Development speed/effort | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 21 | To Irrelevant Variables | ☐ | Computing Resource | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 22 | To Irrelevant Variables | ☐ | Scalability | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 23 | To Missing Values | ☐ | Development speed/effort | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 24 | To Missing Values | ☐ | Computing Resource | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 25 | To Missing Values | ☐ | Scalability | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 26 | Development speed/effort | ☐ | Computing Resource | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 27 | Development speed/effort | ☐ | Scalability | ☐ | ☐ ☐ ☐ ☐ ☐ |
| 28 | Computing Resource | ☐ | Scalability | ☐ | ☐ ☐ ☐ ☐ ☐ |

**Table G. 5 Type-5 Pairwise comparison of criteria with respect to sub-criteria (feedback)**

| | Which criterion is influenced more from the subcriteria Misclassification Rate? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria Kappa? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria CI? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria Stabiity? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria Recall? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria Precision? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria AUROC? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria Interpretability? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Ease of Use of the Model | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria Compactness? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Ease of Use of the Model | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

| | Which criterion is influenced more from the subcriteria Robustness to Categorical and Continuous Data? | | | | How much more? | | | | |
|---|---|---|---|---|---|---|---|---|---|
| # | | | | | 1 | 3 | 5 | 7 | 9 |
| 1 | Predictive Accuracy | ☐ | Robustness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Robustness | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Robustness | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G. 5 (Continued) Type-5 Pairwise comparison of criteria with respect to sub-criteria (feedback)**

Which criterion is influenced more from the subcriteria Robustness to complexitiy?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Ease of Use of the Model | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Robustness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Ease of Use of the Model | ☐ | Robustness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Ease of Use of the Model | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 7 | Ease of Use of the Model | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 8 | Robustness | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 9 | Robustness | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 10 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Which criterion is influenced more from the subcriteria "Robustness to Noise in Data"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Robustness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Robustness | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Robustness | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Which criterion is influenced more from the subcriteria "Robustness to Irrelavent Data"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Robustness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Robustness | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Robustness | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G.5 (Continued) Type-5 Pairwise comparison of criteria with respect to sub-criteria (feedback)**

Which criterion is influenced more from the subcriteria "Robustness to Missing Values"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Robustness | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Robustness | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Robustness | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Which criterion is influenced more from the subcriteria "Learning Curve Requirements"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Which criterion is influenced more from the subcriteria "Development Speed"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Which criterion is influenced more from the subcriteria "Response Speed"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Ease of use of the Model | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Which criterion is influenced more from the subcriteria "Computing Resources"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Ease of use of the Model | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Ease of use of the Model | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Ease of use of the Model | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

**Table G.5 (Continued) Type-5 Pairwise comparison of criteria with respect to sub-criteria (feedback)**

Which criterion is influenced more from the subcriteria "Independence from Experts"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Ease of use of the Model | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Ease of use of the Model | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Ease of use of the Model | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Which criterion is influenced more from the subcriteria "Scalability"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Ease of use of the Model | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Ease of use of the Model | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Ease of use of the Model | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Which criterion is influenced more from the subcriteria "Flexibility"?

| # | | | | | How much more? 1 | 3 | 5 | 7 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Predictive Accuracy | ☐ | Ease of use of the Model | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 2 | Predictive Accuracy | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 3 | Predictive Accuracy | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 4 | Ease of use of the Model | ☐ | Speed | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 5 | Ease of use of the Model | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| 6 | Speed | ☐ | Ease of Modelling | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

# APPENDIX H

# SUPERMATRICES

Unweighted Supermatrix, Weighted Supermatrix and Limit Matrix for both classification and prediction methods are as follows:

| | | 1 | 2 | 3 | 4 | 5 | | 1.1. | 1.2. | 1.3. | 1.4. | 1.5. | 1.6. | 1.7. | 2.1. | 2.2. | 2.3. | 3.1. | 3.2. | 3.3. | 3.4. | 3.5. | 4.1. | 4.2. | 4.3. | 5.1. | 5.2. | 5.3. | 5.4. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction Accuracy | Ease of Use of the Model | Robustness | Speed | Ease of Modelling | Goal | Misclassification Rate | Kappa | CI | Stability | Recall | Precision | AUROC | Interpretability | Compactness | Embaddability | Robustness to categorical and continuous variables | Robustness to complexity | Robustness to Noise in Data | Robustness to Irrelevant Variables | Robustness to Missing Values | Learning Curve Requirements | Development speed/effort | Response Speed | Computing Resource | Independence From Expert | Scalability | Flexibility |
| 1 | Prediction Accuracy | 0.00000 | 0.00000 | 0.00000 | 0.24263 | 0.00000 | 0.49903 | 0.87500 | 0.87500 | 0.87500 | 0.87500 | 0.87500 | 0.87500 | 0.87500 | 0.00000 | 0.00000 | 0.00000 | 0.07520 | 0.05758 | 0.07809 | 0.07809 | 0.07809 | 0.10473 | 0.10473 | 0.00000 | 0.08637 | 0.05330 | 0.04891 | 0.04496 |
| 2 | Ease of Use of the Model | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.07194 | 0.06302 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.83333 | 0.83333 | 1.0 | 0.00000 | 0.07300 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.24998 | 0.09903 | 0.16850 | 0.13640 | 0.15251 |
| 3 | Robustness | 0.64912 | 0.40538 | 0.00000 | 0.08794 | 0.64909 | 0.25321 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.50828 | 0.46120 | 0.52224 | 0.52224 | 0.52224 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 4 | Speed | 0.07193 | 0.11397 | 0.00000 | 0.00000 | 0.27897 | 0.04294 | 0.12500 | 0.12500 | 0.12500 | 0.12500 | 0.12500 | 0.12500 | 0.12500 | 0.00000 | 0.00000 | 0.00000 | 0.15117 | 0.20411 | 0.19983 | 0.19983 | 0.19983 | 0.63699 | 0.63699 | 0.75002 | 0.43186 | 0.40203 | 0.25891 | 0.29195 |
| 5 | Ease of Modelling | 0.27895 | 0.48064 | 0.00000 | 0.66943 | 0.00000 | 0.14181 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.16667 | 0.16667 | 0.00000 | 0.26534 | 0.20411 | 0.19983 | 0.19983 | 0.19983 | 0.25828 | 0.25828 | 0.00000 | 0.38275 | 0.37616 | 0.55578 | 0.51058 |
| | Goal | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.1. | Misclassification Rate | 0.21718 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.13295 | 0.10046 | 0.10943 | 0.18221 | 0.16884 | 0.10109 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03304 | 0.04112 | 0.05367 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.2. | Kappa | 0.10903 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.11300 | 0.00000 | 0.06080 | 0.10309 | 0.13539 | 0.13577 | 0.08317 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03114 | 0.03442 | 0.05367 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.3. | CI | 0.03679 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04188 | 0.04125 | 0.00000 | 0.03321 | 0.05751 | 0.05733 | 0.12457 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01965 | 0.01257 | 0.02275 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.4. | Stability | 0.02925 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04188 | 0.04217 | 0.04365 | 0.00000 | 0.03695 | 0.03684 | 0.03751 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01789 | 0.01346 | 0.02185 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.5. | Recall | 0.36798 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.13256 | 0.13043 | 0.13540 | 0.06074 | 0.00000 | 0.16023 | 0.13970 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05249 | 0.04246 | 0.08580 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.6. | Precision | 0.17004 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.11588 | 0.10993 | 0.11274 | 0.05318 | 0.14571 | 0.00000 | 0.12198 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04470 | 0.03708 | 0.07416 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.7. | AUROC | 0.06973 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.11288 | 0.10704 | 0.09175 | 0.07674 | 0.09665 | 0.09636 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03214 | 0.01832 | 0.05928 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 2.1. | Interpretability | 0.00000 | 0.60000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05255 | 0.20993 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.10095 | 0.00000 | 0.00000 |
| 2.2. | Compactness | 0.00000 | 0.20000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.27327 | 0.00000 | 0.13875 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.06785 | 0.00000 | 0.00000 | 0.00000 |
| 2.3. | Embaddability | 0.00000 | 0.20000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04585 | 0.08341 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 3.1. | Robustness to categorical and continuous variables | 0.00000 | 0.00000 | 0.05376 | 0.00000 | 0.00000 | 0.00000 | 0.01790 | 0.01768 | 0.01799 | 0.01918 | 0.01908 | 0.01903 | 0.02190 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03147 | 0.02828 | 0.02307 | 0.04582 | 0.04195 | 0.00000 | 0.03697 |
| 3.2. | Robustness to complexity | 0.00000 | 0.00000 | 0.45598 | 0.00000 | 0.00000 | 0.00000 | 0.09943 | 0.09843 | 0.10380 | 0.12845 | 0.07711 | 0.07686 | 0.09644 | 0.00000 | 0.35538 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.16715 | 0.14785 | 0.14823 | 0.36240 | 0.28113 | 0.35810 | 0.18138 |
| 3.3. | Robustness to Noise in Data | 0.00000 | 0.00000 | 0.20972 | 0.00000 | 0.00000 | 0.00000 | 0.04457 | 0.04409 | 0.04740 | 0.08110 | 0.04452 | 0.04438 | 0.05053 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.08923 | 0.06310 | 0.06827 | 0.15202 | 0.10065 | 0.07246 | 0.08887 |
| 3.4. | Robustness to Irrelevant Variables | 0.00000 | 0.00000 | 0.09142 | 0.00000 | 0.00000 | 0.00000 | 0.02473 | 0.02448 | 0.02483 | 0.02764 | 0.02540 | 0.02533 | 0.02556 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05810 | 0.02723 | 0.02601 | 0.05514 | 0.04382 | 0.03501 | 0.06295 |
| 3.5. | Robustness to Missing Values | 0.00000 | 0.00000 | 0.18911 | 0.00000 | 0.00000 | 0.00000 | 0.04889 | 0.04836 | 0.05171 | 0.04679 | 0.03561 | 0.03551 | 0.02803 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.07168 | 0.06310 | 0.06827 | 0.15202 | 0.10065 | 0.06058 | 0.08887 |
| 4.1. | Learning Curve Requirements | 0.00000 | 0.00000 | 0.00000 | 0.10472 | 0.00000 | 0.00000 | 0.03107 | 0.02987 | 0.03112 | 0.04561 | 0.03391 | 0.03381 | 0.03839 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.06593 | 0.00000 | 0.00000 | 0.30062 | 0.00000 | 0.00000 |
| 4.2. | Development speed/effort | 0.00000 | 0.00000 | 0.00000 | 0.25828 | 0.00000 | 0.00000 | 0.02849 | 0.02825 | 0.02859 | 0.03242 | 0.03124 | 0.03116 | 0.04067 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.15521 | 0.00000 | 0.06028 | 0.00000 | 0.00000 | 0.13508 | 0.41757 |
| 4.3. | Response Speed | 0.00000 | 0.00000 | 0.00000 | 0.63700 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.30057 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 5.1. | Computing Resource | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05962 | 0.00000 | 0.01336 | 0.01328 | 0.01337 | 0.01232 | 0.01209 | 0.01207 | 0.01762 | 0.12202 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.07337 | 0.15195 | 0.00000 | 0.03023 | 0.29384 | 0.03388 |
| 5.2. | Independence From Expert | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.17534 | 0.00000 | 0.05363 | 0.05286 | 0.05509 | 0.10263 | 0.03092 | 0.03085 | 0.03110 | 0.48968 | 0.00000 | 0.06348 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.19611 | 0.15526 | 0.00000 | 0.16473 | 0.00000 | 0.00000 | 0.00000 |
| 5.3. | Scalability | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.38252 | 0.00000 | 0.01670 | 0.01669 | 0.01671 | 0.01559 | 0.01455 | 0.01453 | 0.01956 | 0.06918 | 0.15328 | 0.22379 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.07697 | 0.02542 | 0.00000 | 0.00000 | 0.00000 | 0.08951 |
| 5.4. | Flexibility | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.38252 | 0.00000 | 0.06315 | 0.06224 | 0.06459 | 0.05189 | 0.02117 | 0.02112 | 0.02218 | 0.00000 | 0.35538 | 0.06348 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.09949 | 0.05734 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |

**Figure H. 1 Unweighted Supermatrix for the Classification methods**

| | | 1 | 2 | 3 | 4 | 5 | | 1.1. | 1.2. | 1.3. | 1.4. | 1.5. | 1.6. | 1.7. | 2.1. | 2.2. | 2.3. | 3.1. | 3.2. | 3.3. | 3.4. | 3.5. | 4.1. | 4.2. | 4.3. | 5.1. | 5.2. | 5.3. | 5.4. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction Accuracy | Ease of Use of the Model | Robustness | Speed | Ease of Modelli | Goal | Misclassificatio n Rate | Kappa | CI | Stability | Recall | Precision | AUROC | Interpretability | Compactness | Embaddability | Robustness to categorical and continuous variables | Robustness to complexity | Robustness to Noise in Data | Robustness to Irrelevant Variables | Robustness to Missing Values | Learning Curve Requirements | Development speed/effort | Response Speed | Computing Resource | Independence From Expert | Scalability | Flexibility |
| 1 | Prediction Accuracy | 0.00000 | 0.00000 | 0.00000 | 0.12131 | 0.00000 | 0.49903 | 0.43750 | 0.43750 | 0.43750 | 0.43750 | 0.43750 | 0.43750 | 0.43750 | 0.00000 | 0.00000 | 0.00000 | 0.07520 | 0.05758 | 0.07809 | 0.07809 | 0.07809 | 0.05237 | 0.05237 | 0.00000 | 0.04319 | 0.02665 | 0.02445 | 0.02248 |
| 2 | Ease of Use of the Model | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03597 | 0.06302 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.41667 | 0.41667 | 0.50000 | 0.00000 | 0.07300 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.12499 | 0.04951 | 0.08425 | 0.06820 | 0.07625 |
| 3 | Robustness | 0.32456 | 0.20269 | 0.00000 | 0.04397 | 0.32454 | 0.25321 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.50828 | 0.46120 | 0.52224 | 0.52224 | 0.52224 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 4 | Speed | 0.03596 | 0.05699 | 0.00000 | 0.00000 | 0.13949 | 0.04294 | 0.06250 | 0.06250 | 0.06250 | 0.06250 | 0.06250 | 0.06250 | 0.06250 | 0.00000 | 0.00000 | 0.00000 | 0.15117 | 0.20411 | 0.19983 | 0.19983 | 0.19983 | 0.31849 | 0.31849 | 0.37501 | 0.21593 | 0.20102 | 0.12946 | 0.14597 |
| 5 | Ease of Modelling | 0.13948 | 0.24032 | 0.00000 | 0.33471 | 0.00000 | 0.14181 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.08333 | 0.08333 | 0.00000 | 0.26534 | 0.20411 | 0.19983 | 0.19983 | 0.19983 | 0.12914 | 0.12914 | 0.00000 | 0.19137 | 0.18808 | 0.27789 | 0.25529 |
| | Goal | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.1. | Misclassification Rate | 0.10859 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.06648 | 0.05023 | 0.05471 | 0.09110 | 0.08442 | 0.05055 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01652 | 0.02056 | 0.02683 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.2. | Kappa | 0.05452 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05650 | 0.00000 | 0.03040 | 0.05155 | 0.06769 | 0.06789 | 0.04158 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01557 | 0.01721 | 0.02683 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.3. | CI | 0.01840 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02094 | 0.02063 | 0.00000 | 0.01661 | 0.02875 | 0.02866 | 0.06229 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00982 | 0.00628 | 0.01138 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.4. | Stability | 0.01462 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02094 | 0.02109 | 0.02182 | 0.00000 | 0.01847 | 0.01842 | 0.01875 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00894 | 0.00673 | 0.01092 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.5. | Recall | 0.18399 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.06628 | 0.06522 | 0.06770 | 0.03037 | 0.00000 | 0.08011 | 0.06985 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02625 | 0.02123 | 0.04290 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.6. | Precision | 0.08502 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05794 | 0.05497 | 0.05637 | 0.02659 | 0.07285 | 0.00000 | 0.06099 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02235 | 0.01854 | 0.03708 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.7. | AUROC | 0.03487 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05644 | 0.05352 | 0.04588 | 0.03837 | 0.04833 | 0.04818 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01607 | 0.00916 | 0.02964 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 2.1. | Interpretability | 0.00000 | 0.30000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02627 | 0.10497 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05048 | 0.00000 | 0.00000 |
| 2.2. | Compactness | 0.00000 | 0.10000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.13663 | 0.00000 | 0.06938 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03393 | 0.00000 | 0.00000 | 0.00000 |
| 2.3. | Embaddability | 0.00000 | 0.10000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02293 | 0.04171 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 3.1. | Robustness to categorical and continuous variables | 0.00000 | 0.00000 | 0.05376 | 0.00000 | 0.00000 | 0.00000 | 0.00895 | 0.00884 | 0.00900 | 0.00959 | 0.00954 | 0.00951 | 0.01095 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01573 | 0.01414 | 0.01154 | 0.02291 | 0.02097 | 0.00000 | 0.01849 |
| 3.2. | Robustness to complexity | 0.00000 | 0.00000 | 0.45598 | 0.00000 | 0.00000 | 0.00000 | 0.04972 | 0.04921 | 0.05190 | 0.06422 | 0.03855 | 0.03843 | 0.04822 | 0.00000 | 0.17769 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.08357 | 0.07392 | 0.07411 | 0.18120 | 0.14057 | 0.17905 | 0.09069 |
| 3.3. | Robustness to Noise in Data | 0.00000 | 0.00000 | 0.20972 | 0.00000 | 0.00000 | 0.00000 | 0.02228 | 0.02204 | 0.02370 | 0.04055 | 0.02226 | 0.02219 | 0.02527 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04462 | 0.03155 | 0.03413 | 0.07601 | 0.05033 | 0.03623 | 0.04444 |
| 3.4. | Robustness to Irrelevant Variables | 0.00000 | 0.00000 | 0.09142 | 0.00000 | 0.00000 | 0.00000 | 0.01237 | 0.01224 | 0.01242 | 0.01382 | 0.01270 | 0.01266 | 0.01278 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02905 | 0.01362 | 0.01301 | 0.02757 | 0.02191 | 0.01750 | 0.03147 |
| 3.5. | Robustness to Missing Values | 0.00000 | 0.00000 | 0.18911 | 0.00000 | 0.00000 | 0.00000 | 0.02444 | 0.02418 | 0.02586 | 0.02340 | 0.01781 | 0.01775 | 0.01401 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03584 | 0.03155 | 0.03413 | 0.07601 | 0.05033 | 0.03029 | 0.04444 |
| 4.1. | Learning Curve Requirements | 0.00000 | 0.00000 | 0.00000 | 0.05236 | 0.00000 | 0.00000 | 0.01554 | 0.01494 | 0.01556 | 0.02280 | 0.01695 | 0.01691 | 0.01919 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03297 | 0.00000 | 0.00000 | 0.15031 | 0.00000 | 0.00000 |
| 4.2. | Development speed/effort | 0.00000 | 0.00000 | 0.00000 | 0.12914 | 0.00000 | 0.00000 | 0.01425 | 0.01412 | 0.01429 | 0.01621 | 0.01562 | 0.01558 | 0.02033 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.07761 | 0.00000 | 0.03014 | 0.00000 | 0.00000 | 0.06754 | 0.20879 |
| 4.3. | Response Speed | 0.00000 | 0.00000 | 0.00000 | 0.31850 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.15028 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 5.1. | Computing Resource | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02981 | 0.00000 | 0.00668 | 0.00664 | 0.00668 | 0.00616 | 0.00605 | 0.00603 | 0.00881 | 0.06101 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03669 | 0.07598 | 0.00000 | 0.01511 | 0.14692 | 0.01694 |
| 5.2. | Independence From Expert | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.08767 | 0.00000 | 0.02681 | 0.02643 | 0.02755 | 0.05131 | 0.01546 | 0.01542 | 0.01555 | 0.24484 | 0.00000 | 0.03174 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.09806 | 0.07763 | 0.00000 | 0.08237 | 0.00000 | 0.00000 | 0.00000 |
| 5.3. | Scalability | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.19126 | 0.00000 | 0.00835 | 0.00835 | 0.00835 | 0.00780 | 0.00728 | 0.00726 | 0.00978 | 0.03459 | 0.07664 | 0.11189 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03849 | 0.01271 | 0.00000 | 0.00000 | 0.00000 | 0.04475 |
| 5.4. | Flexibility | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.19126 | 0.00000 | 0.03157 | 0.03112 | 0.03229 | 0.02595 | 0.01058 | 0.01056 | 0.01109 | 0.00000 | 0.17769 | 0.03174 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04974 | 0.02867 | 0.00000 | 0.00000 | 0.02247 | 0.00000 |

**Figure H. 2 Weighted Supermatrix for the Classification methods**

| | | 1 | 2 | 3 | 4 | 5 | | 1.1. | 1.2. | 1.3. | 1.4. | 1.5. | 1.6. | 1.7. | 2.1. | 2.2. | 2.3. | 3.1. | 3.2. | 3.3. | 3.4. | 3.5. | 4.1. | 4.2. | 4.3. | 5.1. | 5.2. | 5.3. | 5.4. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction Accuracy | Ease of Use of the Model | Robustness | Speed | Ease of Modelling | Goal | Misclassification Rate | Kappa | CI | Stability | Recall | Precision | AUROC | Interpretability | Compactness | Embaddability | Robustness to categorical and continuous variables | Robustness to complexity | Robustness to Noise in Data | Robustness to Irrelevant Variables | Robustness to Missing Values | Learning Curve Requirements | Development speed/effort | Response Speed | Computing Resource | Independence From Expert | Scalability | Flexibility |
| 1 | Prediction Accuracy | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 | 0.05918 |
| 2 | Ease of Use of the Model | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 | 0.03268 |
| 3 | Robustness | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 | 0.18821 |
| 4 | Speed | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 | 0.11258 |
| 5 | Ease of Modelling | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 | 0.13006 |
| | Goal | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.1. | Misclassification Rate | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 | 0.01155 |
| 1.2. | Kappa | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 | 0.00772 |
| 1.3. | CI | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 | 0.00334 |
| 1.4. | Stability | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 | 0.00260 |
| 1.5. | Recall | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 | 0.01608 |
| 1.6. | Precision | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 | 0.00999 |
| 1.7. | AUROC | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 | 0.00613 |
| 2.1. | Interpretability | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 | 0.01134 |
| 2.2. | Compactness | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 | 0.00554 |
| 2.3. | Embaddability | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 | 0.00376 |
| 3.1. | Robustness to categorical and continuous variables | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 | 0.01290 |
| 3.2. | Robustness to complexity | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 | 0.10814 |
| 3.3. | Robustness to Noise in Data | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 | 0.04779 |
| 3.4. | Robustness to Irrelevant Variables | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 | 0.02134 |
| 3.5. | Robustness to Missing Values | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 | 0.04346 |
| 4.1. | Learning Curve Requirements | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 | 0.01067 |
| 4.2. | Development speed/effort | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 | 0.02562 |
| 4.3. | Response Speed | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 | 0.03642 |
| 5.1. | Computing Resource | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 | 0.01378 |
| 5.2. | Independence From Expert | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 | 0.01970 |
| 5.3. | Scalability | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 | 0.02937 |
| 5.4. | Flexibility | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 | 0.03008 |

**Figure H. 3 Limit Matrix for the Classification methods**

| | | 1 | 2 | 3 | 4 | 5 | | 1.1. | 1.2. | 1.3. | 2.1. | 2.2. | 2.3. | 3.1. | 3.2. | 3.3. | 3.4. | 3.5. | 4.1. | 4.2. | 4.3. | 5.1. | 5.2. | 5.3. | 5.4. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction Accuracy | Ease of Use of the Model | Robustness | Speed | Ease of Modelling | Goal | RMSE | Stability of RMSE | R Square | Interpretability | Compactness | Embaddability | Robustness to categorical and continuous variables | Robustness to complexity | Robustness to Noise in Data | Robustness to Irrelevant Variables | Robustness to Missing Values | Learning Curve Requirements | Development speed/effort | Response Speed | Computing Resource | Independence From Expert | Scalability | Flexibility |
| 1 | Prediction Accuracy | 0.00000 | 0.00000 | 0.00000 | 0.24263 | 0.00000 | 0.49903 | 0.87500 | 0.87500 | 0.87500 | 0.00000 | 0.00000 | 0.00000 | 0.07520 | 0.05758 | 0.07809 | 0.07809 | 0.07809 | 0.10473 | 0.10473 | 0.00000 | 0.08637 | 0.05330 | 0.04891 | 0.04496 |
| 2 | Ease of Use of the Model | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.07194 | 0.06302 | 0.00000 | 0.00000 | 0.00000 | 0.83333 | 0.83333 | 1.00 | 0.00000 | 0.07300 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.24998 | 0.09903 | 0.16850 | 0.13640 | 0.15251 |
| 3 | Robustness | 0.64912 | 0.40538 | 0.00000 | 0.08794 | 0.64909 | 0.25321 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.50828 | 0.46120 | 0.52224 | 0.52224 | 0.52224 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 4 | Speed | 0.07193 | 0.11397 | 0.00000 | 0.00000 | 0.27897 | 0.04294 | 0.12500 | 0.12500 | 0.12500 | 0.00000 | 0.00000 | 0.00000 | 0.15117 | 0.20411 | 0.19983 | 0.19983 | 0.19983 | 0.63699 | 0.63699 | 0.75002 | 0.43186 | 0.40203 | 0.25891 | 0.29195 |
| 5 | Ease of Modelling | 0.27895 | 0.48064 | 0.00000 | 0.66943 | 0.00000 | 0.14181 | 0.00000 | 0.00000 | 0.00000 | 0.16667 | 0.16667 | 0.00000 | 0.26534 | 0.20411 | 0.19983 | 0.19983 | 0.19983 | 0.25828 | 0.25828 | 0.00000 | 0.38275 | 0.37616 | 0.55578 | 0.51058 |
| | Goal | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.1. | RMSE | 0.48679 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.14886 | 0.25558 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02837 | 0.04489 | 0.06750 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.2. | Stability of RMSE | 0.07782 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.08166 | 0.00000 | 0.07708 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02526 | 0.01526 | 0.03220 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.3. | R Square | 0.43539 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.17620 | 0.12656 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02837 | 0.04280 | 0.06750 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 2.1. | Interpretability | 0.00000 | 0.60000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05255 | 0.20993 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.10095 | 0.00000 | 0.00000 |
| 2.2. | Compactness | 0.00000 | 0.20000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.27327 | 0.00000 | 0.13875 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.06785 | 0.00000 | 0.00000 | 0.00000 |
| 2.3. | Embaddability | 0.00000 | 0.20000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04585 | 0.08341 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 3.1. | Robustness to categorical and continuous variables | 0.00000 | 0.00000 | 0.05376 | 0.00000 | 0.00000 | 0.00000 | 0.02883 | 0.06247 | 0.02836 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03282 | 0.02888 | 0.03148 | 0.04582 | 0.04195 | 0.00000 | 0.03697 |
| 3.2. | Robustness to complexity | 0.00000 | 0.00000 | 0.45598 | 0.00000 | 0.00000 | 0.00000 | 0.17417 | 0.11716 | 0.13570 | 0.00000 | 0.35538 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.18360 | 0.16636 | 0.19353 | 0.36240 | 0.28113 | 0.35810 | 0.18138 |
| 3.3. | Robustness to Noise in Data | 0.00000 | 0.00000 | 0.20972 | 0.00000 | 0.00000 | 0.00000 | 0.05665 | 0.10180 | 0.05467 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.09846 | 0.06354 | 0.09247 | 0.15202 | 0.10065 | 0.07246 | 0.08887 |
| 3.4. | Robustness to Irrelevant Variables | 0.00000 | 0.00000 | 0.09142 | 0.00000 | 0.00000 | 0.00000 | 0.04260 | 0.04473 | 0.04008 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05213 | 0.02693 | 0.03659 | 0.05514 | 0.04382 | 0.03501 | 0.06295 |
| 3.5. | Robustness to Missing Values | 0.00000 | 0.00000 | 0.18911 | 0.00000 | 0.00000 | 0.00000 | 0.07822 | 0.06916 | 0.07359 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.09250 | 0.06354 | 0.09247 | 0.15202 | 0.10065 | 0.06058 | 0.08887 |
| 4.1. | Learning Curve Requirements | 0.00000 | 0.00000 | 0.00000 | 0.10472 | 0.00000 | 0.00000 | 0.05228 | 0.05669 | 0.05041 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.07187 | 0.00000 | 0.00000 | 0.30062 | 0.00000 | 0.00000 |
| 4.2. | Development speed/effort | 0.00000 | 0.00000 | 0.00000 | 0.25828 | 0.00000 | 0.00000 | 0.04725 | 0.04793 | 0.04740 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.21001 | 0.00000 | 0.07264 | 0.00000 | 0.00000 | 0.13508 | 0.41757 |
| 4.3. | Response Speed | 0.00000 | 0.00000 | 0.00000 | 0.63700 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.30057 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 5.1. | Computing Resource | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05962 | 0.00000 | 0.02045 | 0.01597 | 0.01950 | 0.12202 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.06953 | 0.19364 | 0.00000 | 0.03023 | 0.29384 | 0.03388 |
| 5.2. | Independence From Expert | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.17534 | 0.00000 | 0.10265 | 0.11400 | 0.08757 | 0.48968 | 0.00000 | 0.06348 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.24847 | 0.19621 | 0.00000 | 0.16473 | 0.00000 | 0.00000 | 0.00000 |
| 5.3. | Scalability | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.38252 | 0.00000 | 0.02366 | 0.01887 | 0.02258 | 0.06918 | 0.15328 | 0.22379 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.08877 | 0.03701 | 0.00000 | 0.00000 | 0.00000 | 0.08951 |
| 5.4. | Flexibility | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.38252 | 0.00000 | 0.11537 | 0.07580 | 0.10748 | 0.00000 | 0.35538 | 0.06348 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.12142 | 0.08296 | 0.00000 | 0.00000 | 0.04494 | 0.00000 |

**Figure H. 4 Unweighted Supermatrix for the Prediction methods**

| | | 1 | 2 | 3 | 4 | 5 | | 1.1. | 1.2. | 1.3. | 2.1. | 2.2. | 2.3. | 3.1. | 3.2. | 3.3. | 3.4. | 3.5. | 4.1. | 4.2. | 4.3. | 5.1. | 5.2. | 5.3. | 5.4. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction Accuracy | Ease of Use of the Model | Robustness | Speed | Ease of Modelling | Goal | RMSE | Stability of RMSE | R Square | Interpretability | Compactness | Embaddability | Robustness to categorical and continuous variables | Robustness to complexity | Robustness to Noise in Data | Robustness to Irrelevant Variables | Robustness to Missing Values | Learning Curve Requirements | Development speed/effort | Response Speed | Computing Resource | Independence From Expert | Scalability | Flexibility |
| 1 | Prediction Accuracy | 0.00000 | 0.00000 | 0.00000 | 0.12131 | 0.00000 | 0.49903 | 0.43750 | 0.43750 | 0.43750 | 0.00000 | 0.00000 | 0.00000 | 0.07520 | 0.05758 | 0.07809 | 0.07809 | 0.07809 | 0.05237 | 0.05237 | 0.00000 | 0.04319 | 0.02665 | 0.02445 | 0.02248 |
| 2 | Ease of Use of the Model | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03597 | 0.06302 | 0.00000 | 0.00000 | 0.00000 | 0.41667 | 0.41667 | 0.50000 | 0.00000 | 0.07300 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.12499 | 0.04951 | 0.08425 | 0.06820 | 0.07625 |
| 3 | Robustness | 0.32456 | 0.20269 | 0.00000 | 0.04397 | 0.32454 | 0.25321 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.50828 | 0.46120 | 0.52224 | 0.52224 | 0.52224 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 4 | Speed | 0.03596 | 0.05699 | 0.00000 | 0.00000 | 0.13949 | 0.04294 | 0.06250 | 0.06250 | 0.06250 | 0.00000 | 0.00000 | 0.00000 | 0.15117 | 0.20411 | 0.19983 | 0.19983 | 0.19983 | 0.31849 | 0.31849 | 0.37501 | 0.21593 | 0.20102 | 0.12946 | 0.14597 |
| 5 | Ease of Modelling | 0.13948 | 0.24032 | 0.00000 | 0.33471 | 0.00000 | 0.14181 | 0.00000 | 0.00000 | 0.00000 | 0.08333 | 0.08333 | 0.00000 | 0.26534 | 0.20411 | 0.19983 | 0.19983 | 0.19983 | 0.12914 | 0.12914 | 0.00000 | 0.19137 | 0.18808 | 0.27789 | 0.25529 |
| | Goal | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.1. | RMSE | 0.24339 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.07443 | 0.12779 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01419 | 0.02244 | 0.03375 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.2. | Stability of RMSE | 0.03891 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04083 | 0.00000 | 0.03854 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01263 | 0.00763 | 0.01610 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.3. | R Square | 0.21770 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.08810 | 0.06328 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01419 | 0.02140 | 0.03375 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 2.1. | Interpretability | 0.00000 | 0.30000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02627 | 0.10497 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.05048 | 0.00000 | 0.00000 |
| 2.2. | Compactness | 0.00000 | 0.10000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.13663 | 0.00000 | 0.06938 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03393 | 0.00000 | 0.00000 | 0.00000 |
| 2.3. | Embaddability | 0.00000 | 0.10000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02293 | 0.04171 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 3.1. | Robustness to categorical and continuous variables | 0.00000 | 0.00000 | 0.05376 | 0.00000 | 0.00000 | 0.00000 | 0.01442 | 0.03123 | 0.01418 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.01641 | 0.01444 | 0.01574 | 0.02291 | 0.02097 | 0.00000 | 0.01849 |
| 3.2. | Robustness to complexity | 0.00000 | 0.00000 | 0.45598 | 0.00000 | 0.00000 | 0.00000 | 0.08709 | 0.05858 | 0.06785 | 0.00000 | 0.17769 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.09180 | 0.08318 | 0.09677 | 0.18120 | 0.14057 | 0.17905 | 0.09069 |
| 3.3. | Robustness to Noise in Data | 0.00000 | 0.00000 | 0.20972 | 0.00000 | 0.00000 | 0.00000 | 0.02833 | 0.05090 | 0.02734 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04923 | 0.03177 | 0.04623 | 0.07601 | 0.05033 | 0.03623 | 0.04444 |
| 3.4. | Robustness to Irrelevant Variables | 0.00000 | 0.00000 | 0.09142 | 0.00000 | 0.00000 | 0.00000 | 0.02130 | 0.02236 | 0.02004 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02607 | 0.01347 | 0.01829 | 0.02757 | 0.02191 | 0.01750 | 0.03147 |
| 3.5. | Robustness to Missing Values | 0.00000 | 0.00000 | 0.18911 | 0.00000 | 0.00000 | 0.00000 | 0.03911 | 0.03458 | 0.03679 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04625 | 0.03177 | 0.04623 | 0.07601 | 0.05033 | 0.03029 | 0.04444 |
| 4.1 | Learning Curve Requirements | 0.00000 | 0.00000 | 0.00000 | 0.05236 | 0.00000 | 0.00000 | 0.02614 | 0.02834 | 0.02520 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03594 | 0.00000 | 0.00000 | 0.15031 | 0.00000 | 0.00000 |
| 4.2. | Development speed/effort | 0.00000 | 0.00000 | 0.00000 | 0.12914 | 0.00000 | 0.00000 | 0.02363 | 0.02397 | 0.02370 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.10501 | 0.00000 | 0.03632 | 0.00000 | 0.00000 | 0.06754 | 0.20879 |
| 4.3. | Response Speed | 0.00000 | 0.00000 | 0.00000 | 0.31850 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.15028 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 5.1. | Computing Resource | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.02981 | 0.00000 | 0.01023 | 0.00798 | 0.00975 | 0.06101 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.03476 | 0.09682 | 0.00000 | 0.01511 | 0.14692 | 0.01694 |
| 5.2. | Independence From Expert | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.08767 | 0.00000 | 0.05132 | 0.05700 | 0.04379 | 0.24484 | 0.00000 | 0.03174 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.12423 | 0.09810 | 0.00000 | 0.08237 | 0.00000 | 0.00000 | 0.00000 |
| 5.3. | Scalability | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.19126 | 0.00000 | 0.01183 | 0.00944 | 0.01129 | 0.03459 | 0.07664 | 0.11189 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.04439 | 0.01851 | 0.00000 | 0.00000 | 0.00000 | 0.04475 |
| 5.4. | Flexibility | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.19126 | 0.00000 | 0.05768 | 0.03790 | 0.05374 | 0.00000 | 0.17769 | 0.03174 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.06071 | 0.04148 | 0.00000 | 0.00000 | 0.02247 | 0.00000 |

**Figure H. 5 Weighted Supermatrix for the Prediction methods**

| | | 1 | 2 | 3 | 4 | 5 | | 1.1. | 1.2. | 1.3. | 2.1. | 2.2. | 2.3. | 3.1. | 3.2. | 3.3. | 3.4. | 3.5. | 4.1. | 4.2. | 4.3. | 5.1. | 5.2. | 5.3. | 5.4. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prediction Accuracy | Ease of Use of the Model | Robustness | Speed | Ease of Modelling | Goal | RMSE | Stability of RMSE | R Square | Interpretability | Compactness | Embaddability | Robustness to categorical and continuous variables | Robustness to complexity | Robustness to Noise in Data | Robustness to Irrelevant Variables | Robustness to Missing Values | Learning Curve Requirements | Development speed/effort | Response Speed | Computing Resource | Independence From Expert | Scalability | Flexibility |
| 1 | Prediction Accuracy | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 | 0.05081 |
| 2 | Ease of Use of the Model | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 | 0.03417 |
| 3 | Robustness | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 | 0.19115 |
| 4 | Speed | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 | 0.11536 |
| 5 | Ease of Modelling | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 | 0.13381 |
| | Goal | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| 1.1. | RMSE | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 | 0.01660 |
| 1.2. | Stability of RMSE | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 | 0.00417 |
| 1.3. | R Square | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 | 0.01479 |
| 2.1. | Interpretability | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 | 0.01192 |
| 2.2. | Compactness | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 | 0.00583 |
| 2.3. | Embaddability | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 | 0.00393 |
| 3.1. | Robustness to categorical and continuous variables | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 | 0.01342 |
| 3.2. | Robustness to complexity | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 | 0.11201 |
| 3.3. | Robustness to Noise in Data | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 | 0.04911 |
| 3.4. | Robustness to Irrelevant Variables | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 | 0.02201 |
| 3.5. | Robustness to Missing Values | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 | 0.04521 |
| 4.1. | Learning Curve Requirements | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 | 0.01122 |
| 4.2. | Development speed/effort | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 | 0.02714 |
| 4.3. | Response Speed | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 | 0.03733 |
| 5.1. | Computing Resource | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 | 0.01500 |
| 5.2. | Independence From Expert | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 | 0.02180 |
| 5.3. | Scalability | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 | 0.03065 |
| 5.4. | Flexibility | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 | 0.03255 |

**Figure H. 6 Limit Matrix for the Prediction methods**

# APPENDIX I

# RANOVA AND FISHER'S LSD TEST RESULTS

For classification measures there are two data sets available. RANOVA analysis is conducted to these data sets.

Since there are two data sets (Customer Satisfaction Data and Casting Data), at first, effects of the data sets are analyzed with the "test of between subjects". This test shows that for criteria "Precision", "F0.5", "F1", "Kappa", "Specificity" and "Stability" data set is not significant and for these criteria remaining analyses are conducted with the combination of these two data sets. Otherwise, analyses are conducted for each data set. In Table I.1 resulting p values and their interpretations are illustrated for each criterion.

**Table I. 1 Tests of Between-Subjects Effects**

| Comparison Measures | Tests of Between-Subjects Effects p-value | Interpretation of the test result |
|---|---|---|
| PCC (1-MCR) | **.013 < .05** | Two data sets are statistically different |
| PRECISION | .165 > .05 | Data sets are not statistically different |
| RECALL | **.032 < .05** | Two data sets are statistically different |
| $F_{0.5}$ | .995 > .05 | Data sets are not statistically different |
| $F_1$ | .186 > .05 | Data sets are not statistically different |
| $F_2$ | **.025 < .05** | Two data sets are statistically different |
| KAPPA | .144 > .05 | Data sets are not statistically different |
| SPECIFICITY | .198 > .05 | Data sets are not statistically different |
| STABILITY OF PCC | .983 > .05 | Data sets are not statistically different |
| AUC | **.036 < .05** | Two data sets are statistically different |

Then, according to results of the "Test of between Subject Effects", "Tests of Within-Subjects Effects" are conducted for each criterion. For PCC, recall, F2 and AUC two data sets are evaluated separately. For each criterion mean scores of the alternative methods are compared. For instance; according to criterion Percent of Correctly Classified (PCC) stated hypotheses are as follows:

**$H_0$:** $\mu_{DT} = \mu_{NN} = \mu_{MARS} = \mu_{LR} = \mu_{SVM} = \mu_{MTS} = \mu_{FCF}$

(Mean PCC scores of these classification methods are equal to each other)

**$H_1$:** At least one of them is different

Used data sets and resulting p-values for each criterion are illustrated in Table I.2 Since all of the p-values are less than 0.05, one can conclude that for all of these criteria at least one method's mean is different from others and it is worth to construct Fisher's LSD test to compare the alternative methods' performance according to each of these criteria. In another words the Fisher's LSD multiple comparison tests are conducted only for the measures which are found statistically different in the RANOVA test, and the results are illustrated in Table I.3 and I.4.

**Table I. 2 Tests of Within-Subjects Effects**

| Comparison Measures | DataSet | p-value |
|---|---|---|
| PCC | Casting data | .000 |
| | Customer Satisfaction Data | .002 |
| PRECISION* | Combination of Two Data Sets | .000 |
| RECALL | Casting Data | .000 |
| | Customer Satisfaction Data | .000 |
| F05 | Combination of Two Data Sets | .000 |
| F1 | Combination of Two Data Sets | .000 |
| F2 | Casting Data | .000 |
| | Customer Satisfaction Data | .002 |
| KAPPA | Combination of Two Data Sets | .000 |
| SPECIFICITY | Combination of Two Data Sets | .000 |
| STABILITY OF PCC | Combination of Two Data Sets | .000 |
| AUC | Casting Data | .001 |
| | Customer Satisfaction Data | .000 |

**Table I. 3 Fisher's LSD multiple comparison tests results for classification methods**

| Comparison Measures | Methods Compared | | P-value Customer Satisfaction Data | P-value Casting Data | P-value combination of two data sets | Customer Satisfaction Data Mean Difference | Casting Data Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| PCC (1- MCR) | Method 1 | Method 2 | | | | | | |
| | DT | NN | .667 | .095 | - | -.009 | .029 | - |
| | DT | MARS | .187 | .279 | - | .047 | .050 | - |
| | DT | LR | .212 | .073 | - | -.047 | .069 | - |
| | DT | SVM* | .497 | **.017** | - | .021 | -.138* | - |
| | DT* | MTS | .998 | **.000** | - | 3.333E-5 | .281* | - |
| | DT | FCF* | **.011** | **.047** | - | -.081* | -.080* | - |
| | NN | MARS | .096 | .607 | - | .056 | .021 | - |
| | NN | LR | .449 | .174 | - | -.038 | .040 | - |
| | NN | SVM* | .423 | **.024** | - | .030 | -.167* | - |
| | NN* | MTS | .773 | **.001** | - | .009 | .252* | - |
| | NN | FCF* | .085 | **.030** | - | -.073 | -.109* | - |
| | MARS | LR | .124 | .346 | - | -.094 | .019 | - |
| | MARS | SVM* | .226 | **.022** | - | -.026 | -.188* | - |
| | MARS * | MTS | .173 | **.028** | - | -.047 | .231* | - |
| | MARS | FCF* | **.029** | **.015** | - | -.128* | -.130* | - |
| | LR | SVM* | .119 | **.011** | - | .068 | -.207* | - |
| | LR* | MTS | .093 | **.013** | - | .047 | .212* | - |
| | LR | FCF* | .208 | **.001** | - | -.034 | -.149* | - |
| | SVM* | MTS | .338 | **.003** | - | -.021 | .419* | - |
| | SVM | FCF | **.034** | .083 | - | -.103* | .058 | - |
| | MTS | FCF* | **.003** | **.004** | - | -.081* | -.361* | - |

**Table I. 3 (Continued) Fisher's LSD multiple comparison tests results for classification methods**

| Comparison Measures | Methods Compared | | P-value Customer Satisfaction Data | P-value Casting Data | P-value combination of two data sets | Customer Satisfaction Data Mean Difference | Casting Data Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| PRECISION | DT | NN | - | - | .439 | - | - | .038 |
| | DT* | MARS | - | - | **.021** | - | - | .153* |
| | DT | LR | - | - | .200 | - | - | .084 |
| | DT | SVM | - | - | .095 | - | - | -.140 |
| | DT | MTS | - | - | .476 | - | - | .052 |
| | DT | FCF* | - | - | **.020** | - | - | -.163* |
| | NN* | MARS | - | - | **.020** | - | - | .116* |
| | NN | LR | - | - | .503 | - | - | .046 |
| | NN | SVM | - | - | .054 | - | - | -.178 |
| | NN | MTS | - | - | .896 | - | - | .014 |
| | NN | FCF* | - | - | **.013** | - | - | -.201* |
| | MARS | LR | - | - | .128 | - | - | -.069 |
| | MARS | SVM* | - | - | **.005** | - | - | -.294* |
| | MARS | MTS | - | - | .324 | - | - | -.101 |
| | MARS | FCF* | - | - | **.000** | - | - | -.317* |
| | LR | SVM* | - | - | **.030** | - | - | -.224* |
| | LR | MTS | - | - | .672 | - | - | -.032 |
| | LR | FCF* | - | - | .001 | - | - | -.247* |
| | SVM | MTS | - | - | .159 | - | - | .192 |
| | SVM | FCF | - | - | .674 | - | - | -.023 |
| | MTS | FCF | - | - | .064 | - | - | -.215 |
| RECALL | DT | NN | .438 | .184 | - | .083 | .133 | |
| | DT | MARS | .038 | .423 | - | .069* | .111 | |
| | DT | LR | .257 | .199 | - | -.153 | .111 | |
| | DT | SVM* | .111 | **.011** | - | .306 | -.422* | - |
| | DT | MTS* | .070 | **.020** | - | .278 | -.267* | - |
| | DT | FCF | .059 | NA | - | -.250 | -.400 | - |
| | NN | MARS | .902 | .840 | - | -.014 | -.022 | - |
| | NN | LR | .245 | .868 | - | -.236 | -.022 | - |
| | NN | SVM* | .103 | **.011** | - | .222 | -.556* | - |

**Table I. 3 (Continued) Fisher's LSD multiple comparison tests results for classification methods**

| Comparison Measures | Methods Compared | | P-value Customer Satisfaction Data | P-value Casting Data | P-value combination of two data sets | Customer Satisfaction Data Mean Difference | Casting Data Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| RECALL | NN | MTS* | .060 | **.009** | - | .194 | -.400* | - |
| | NN | FCF* | .120 | **.015** | - | -.333 | -.533* | - |
| | MARS | LR | .135 | 1.000 | - | -.222 | .000 | - |
| | MARS | SVM | .185 | .062 | - | .236 | -.533 | - |
| | MARS | MTS* | .138 | **.042** | - | .208 | -.378* | - |
| | MARS | FCF* | .029 | **.044** | - | -.319* | -.511* | - |
| | LR | SVM* | .049 | **.035** | - | .458* | -.533* | - |
| | LR | MTS* | .050 | **.042** | - | .431 | -.378* | - |
| | LR | FCF* | .118 | **.013** | - | -.097 | -.511* | - |
| | SVM | MTS | .529 | .118 | - | -.028 | .156 | - |
| | SVM | FCF | .036 | .666 | - | -.556* | .022 | - |
| | MTS | FCF | .029 | .074 | - | -.528* | -.133 | - |
| F05 | DT | NN | - | - | .379 | - | - | .026 |
| | DT* | MARS | - | - | **.015** | - | - | .095* |
| | DT | LR | - | - | .738 | - | - | .010 |
| | DT | SVM* | - | - | **.034** | - | - | -.175* |
| | DT* | MTS | - | - | **.002** | - | - | .143* |
| | DT | FCF* | - | - | **.002** | - | - | -.202* |
| | NN* | MARS | - | - | **.021** | - | - | .069* |
| | NN | LR | - | - | .732 | - | - | -.016 |
| | NN | SVM* | - | - | **.009** | - | - | -.200* |
| | NN* | MTS | - | - | **.015** | - | - | .117* |
| | NN | FCF* | - | - | **.002** | - | - | -.228* |
| | MARS | LR* | - | - | **.034** | - | - | -.085* |
| | MARS | SVM* | - | - | **.005** | - | - | -.269* |
| | MARS | MTS | - | - | .185 | - | - | .049 |
| | MARS | FCF* | - | - | **.001** | - | - | -.297* |
| | LR | SVM* | - | - | **.051** | - | - | -.185 |
| | LR* | MTS | - | - | **.027** | - | - | .133* |
| | LR | FCF* | - | - | **.006** | - | - | -.213* |

**Table I. 3 (Continued) Fisher's LSD multiple comparison tests results for classification methods**

| Comparison Measures | Methods Compared | | P-value Customer Satisfaction Data | P-value Casting Data | P-value combination of two data sets | Customer Satisfaction Data Mean Difference | Casting Data Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| F05 | SVM* | MTS | - | - | **.001** | - | - | .318* |
| | SVM | FCF | - | - | .580 | - | - | -.028 |
| | MTS | FCF* | - | - | **.000** | - | - | -.346* |
| F01 | DT | NN | - | - | .451 | - | - | .023 |
| | DT | MARS | - | - | .092 | - | - | .055 |
| | DT | LR | - | - | .613 | - | - | -.022 |
| | DT | SVM | - | - | .056 | - | - | -.152 |
| | DT* | MTS | - | - | **.019** | - | - | .135* |
| | DT | FCF* | - | - | **.000** | - | - | -.248* |
| | NN | MARS | - | - | .221 | - | - | .031 |
| | NN | LR | - | - | .407 | - | - | -.046 |
| | NN | SVM* | - | - | **.017** | - | - | -.176* |
| | NN* | MTS | - | - | **.007** | - | - | .111* |
| | NN | FCF* | - | - | **.002** | - | - | -.271* |
| | MARS | LR* | - | - | **.051** | - | - | -.077 |
| | MARS | SVM* | - | - | **.015** | - | - | -.207* |
| | MARS | MTS | - | - | .052 | - | - | .080 |
| | MARS | FCF* | - | - | **.000** | - | - | -.302* |
| | LR | SVM | - | - | .123 | - | - | -.130 |
| | LR* | MTS | - | - | **.035** | - | - | .157* |
| | LR | FCF* | - | - | **.004** | - | - | -.225* |
| | SVM* | MTS | - | - | **.000** | - | - | .287* |
| | SVM | FCF | - | - | .145 | - | - | -.095 |
| | MTS | FCF* | - | - | **.000** | - | - | -.382* |
| | NN* | MTS | - | - | **.007** | - | - | .111* |
| | NN | FCF* | - | - | **.002** | - | - | -.271* |
| | MARS | LR* | - | - | **.051** | - | - | -.077 |
| | MARS | SVM* | - | - | **.015** | - | - | -.207* |
| | MARS | MTS | - | - | .052 | - | - | .080 |
| | MARS | FCF* | - | - | **.000** | - | - | -.302* |

**Table I. 3 (Continued) Fisher's LSD multiple comparison tests results for classification methods**

| Comparison Measures | Methods Compared | | P-value Customer Satisfaction Data | P-value Casting Data | P-value combination of two data sets | Customer Satisfaction Data Mean Difference | Casting Data Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| F01 | LR | SVM | - | - | .123 | - | - | -.130 |
| | LR* | MTS | - | - | **.035** | - | - | .157* |
| | LR | FCF* | - | - | **.004** | - | - | -.225* |
| | SVM* | MTS | - | - | **.000** | - | - | .287* |
| | SVM | FCF | - | - | .145 | - | - | -.095 |
| | MTS | FCF* | - | - | **.000** | - | - | -.382* |
| F02 | DT | NN | .522 | .956 | - | .039 | -.003 | - |
| | DT | MARS | .407 | .948 | - | .038 | .004 | - |
| | DT | LR | .268 | .520 | - | -.135 | .038 | - |
| | DT | SVM* | .273 | **.010** | - | .166 | -.433* | - |
| | DT | MTS | .155 | .272 | - | .188 | -.045 | - |
| | DT | FCF* | **.042** | **.000** | - | -.220* | -.362* | - |
| | NN | MARS | .988 | .602 | - | .000 | .007 | - |
| | NN | LR | .247 | .220 | - | -.174 | .041 | - |
| | NN | SVM* | .202 | **.021** | - | .127 | -.430* | - |
| | NN | MTS | .069 | .245 | - | .149 | -.041 | - |
| | NN | FCF* | .075 | **.027** | - | -.259 | -.359* | - |
| | MARS | LR | .117 | .105 | - | -.173 | .034 | - |
| | MARS | SVM* | .260 | **.024** | - | .128 | -.437* | - |
| | MARS | MTS | .124 | .220 | - | .150 | -.048 | - |
| | MARS | FCF* | **.016** | **.025** | - | -.259* | -.366* | - |
| | LR | SVM* | .128 | **.023** | - | .301 | -.471* | - |
| | LR | MTS | .093 | .115 | - | .323 | -.082 | - |
| | LR | FCF* | .182 | **.018** | - | -.086 | -.400* | - |
| | SVM* | MTS | .493 | **.012** | - | .022 | .388* | - |
| | SVM | FCF | .073 | .258 | - | -.386 | .071 | - |
| | MTS | FCF* | **.045** | **.013** | - | -.409* | -.317* | - |

**Table I. 3 (Continued) Fisher's LSD multiple comparison tests results for classification methods**

| Comparison Measures | Methods Compared | | P-value Customer Satisfaction Data | P-value Casting Data | P-value combination of two data sets | Customer Satisfaction Data Mean Difference | Casting Data Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| KAPPA | DT | NN | - | - | .078 | - | - | .097 |
| | DT* | MARS | - | - | **.013** | - | - | .149* |
| | DT | LR | - | - | .609 | - | - | .031 |
| | DT | SVM* | - | - | **.049** | - | - | -.156* |
| | DT* | MTS | - | - | **.002** | - | - | .189* |
| | DT | FCF* | - | - | **.000** | - | - | -.292* |
| | NN | MARS | - | - | .315 | - | - | .052 |
| | NN | LR | - | - | .482 | - | - | -.066 |
| | NN | SVM* | - | - | **.012** | - | - | -.253* |
| | NN | MTS | - | - | .076 | - | - | .092 |
| | NN | FCF* | - | - | **.002** | - | - | -.389* |
| | MARS | LR | - | - | .083 | - | - | -.118 |
| | MARS | SVM* | - | - | **.009** | - | - | -.305* |
| | MARS | MTS | - | - | .409 | - | - | .040 |
| | MARS | FCF* | - | - | **.000** | - | - | -.441* |
| | LR | SVM* | - | - | **.048** | - | - | -.186* |
| | LR | MTS | - | - | .053 | - | - | .158 |
| | LR | FCF* | - | - | **.001** | - | - | -.323* |
| | SVM* | MTS | - | - | **.000** | - | - | .345* |
| | SVM | FCF* | - | - | **.045** | - | - | -.136* |
| | MTS | FCF* | - | - | **.000** | - | - | -.481* |
| SPECIFICITY | DT | NN | - | - | .176 | - | - | -.020 |
| | DT | MARS | - | - | .329 | - | - | .038 |
| | DT | LR | - | - | .063 | - | - | .030 |
| | DT | SVM* | - | - | **.003** | - | - | -.093* |
| | DT* | MTS | - | - | **.001** | - | - | .132* |
| | DT | FCF | - | - | .436 | - | - | -.012 |
| | NN | MARS | - | - | .193 | - | - | .058 |
| | NN* | LR | - | - | **.004** | - | - | .051* |
| | NN | SVM* | - | - | **.004** | - | - | -.073* |
| | NN* | MTS | - | - | **.000** | - | - | .152* |

**Table I. 3 (Continued) Fisher's LSD multiple comparison tests results for classification methods**

| Comparison Measures | Methods Compared | | P-value Customer Satisfaction Data | P-value Casting Data | P-value combination of two data sets | Customer Satisfaction Data Mean Difference | Casting Data Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| SPECIFICITY | NN | FCF | - | - | .574 | - | - | .009 |
| | MARS | LR | - | - | .846 | - | - | -.007 |
| | MARS | SVM* | - | - | **.020** | - | - | -.131* |
| | MARS | MTS | - | - | .102 | - | - | .094 |
| | MARS | FCF | - | - | .147 | - | - | -.049 |
| | LR | SVM* | - | - | **.000** | - | - | -.124* |
| | LR* | MTS | - | - | **.002** | - | - | .101* |
| | LR | FCF* | - | - | **.023** | - | - | -.042* |
| | SVM* | MTS | - | - | **.000** | - | - | .225* |
| | SVM* | FCF | - | - | **.011** | - | - | .082* |
| | MTS | FCF* | - | - | **.004** | - | - | -.143* |
| STABILITY_PCC | DT | NN | - | - | .677 | - | - | .006 |
| | DT | MARS | - | - | .100 | - | - | -.039 |
| | DT | LR | - | - | .287 | - | - | -.010 |
| | DT* | SVM | - | - | **.020** | - | - | .026* |
| | DT | MTS* | - | - | **.000** | - | - | -.096* |
| | DT* | FCF | - | - | **.020** | - | - | .018* |
| | NN | MARS | - | - | .140 | - | - | -.045 |
| | NN | LR | - | - | .363 | - | - | -.016 |
| | NN | SVM | - | - | .183 | - | - | .020 |
| | NN | MTS* | - | - | **.001** | - | - | -.101* |
| | NN | FCF | - | - | .379 | - | - | .012 |
| | MARS | LR | - | - | .279 | - | - | .029 |
| | MARS* | SVM | - | - | **.020** | - | - | .065* |
| | MARS | MTS* | - | - | **.049** | - | - | -.056* |
| | MARS* | FCF | - | - | **.018** | - | - | .057* |
| | LR | SVM | - | - | .060 | - | - | .010 |

**Table I. 3 (Continued) Fisher's LSD multiple comparison tests results for classification methods**

| Comparison Measures | Methods Compared | | P-value Customer Satisfaction Data | P-value Casting Data | P-value combination of two data sets | Customer Satisfaction Data Mean Difference | Casting Data Mean Difference | Mean Difference |
|---|---|---|---|---|---|---|---|---|
| STABILITY_PCC | LR | MTS* | - | - | **.001** | - | - | .036 |
| | LR | FCF | - | - | .057 | - | - | -.086* |
| | SVM | MTS* | - | - | **.000** | - | - | .028 |
| | SVM | FCF | - | - | .254 | - | - | -.122* |
| | MTS* | FCF | - | - | **.000** | - | - | -.008 |
| AUC | DT | NN | .053 | .082 | - | -.103 | .051 | - |
| | DT* | MARS | .876 | **.030** | - | .003 | .197* | - |
| | DT* | LR | .306 | **.001** | - | -.055 | .205* | - |
| | DT | SVM | .103 | .084 | - | .090 | -.104 | - |
| | DT* | MTS | .057 | **.003** | - | .124 | .314* | - |
| | DT | FCF | **.025** | .633 | - | -.110* | .010 | - |
| | NN | MARS | .102 | .092 | - | .105 | .146 | - |
| | NN* | LR | .484 | **.012** | - | .047 | .153* | - |
| | NN | SVM* | **.005** | **.032** | - | .192* | -.155* | - |
| | NN* | MTS | **.015** | **.015** | - | .227* | .262* | - |
| | NN | FCF* | .869 | **.011** | - | -.007 | -.041* | - |
| | MARS | LR | .174 | .839 | - | -.058 | .007 | - |
| | MARS | SVM* | .154 | **.045** | - | .087 | -.301* | - |
| | MARS* | MTS | .061 | **.027** | - | .122 | .116* | - |
| | MARS | FCF | **.002** | .068 | - | -.112* | -.187 | - |
| | LR | SVM* | .104 | **.014** | - | .145 | -.308* | - |
| | LR* | MTS | **.035** | **.020** | - | .180* | .109* | - |
| | LR | FCF* | .147 | **.011** | - | -.054 | -.194* | - |
| | SVM* | MTS | .186 | **.012** | - | .035 | .418* | - |
| | SVM* | FCF | **.036** | **.045** | - | -.200* | .114* | - |
| | MTS | FCF* | **.016** | **.013** | - | -.234* | -.304* | - |

*. The mean difference is significant at the .05 level.

Test results given in Table I.3 are evaluated to find out the thresholds to prefer one alternative method to other with respect to specified criterion. For each criterion, mean differences significant at the 0.05 level are taken into consideration. Average, minimum and median of the significant mean differences are calculated. For instance, for PCC (1-MCR) average of the significant mean differences is about 0.184, median of the significant mean differences is 0.158 and minimum of the significant mean differences is 0.08. Determined preference threshold is 0.01 for MCR and this value is slightly above the minimum significant mean difference 0.08. Determined indifference threshold is 0.05 and seems suitable with respect to minimum of the significant mean differences that is 0.08.

For Kappa, average of the significant mean differences is about 0.28, median of the significant mean differences is 0.29 and minimum of the significant mean differences is 0.136. In this study for Kappa, determined preference threshold is 0.2 and indifference threshold is 0.1. Test results do not contradict these threshold selections.

Same analyses are also conducted for prediction methods. Since for prediction methods there is only ona data set, there is no need to conduct "test of between subjects" to analyse the effects of the data sets.

"Tests of Within-Subjects Effects" are conducted for each prediction criterion. For each criterion mean scores of the alternative methods are compared. For instance; according to criterion Mean Absolute Error (MAE), stated hypotheses are as follows:

**$H_0$:** $\mu_{DT} = \mu_{NN} = \mu_{MARS} = \mu_{MLR} = \mu_{FR} = \mu_{RR}$

(Mean MAE scores of these pretiction methods are equal to each other)

**$H_1$:** At least one of them is different

Used data sets and resulting p-values for each prediction criterion are illustrated in Table I.4. p-values of R and Stability of MSE are less than 0.05, and for these criteria at least one method's mean is different from the others and it is worth to construct

Fisher's LSD test to compare the alternative methods performance according to these two criteria.

**Table I. 4 Tests of Within-Subjects Effects**

| Comparison Measures | Data Set | P-value |
|---|---|---|
| MAE | Casting Data | 0.166 |
| MSE | | 0.357 |
| RMSE | | 0.244 |
| R | | **0.005** |
| R2 | | 0.088 |
| ADJR2 | | 0.366 |
| PWI-1 | | 0.400 |
| PWI-2 | | 0.374 |
| STABILITY_MSE | | **0.052** |
| STABILITY_RMSE | | 0.087 |

The Fisher's LSD multiple comparison tests are conducted only for these two criteria which are found statistically different in the RANOVA test, and the results are illustrated in Table I.5. None of the selected comparison criteria for prediction statistically different according to RANOVA test and thus Fisher's LSD multiple comparison test results are not available for them (RMSE, Stability of RMSE and R Square).

**Table I. 5 Fisher's LSD multiple comparison tests results for prediction methods**

| Comparison Measures | Methods Compared | | P-value | Mean Difference |
|---|---|---|---|---|
| | Method-1 | Method-2 | | |
| R | DT | NN | .685 | .044 |
| | DT | MARS | .406 | .105 |
| | DT | MLR | .611 | -.087 |
| | DT | MLR_LOGIT | .132 | .411 |
| | DT | HUBER-M | .101 | .435 |
| | DT | FF | .418 | -.172 |
| | NN | MARS | .513 | .061 |
| | NN | MLR | .324 | -.131 |
| | NN | MLR_LOGIT | .108 | .367 |
| | NN | HUBER-M | .088 | .391 |
| | NN | FF | .210 | -.216 |
| | MARS | MLR | .053 | -.193 |
| | MARS* | MLR_LOGIT | .044 | .306* |
| | MARS* | HUBER-M | .024 | .329* |
| | MARS | FF | .058 | -.277 |
| | MLR* | MLR_LOGIT | .004 | .499* |
| | MLR* | HUBER-M | .004 | .522* |
| | MLR | FF | .069 | -.085 |
| | MLR_LOGIT | HUBER-M | .340 | .023 |
| | MLR_LOGIT | FF* | .003 | -.583* |
| | HUBER-M | FF* | .005 | -.607* |
| STABILITY_MSE | DT | NN | .447 | -.369 |
| | DT | MARS | .671 | .134 |
| | DT | MLR | .633 | -.199 |
| | DT | MLR_LOGIT | .216 | -.488 |
| | DT | HUBER-M | .429 | -.296 |
| | DT | FF | .105 | -.868 |
| | NN | MARS | .102 | .503 |
| | NN | MLR | .065 | .170 |
| | NN | MLR_LOGIT | .434 | -.119 |
| | NN | HUBER-M | .511 | .073 |
| | NN* | FF | .033 | -.499* |
| | MARS | MLR | .123 | -.333 |
| | MARS* | MLR_LOGIT | .041 | -.623* |
| | MARS | HUBER-M | .060 | -.431 |
| | MARS* | FF | .020 | -1.003* |
| | MLR | MLR_LOGIT | .098 | -.289 |
| | MLR | HUBER-M | .240 | -.097 |
| | MLR* | FF | .013 | -.669* |
| | MLR_LOGIT | HUBER-M* | .043 | .192* |
| | MLR_LOGIT* | FF | .007 | -.380* |
| | HUBER-M* | FF | .004 | -.572* |

# APPENDIX J

# SENSITIVITY ANALYSES

# FOR SUB-CRITERIA OF THE CLASSIFICATION METHODS

For classification methods, there are 22 sub-criteria in the sub-criteria set I, and we change the each sub-criterion weight ($W$) in the range of [0, 1].

$$Wi_{difference} = Wi_{old} - Wi_{new} \qquad for\ each\ sub\ criterion\ i$$

$$Wj_{new} = Wi_{difference} \frac{Wj_{old}}{\sum_{j \neq i} Wj} \qquad i, j \in \mathrm{I} \quad for\ each\ sub\ criterion\ j \neq i$$

Then according to each generated weight set, net flows are calculated and graphs are conducted.

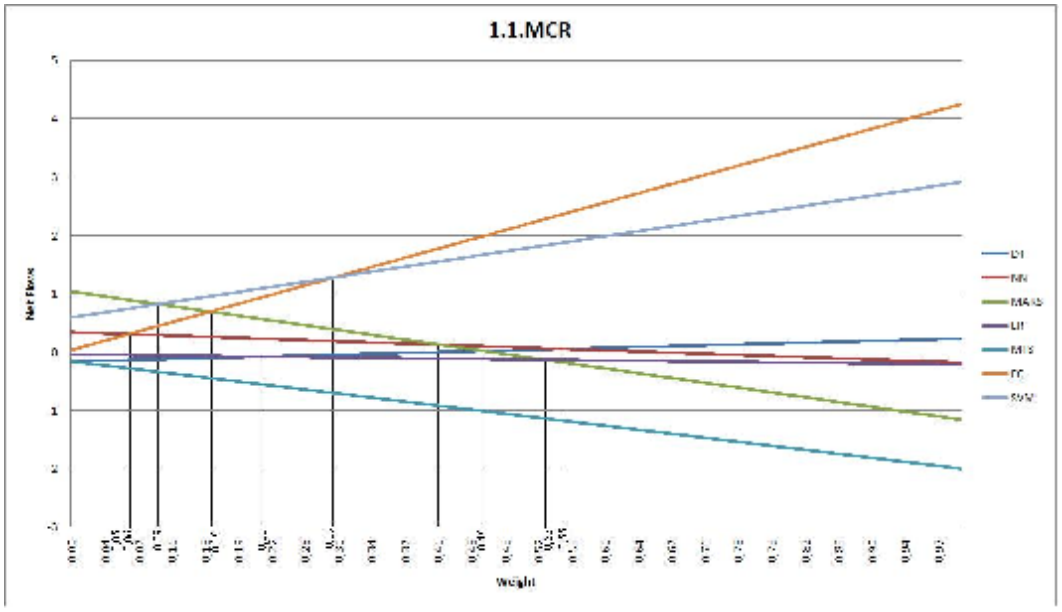



**Figure J. 1 Sensitivity of the net flows with respect to change in the weight of the MCR**

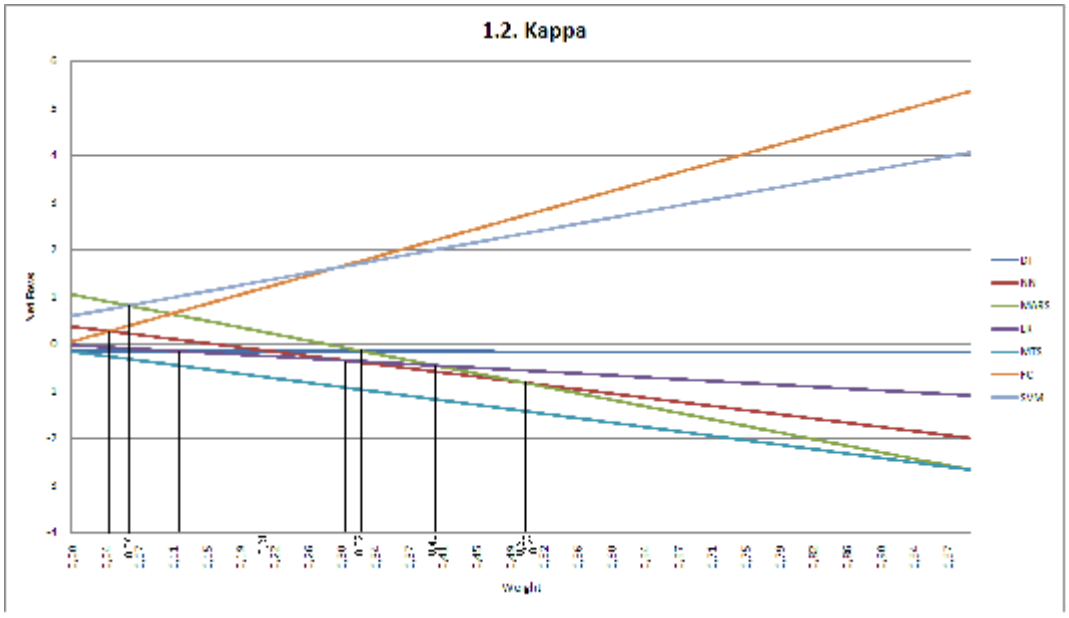



**Figure J. 2 Sensitivity of the net flows with respect to change in the weight of the Kappa**

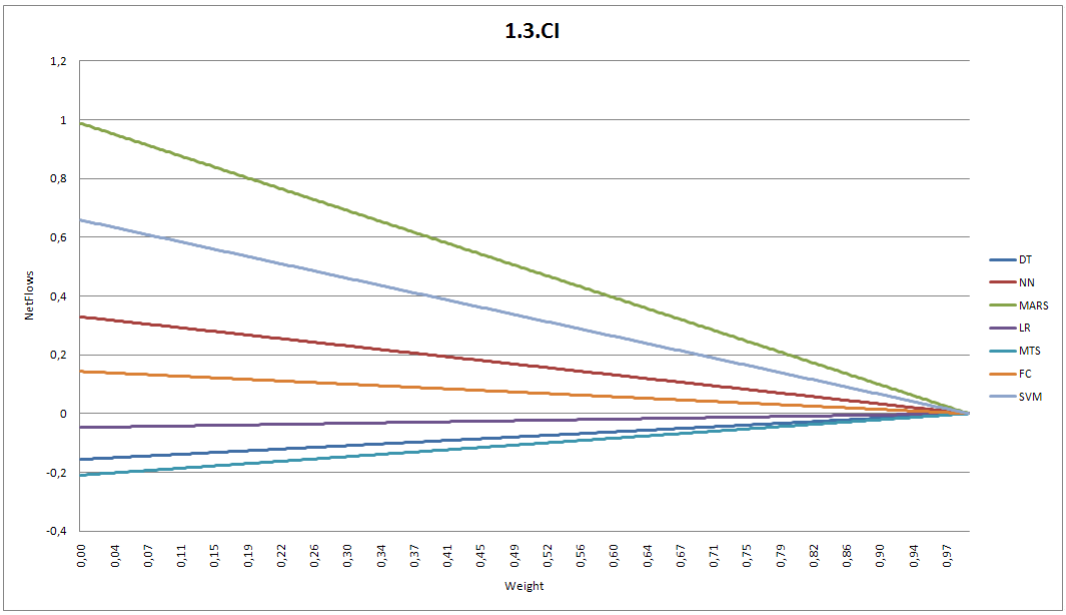



**Figure J. 3 Sensitivity of the net flows with respect to change in the weight of the CI**

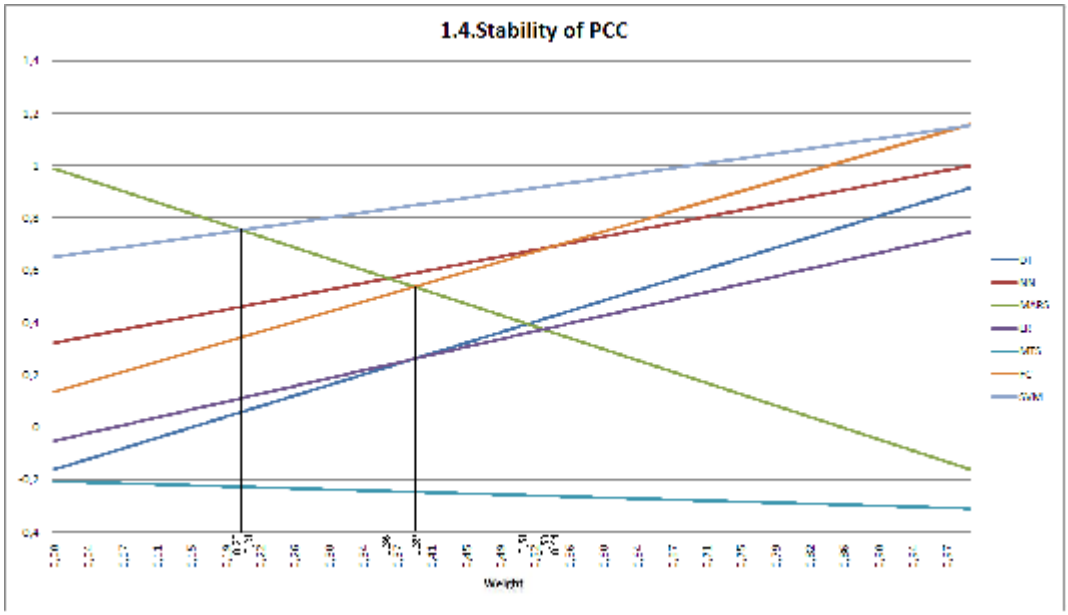



**Figure J. 4 Sensitivity of the net flows with respect to change in the weight of the Stability of PCC**

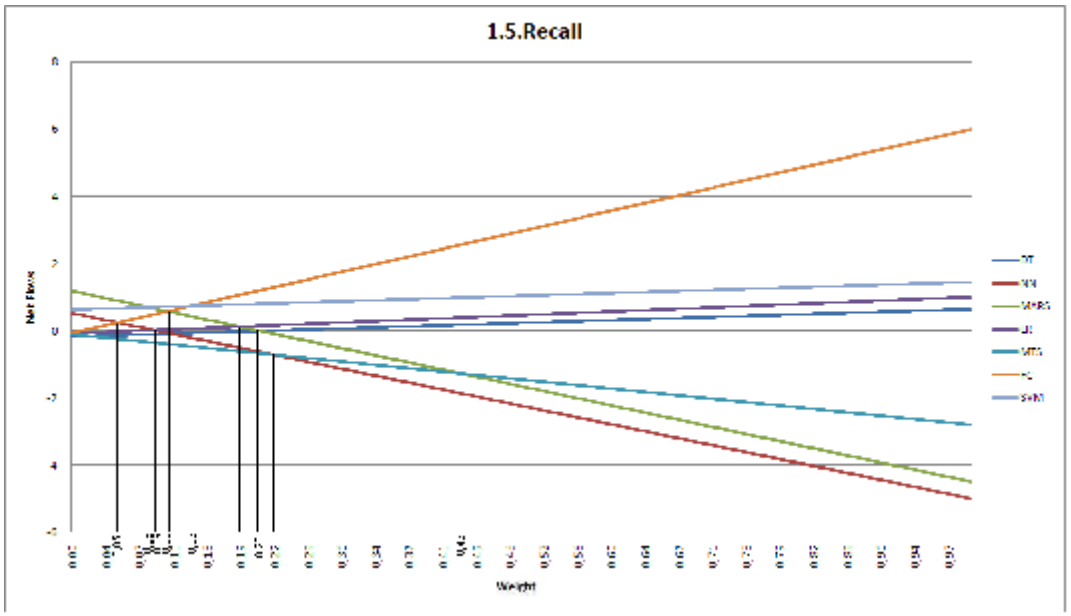



**Figure J. 5 Sensitivity of the net flows with respect to change in the weight of the Recall**

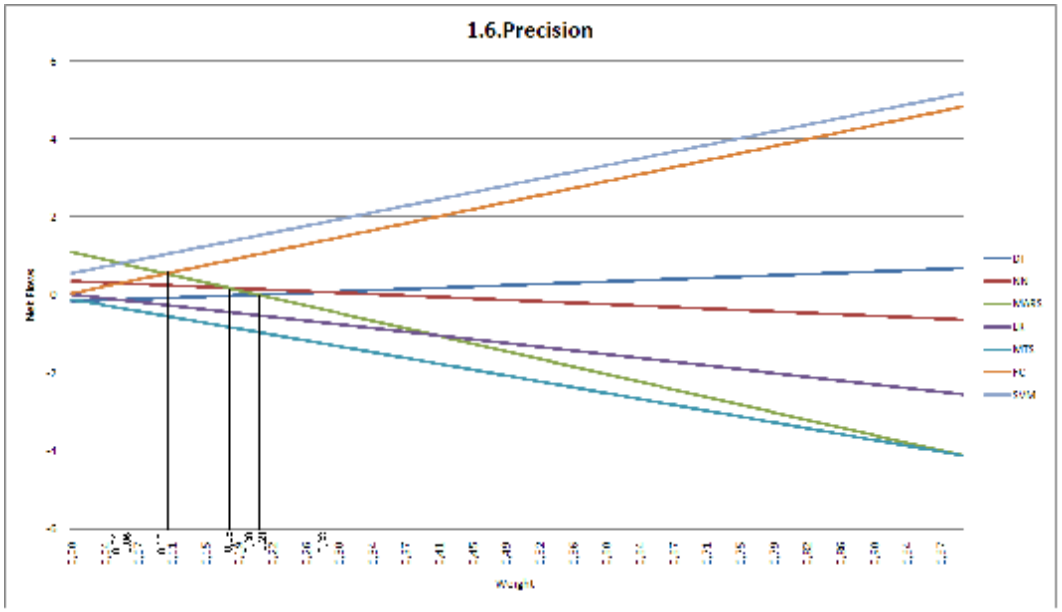



**Figure J. 6 Sensitivity of the net flows with respect to change in the weight of the Precision**

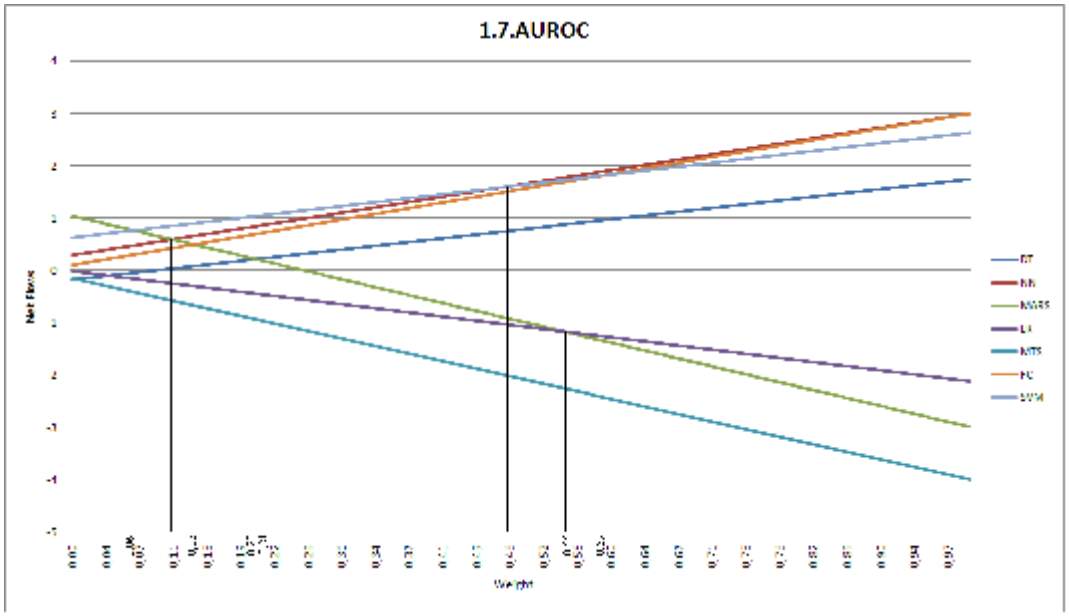



**Figure J. 7 Sensitivity of the net flows with respect to change in the weight of the AUROC**

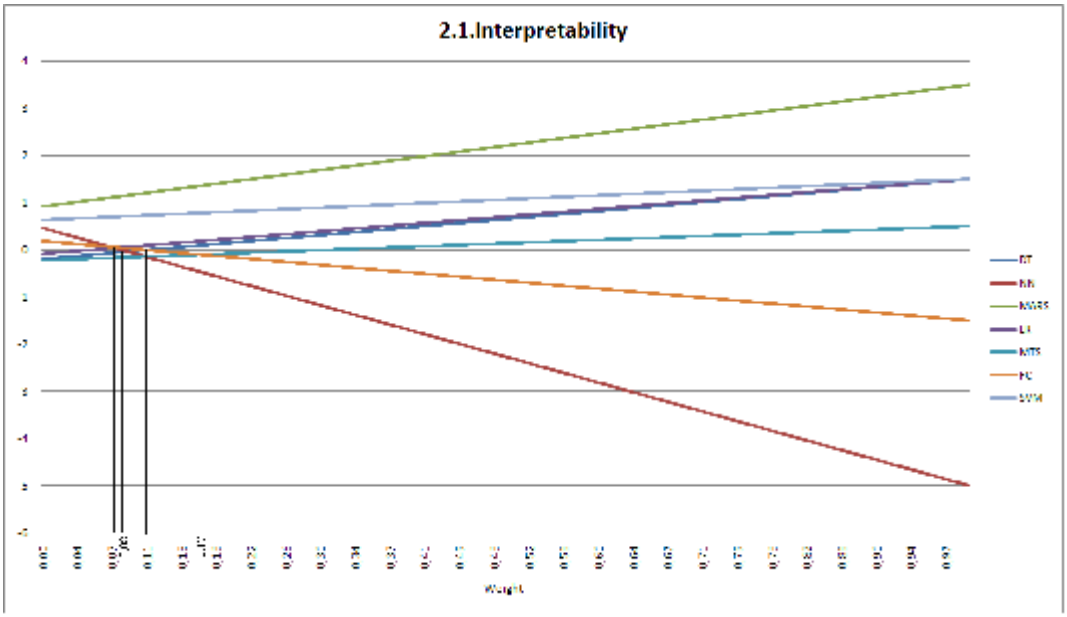



**Figure J. 8 Sensitivity of the net flows with respect to change in the weight of the Interpretability**

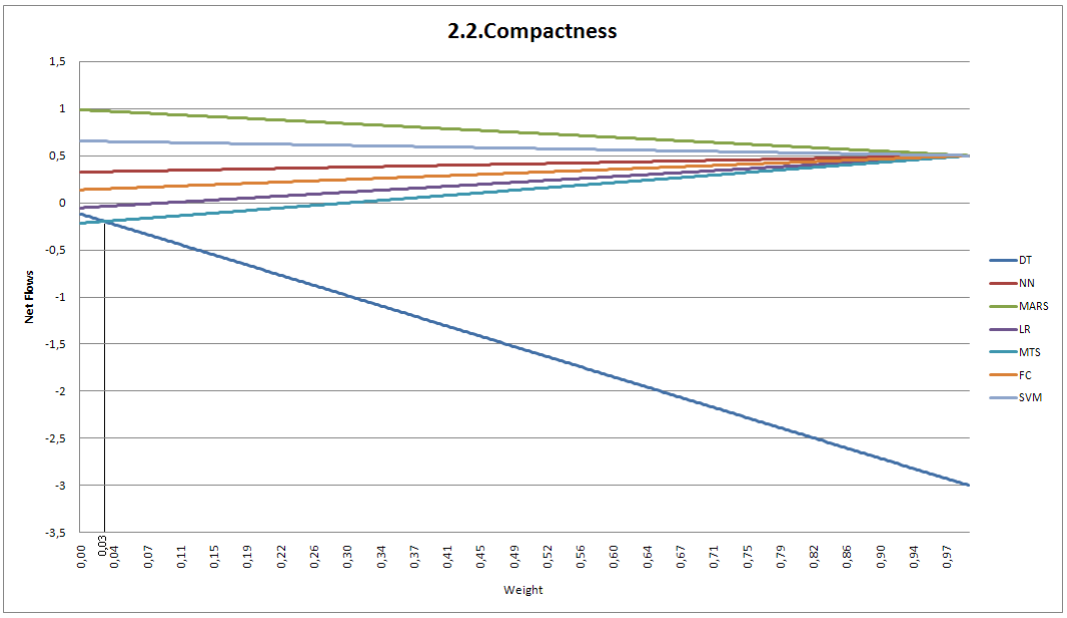



**Figure J. 9 Sensitivity of the net flows with respect to change in the weight of the Compactness**

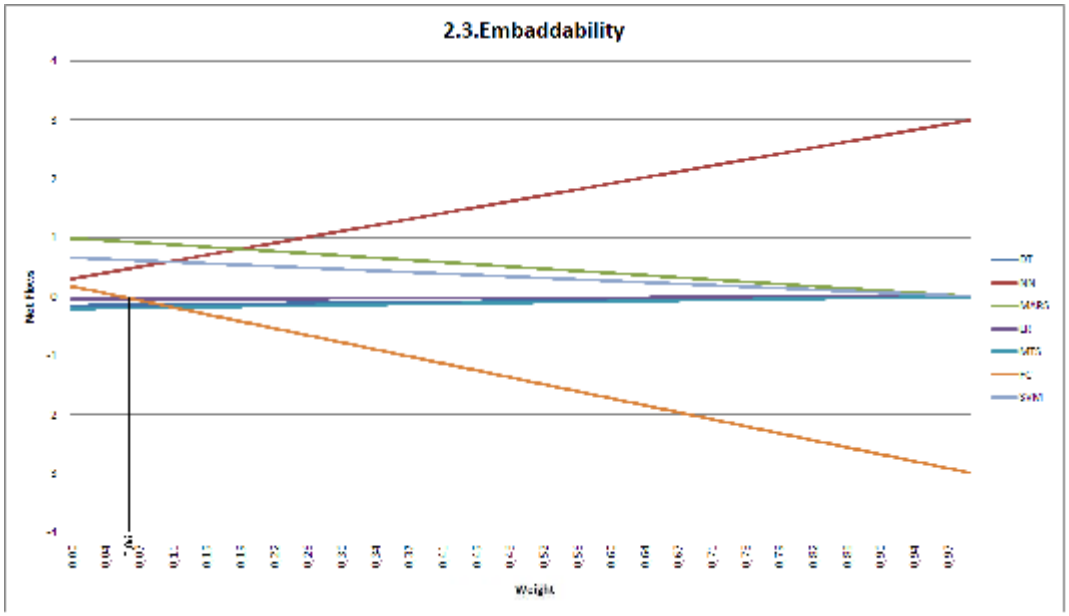



**Figure J. 10 Sensitivity of the net flows with respect to change in the weight of the Embaddability**

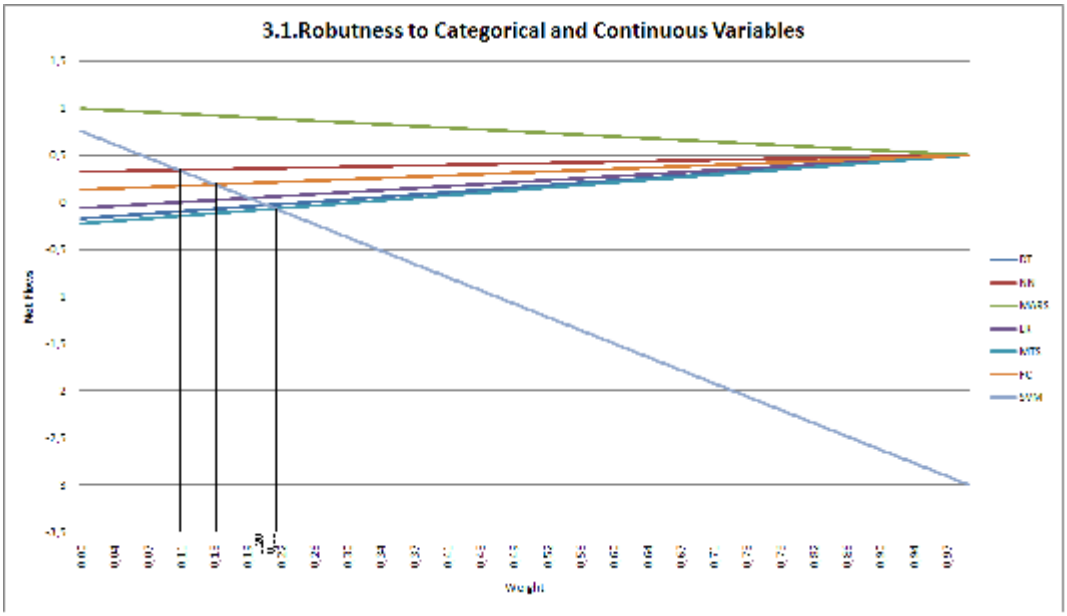



**Figure J. 11 Sensitivity of the net flows with respect to change in the weight of the Robustness to categorical and continuous variables**

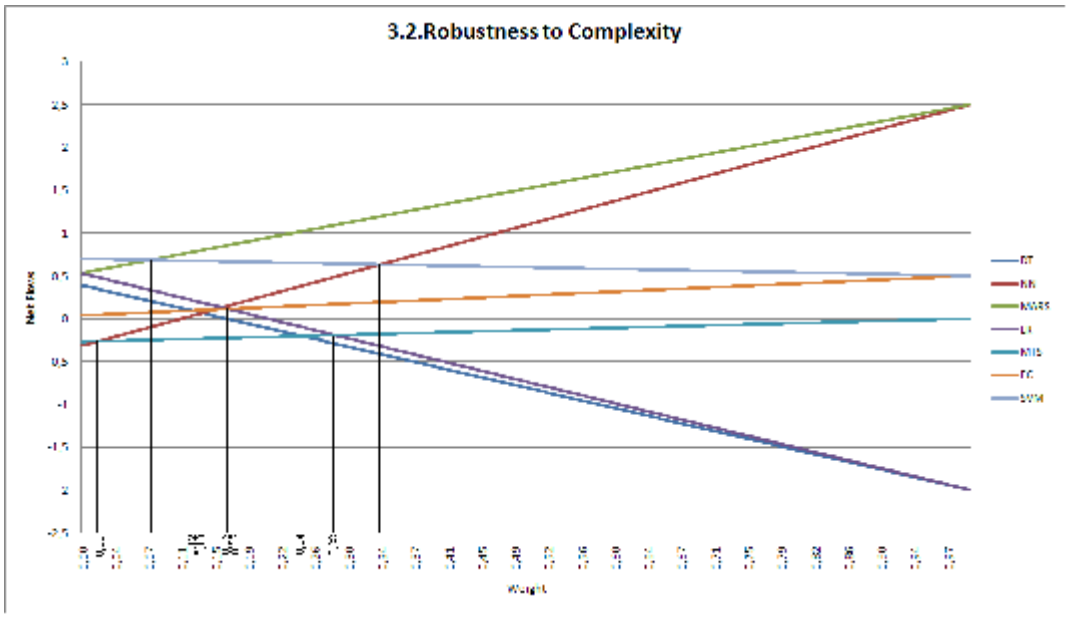



**Figure J. 12 Sensitivity of the net flows with respect to change in the weight of the Robustness to complexitiy**

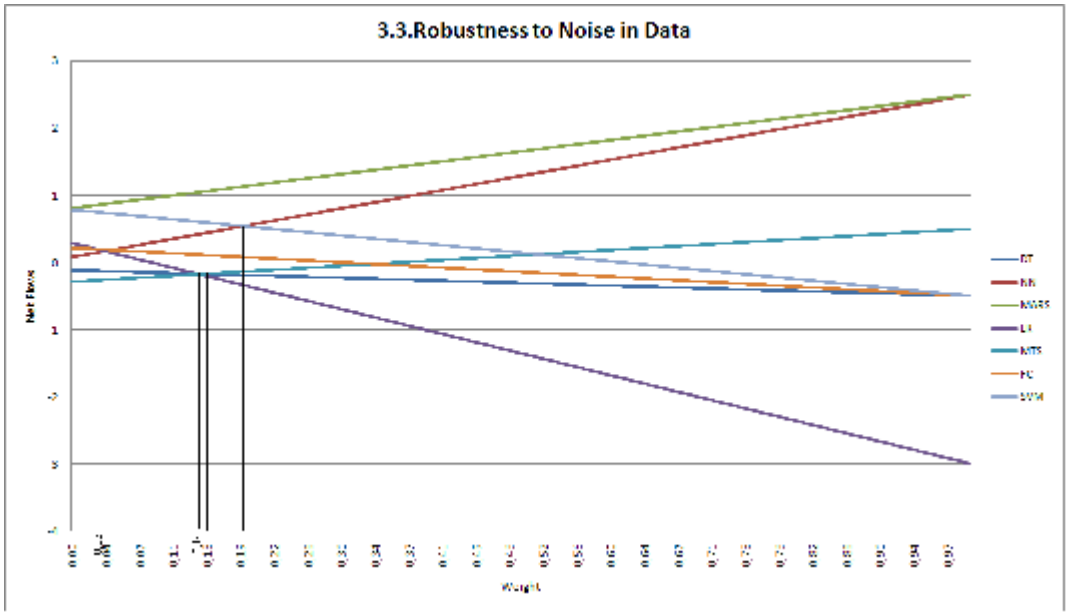



**Figure J. 13 Sensitivity of the net flows with respect to change in the weight of the Robustness to noise in data**

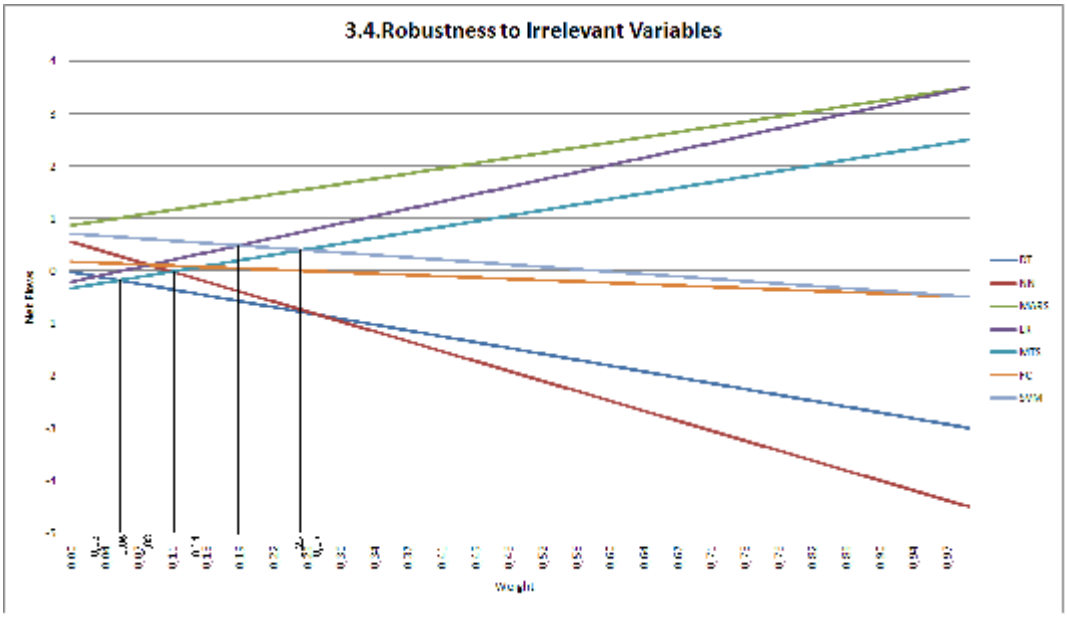



**Figure J. 14 Sensitivity of the net flows with respect to change in the weight of the Robustness to irrelevant variables**

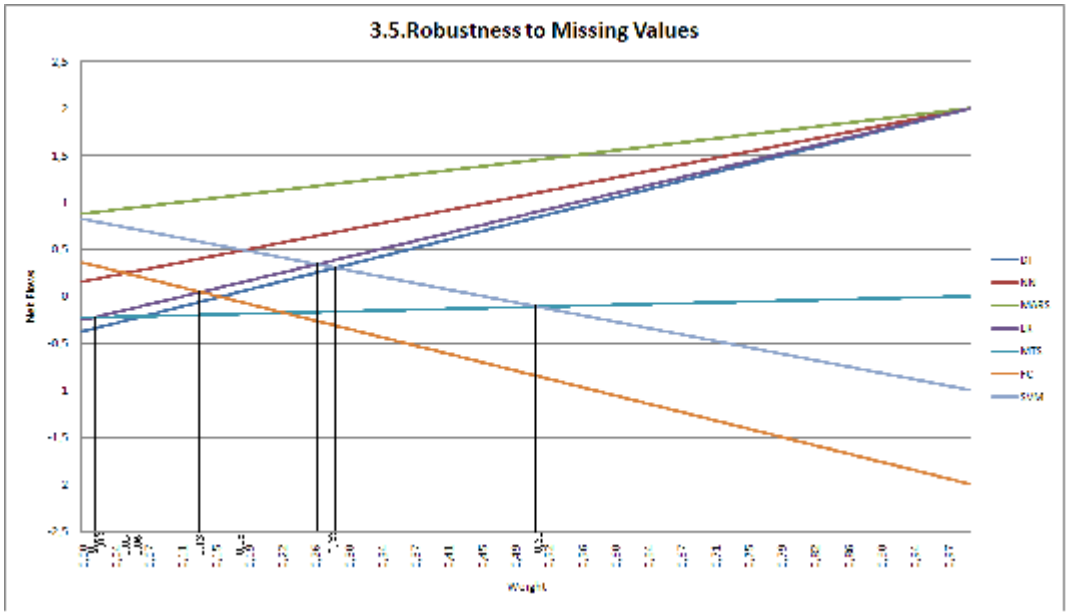



**Figure J. 15 Sensitivity of the net flows with respect to change in the weight of the Robustness to missing values**

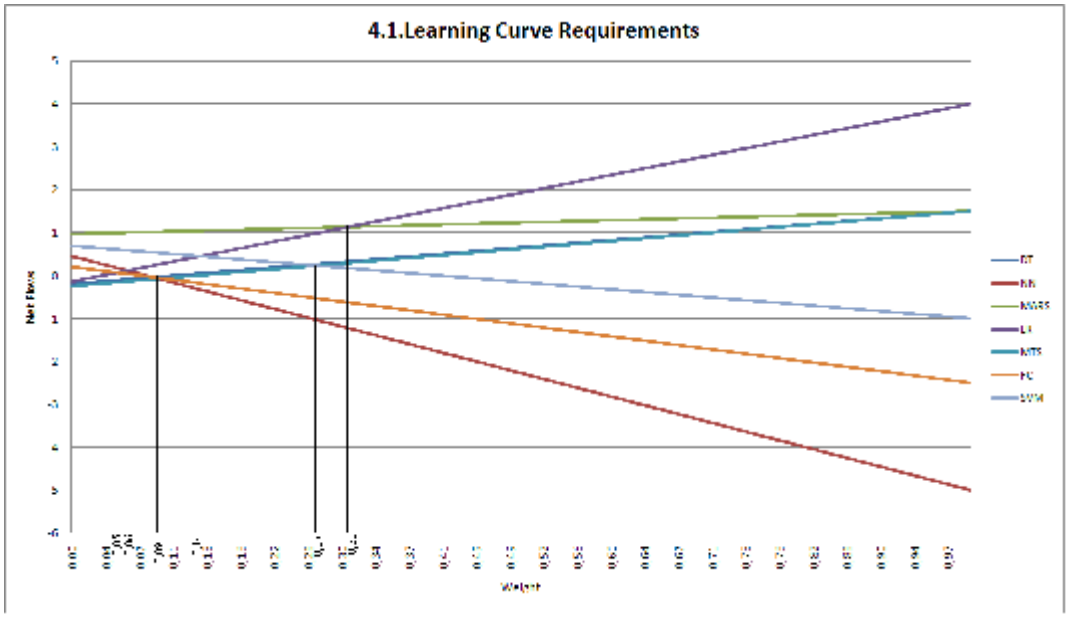



**Figure J. 16 Sensitivity of the net flows with respect to change in the weight of the Learning curve requirements**

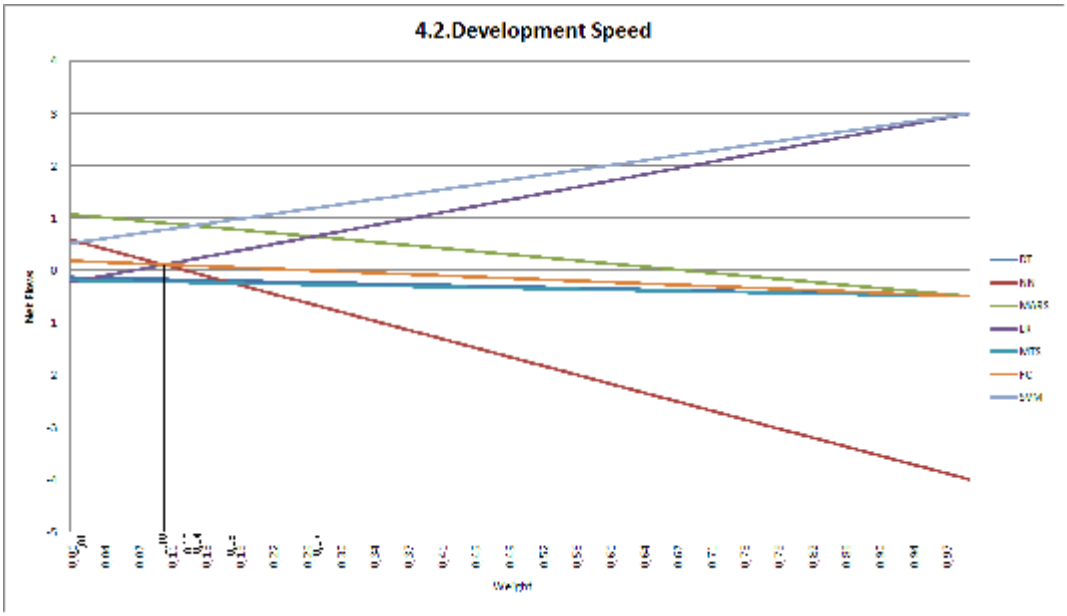



**Figure J. 17 Sensitivity of the net flows with respect to change in the weight of the Development speed**

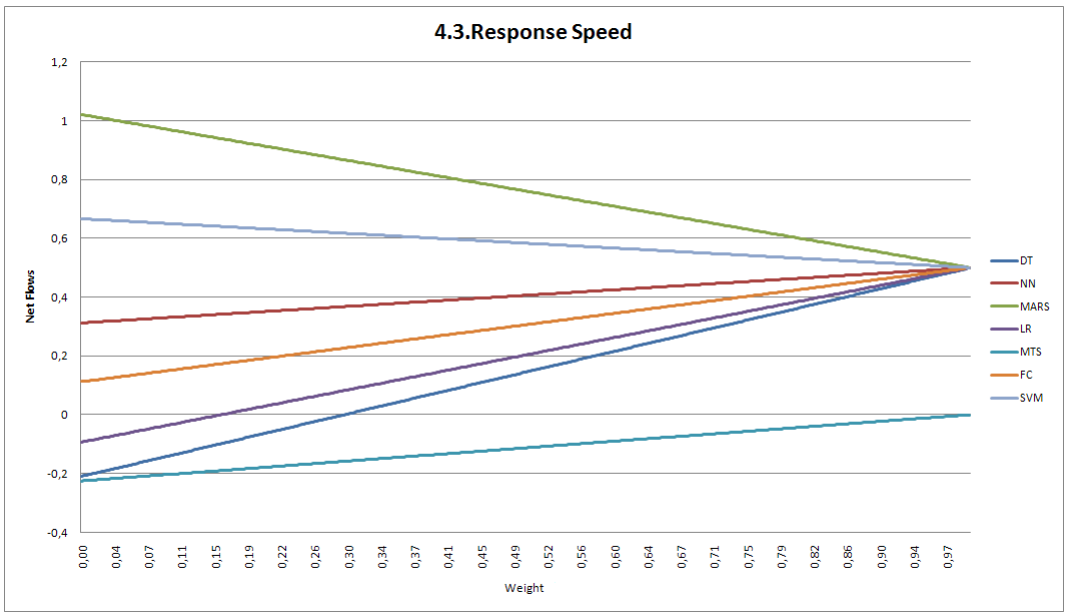



**Figure J. 18 Sensitivity of the net flows with respect to change in the weight of the Response speed**

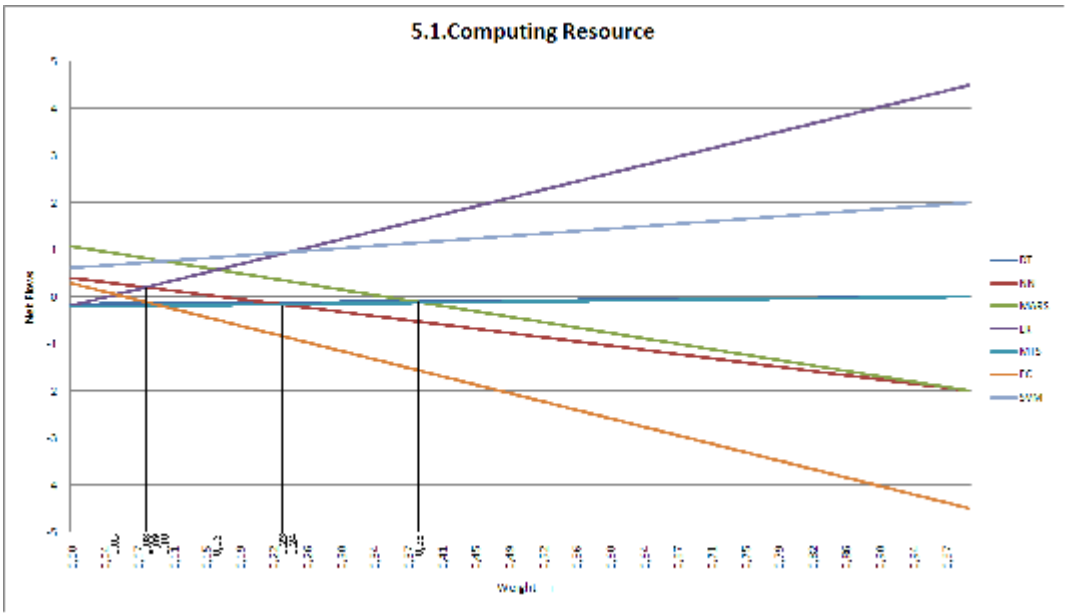



**Figure J. 19 Sensitivity of the net flows with respect to change in the weight of the Computing resource**

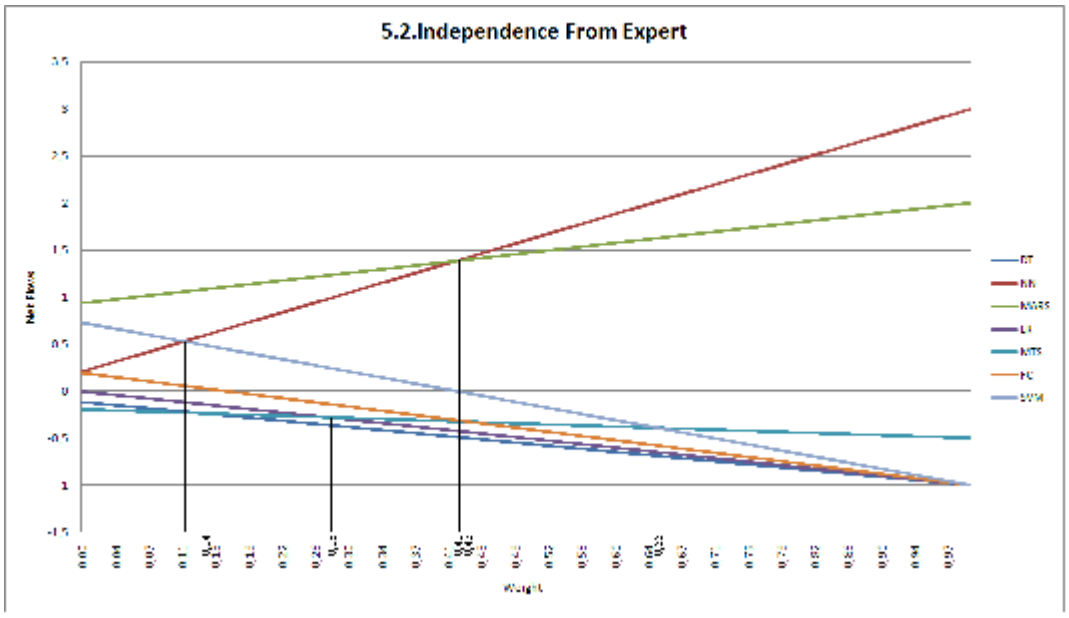



**Figure J. 20 Sensitivity of the net flows with respect to change in the weight of the Computing resource**

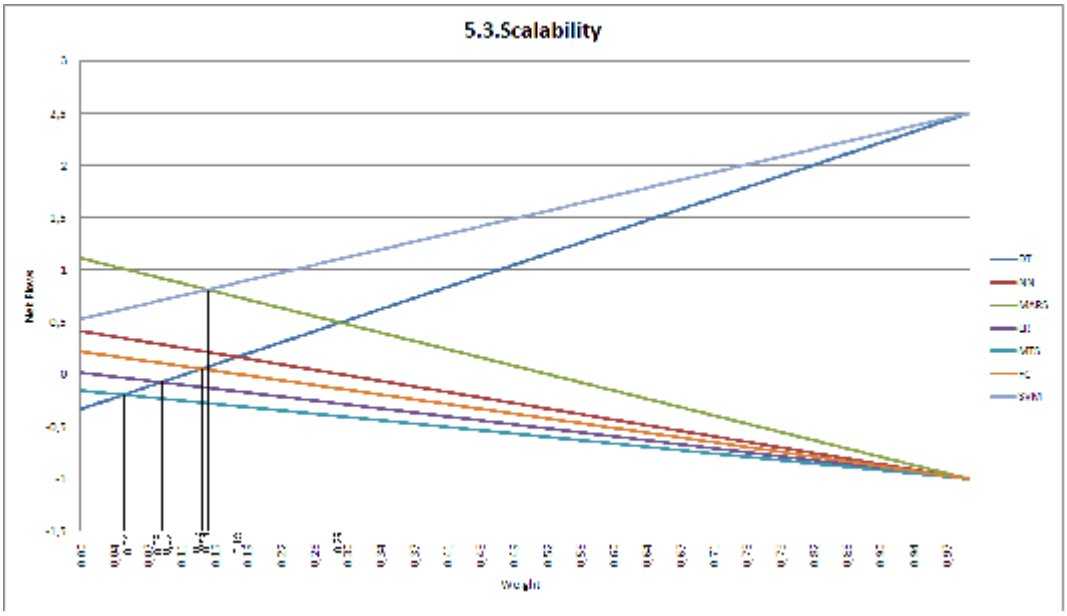



**Figure J. 21 Sensitivity of the net flows with respect to change in the weight of the Scalability**

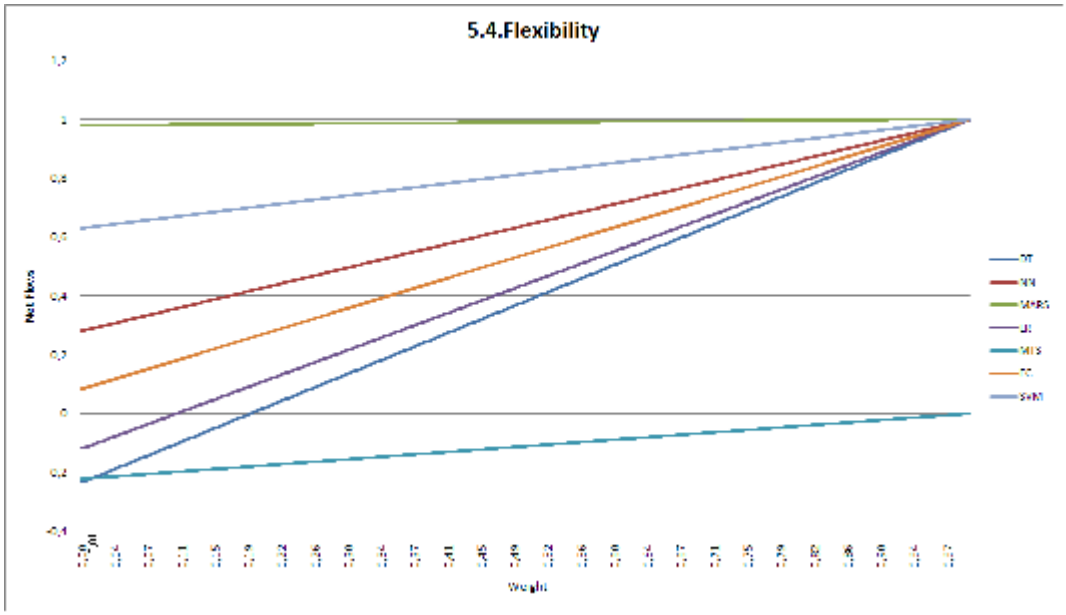



**Figure J. 22 Sensitivity of the net flows with respect to change in the weight of the Flexibility**

## FOR SUB-CRITERIA OF THE PREDICTION METHODS

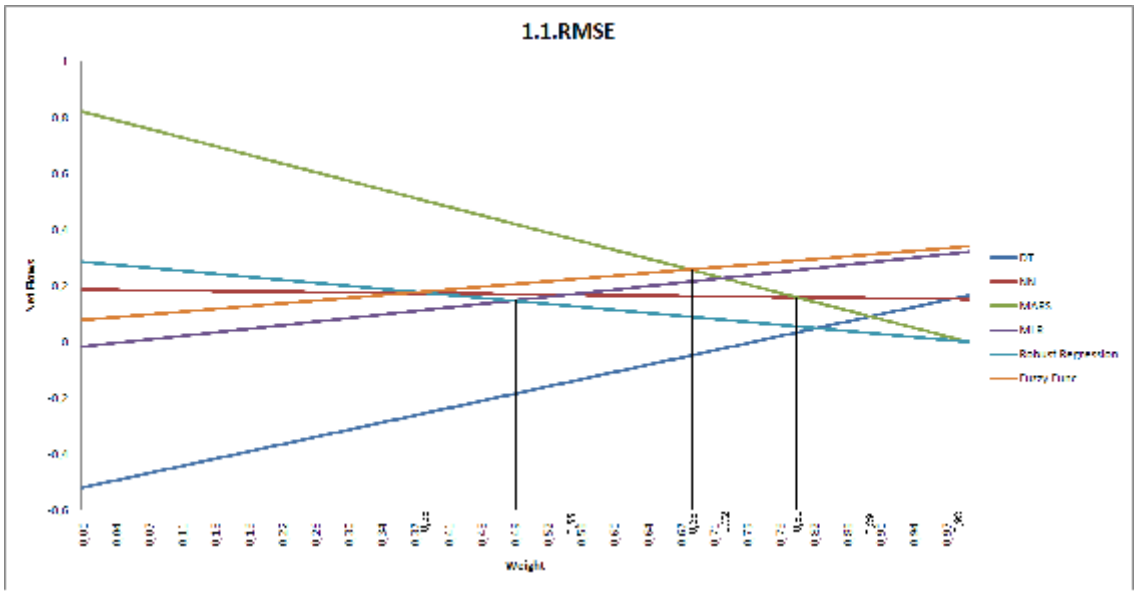



**Figure J. 23 Sensitivity of the net flows with respect to change in the weight of the RMSE**

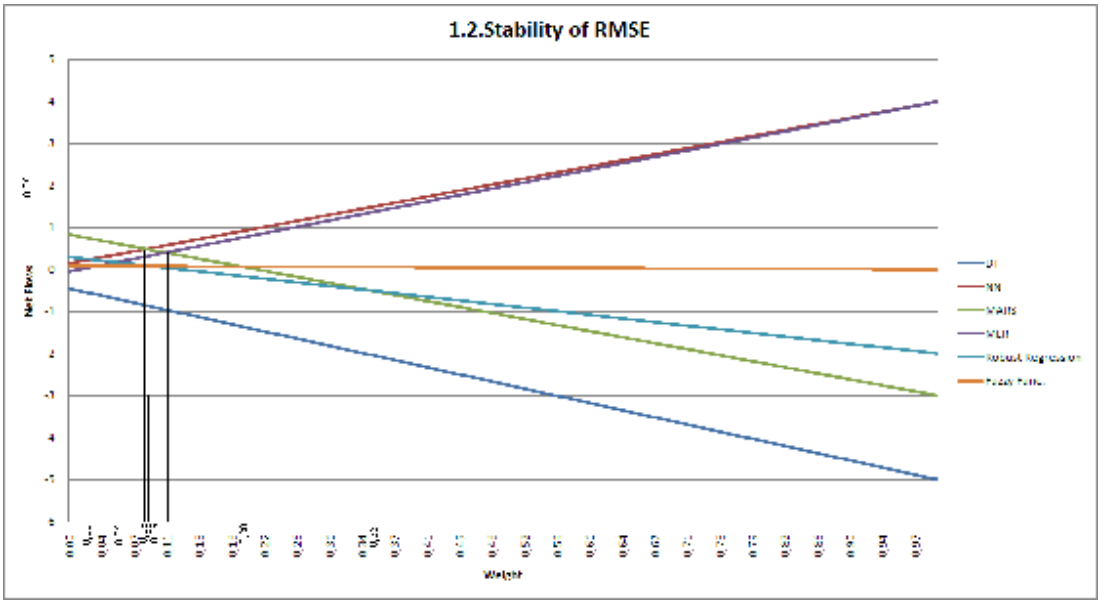



**Figure J. 24 Sensitivity of the net flows with respect to change in the weight of the Stability of RMSE**

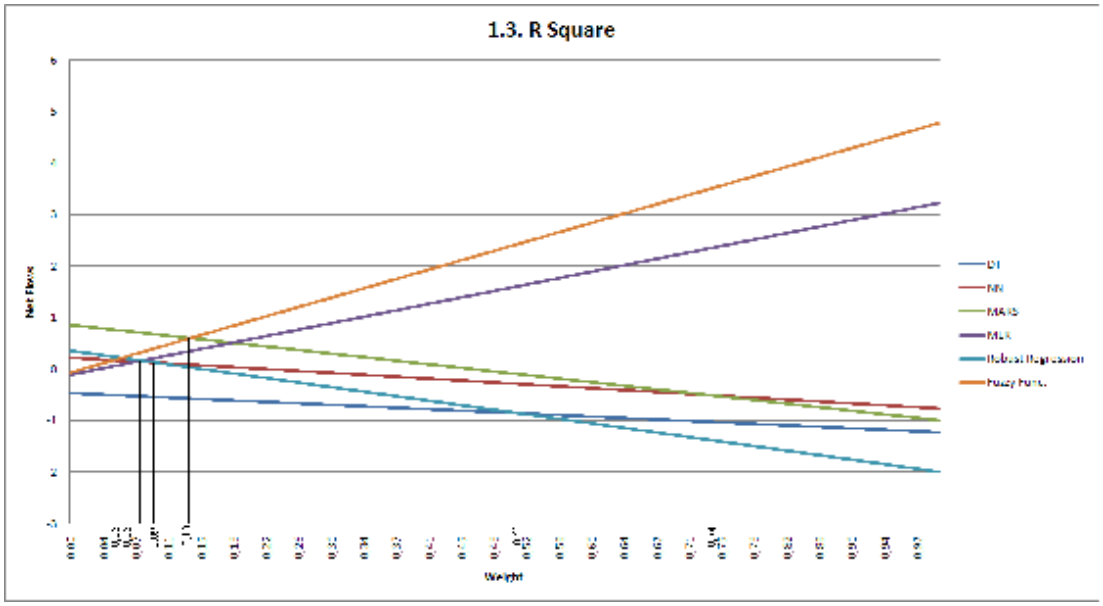



**Figure J. 25 Sensitivity of the net flows with respect to change in the weight of the R Square**

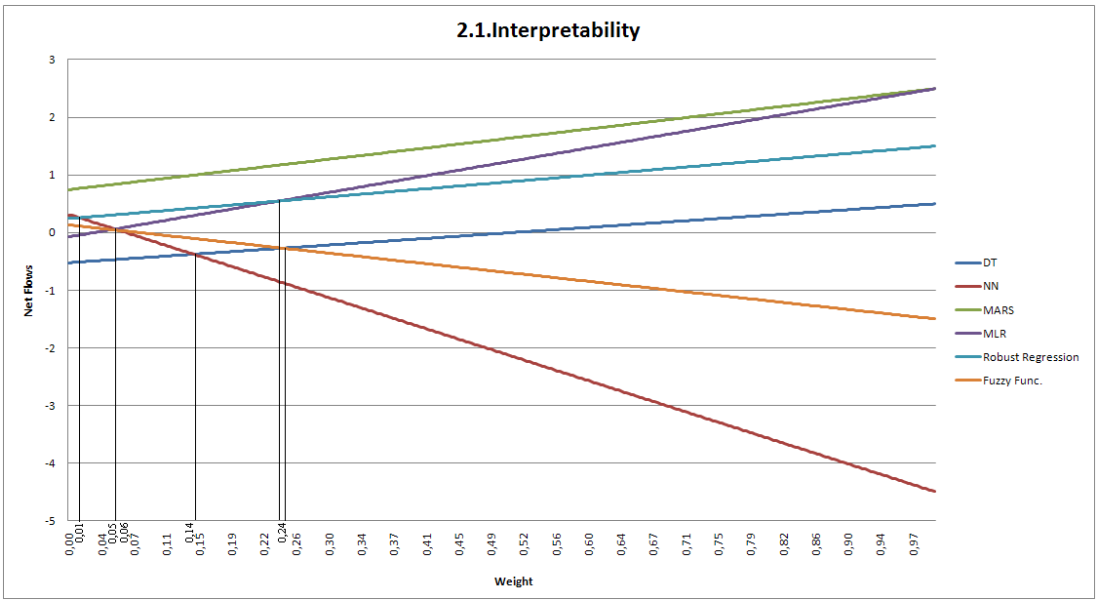



**Figure J. 26 Sensitivity of the net flows with respect to change in the weight of the Interpretability**

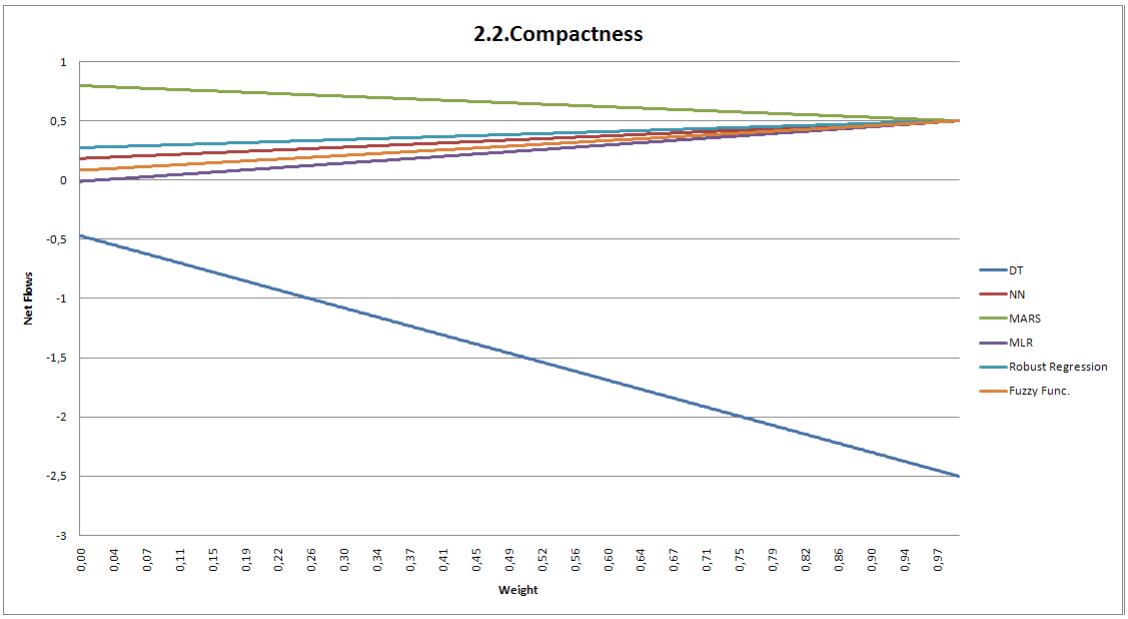



**Figure J. 27 Sensitivity of the net flows with respect to change in the weight of the Compactness**

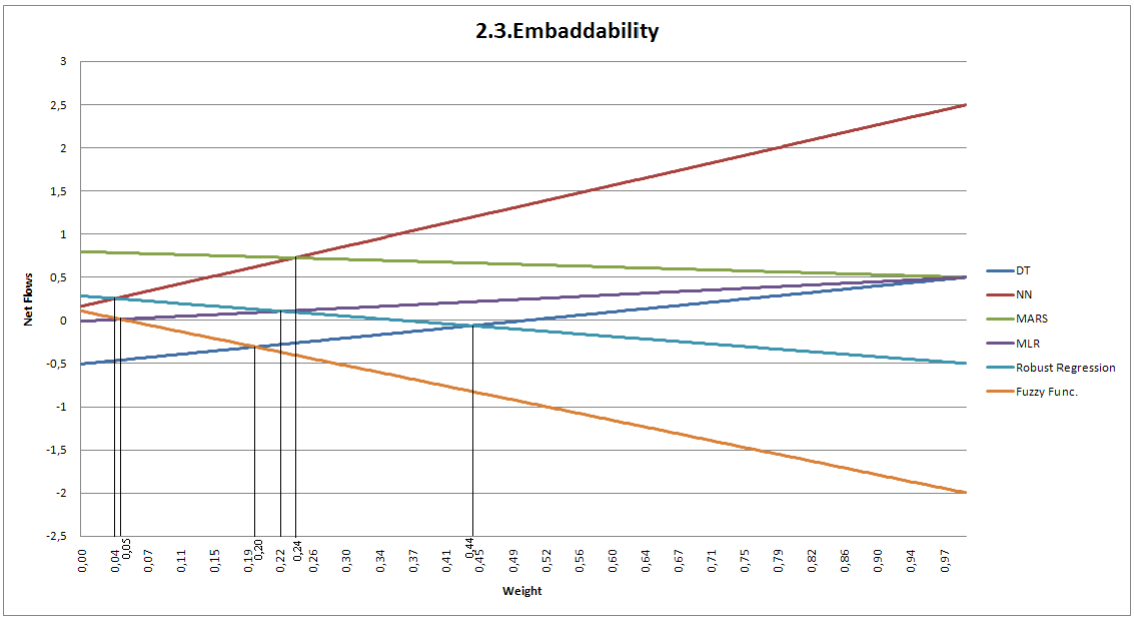



**Figure J. 28 Sensitivity of the net flows with respect to change in the weight of the Embaddability**

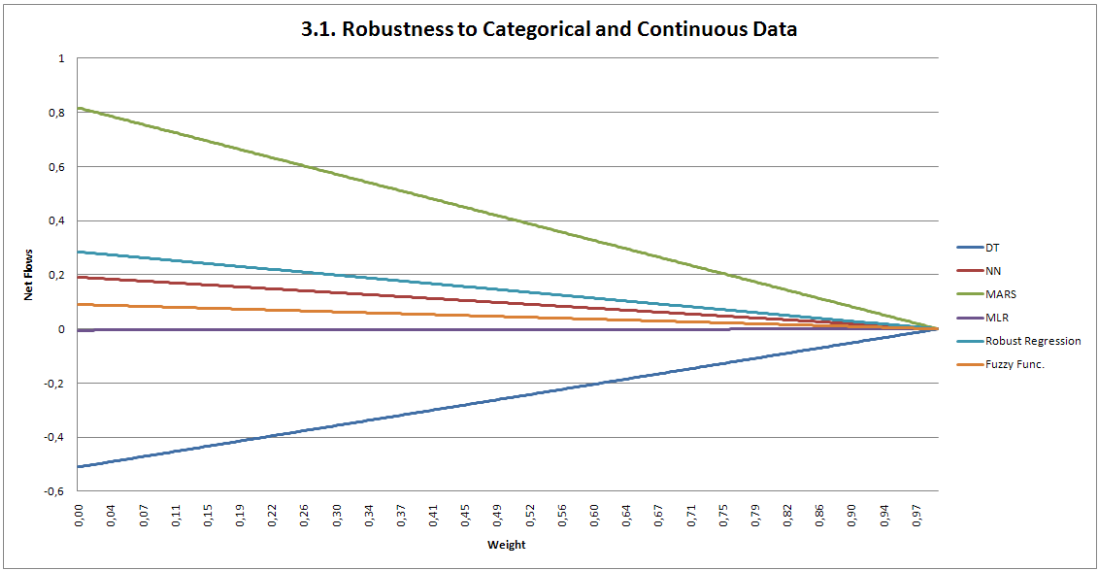



**Figure J. 29 Sensitivity of the net flows with respect to change in the weight of the Robustness to categorical and continuous data**

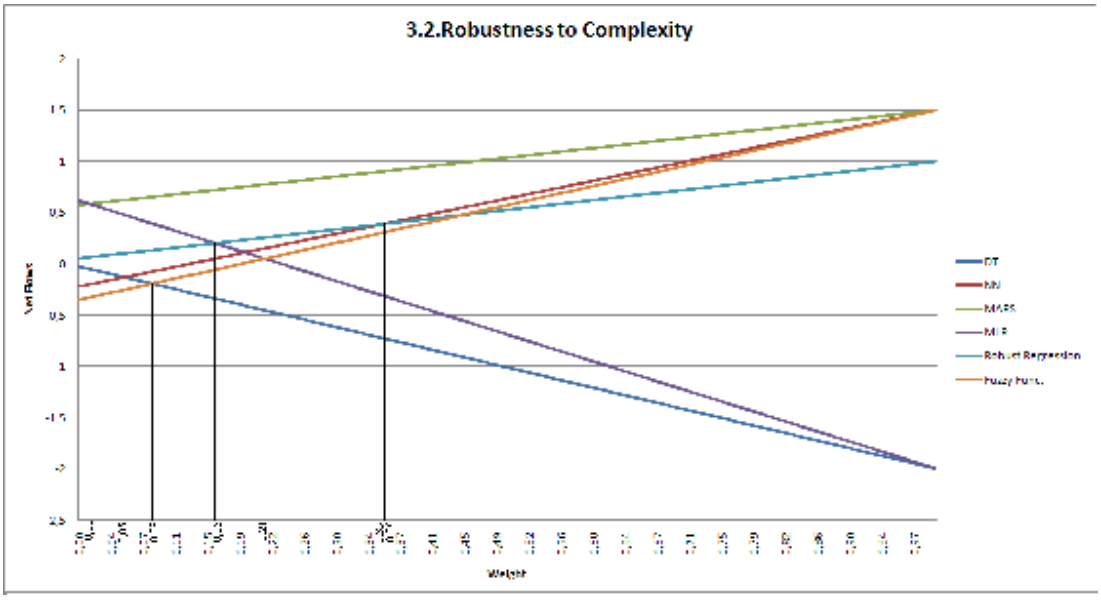



**Figure J. 30 Sensitivity of the net flows with respect to change in the weight of the Robustness to complexitiy**

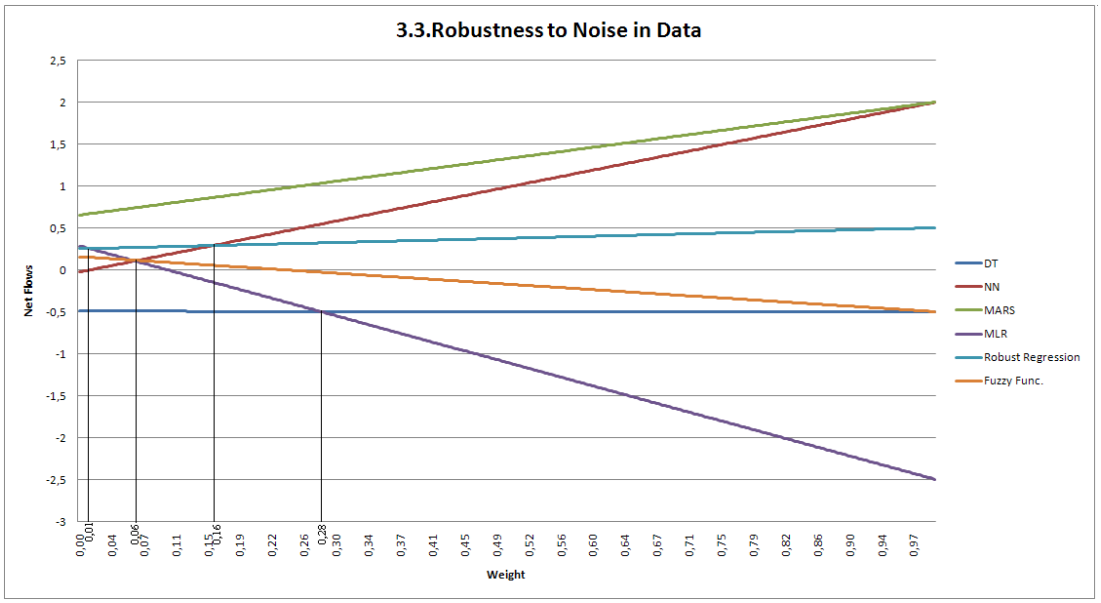



**Figure J. 31 Sensitivity of the net flows with respect to change in the weight of the Robustness to noise in data**

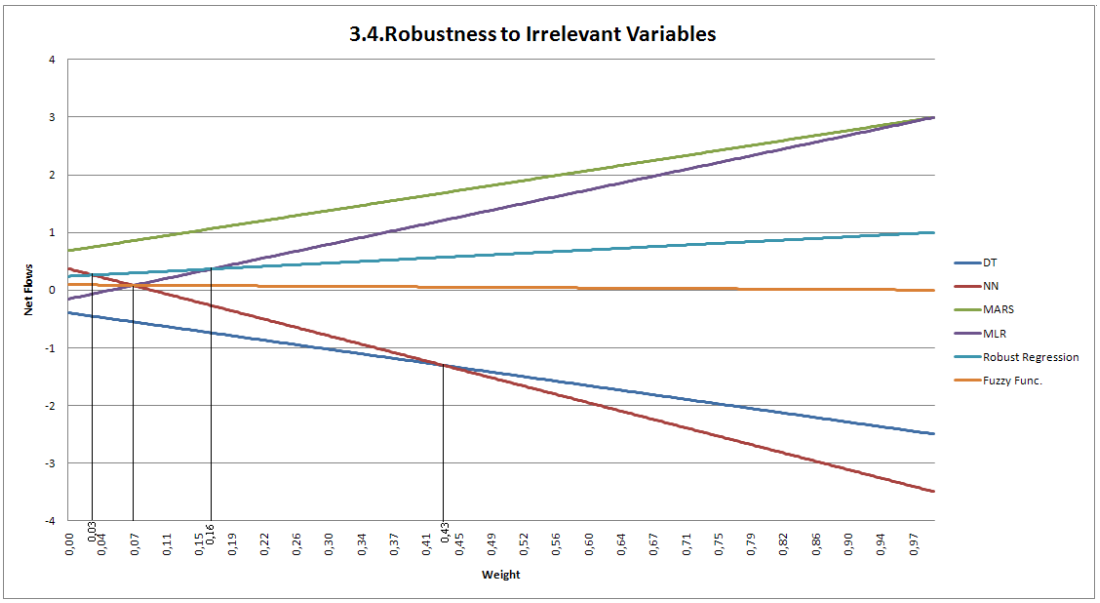



**Figure J. 32 Sensitivity of the net flows with respect to change in the weight of the Robustness to irrelevant variables**

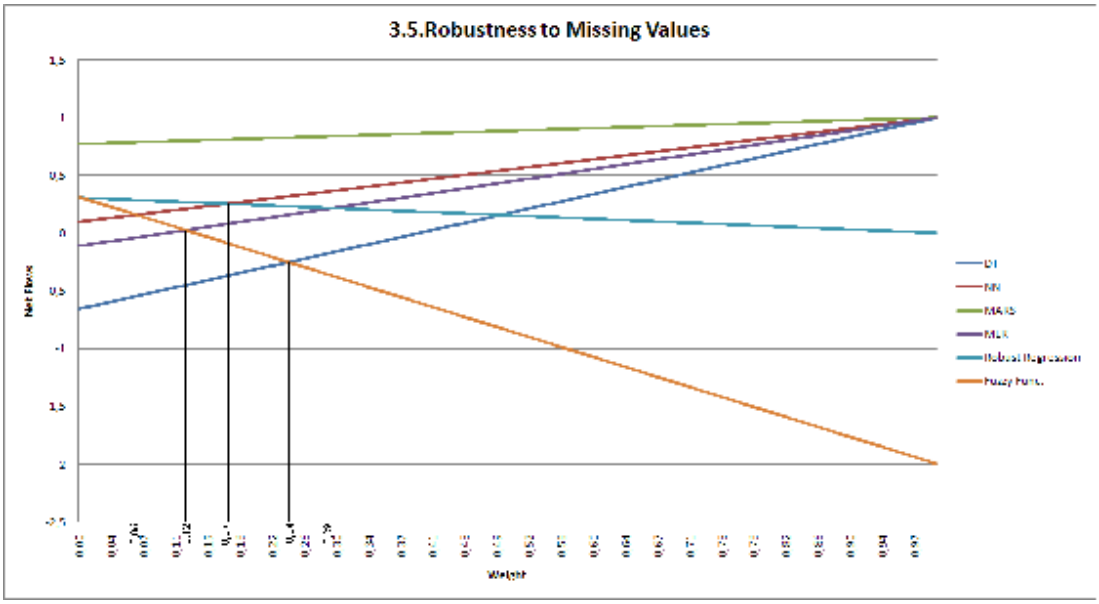



**Figure J. 33 Sensitivity of the net flows with respect to change in the weight of the Robustness to missing values**

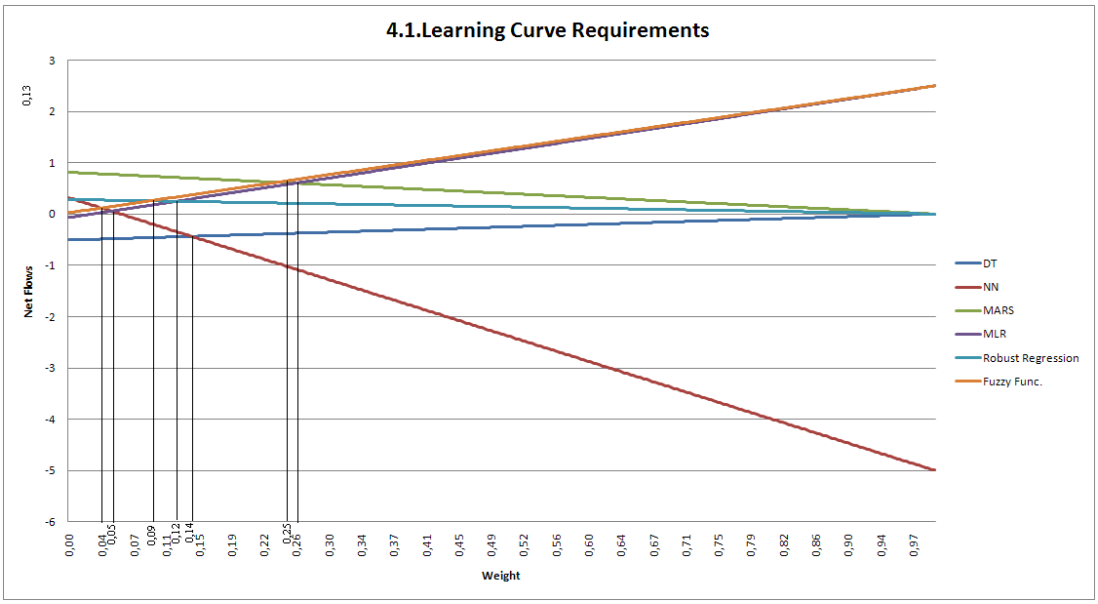



**Figure J. 34Sensitivity of the net flows with respect to change in the weight of the Learning curve requirements**

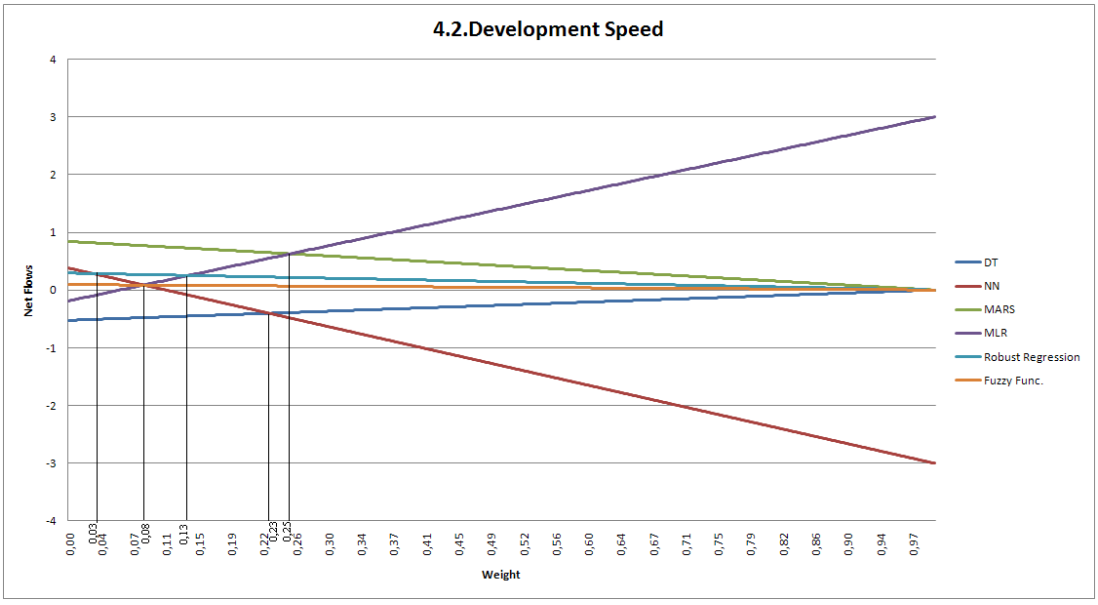



**Figure J. 35 Sensitivity of the net flows with respect to change in the weight of the Development speed**

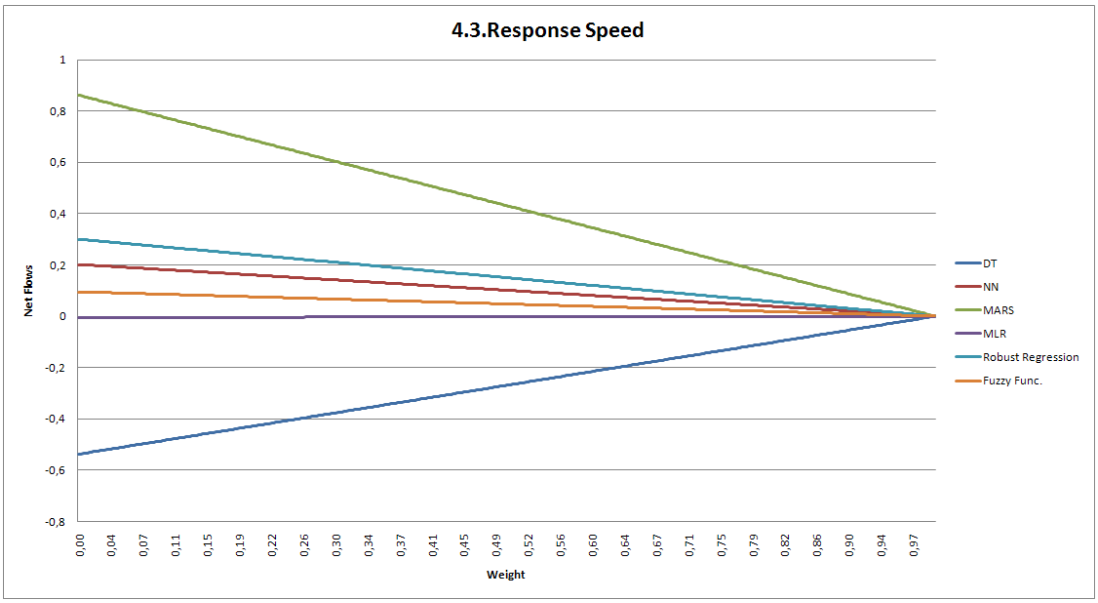



**Figure J. 36 Sensitivity of the net flows with respect to change in the weight of the Response speed**

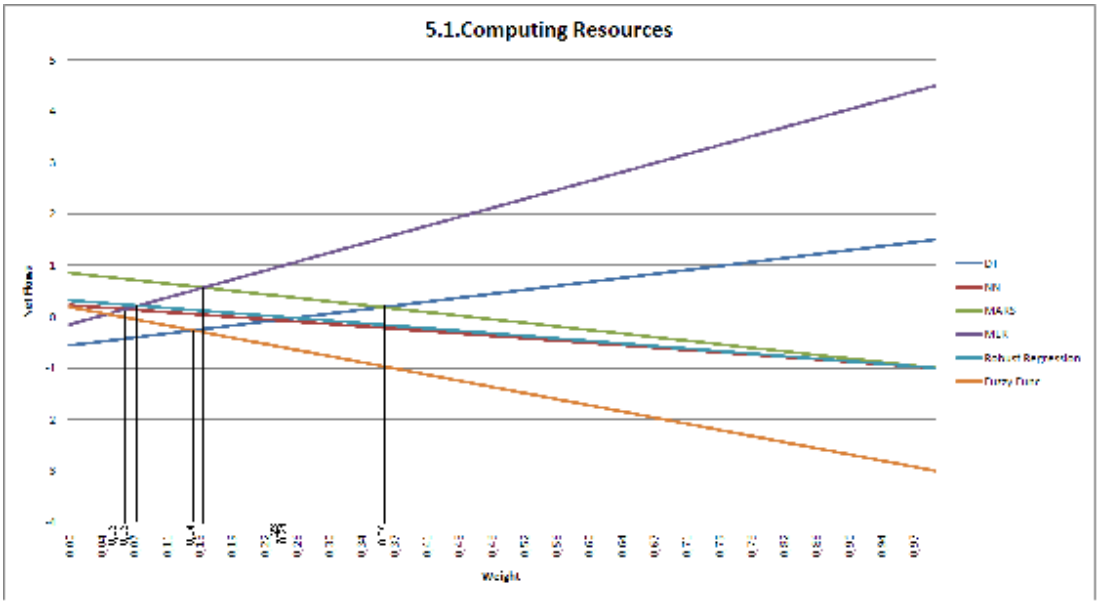



**Figure J. 37 Sensitivity of the net flows with respect to change in the weight of the Computing resource**

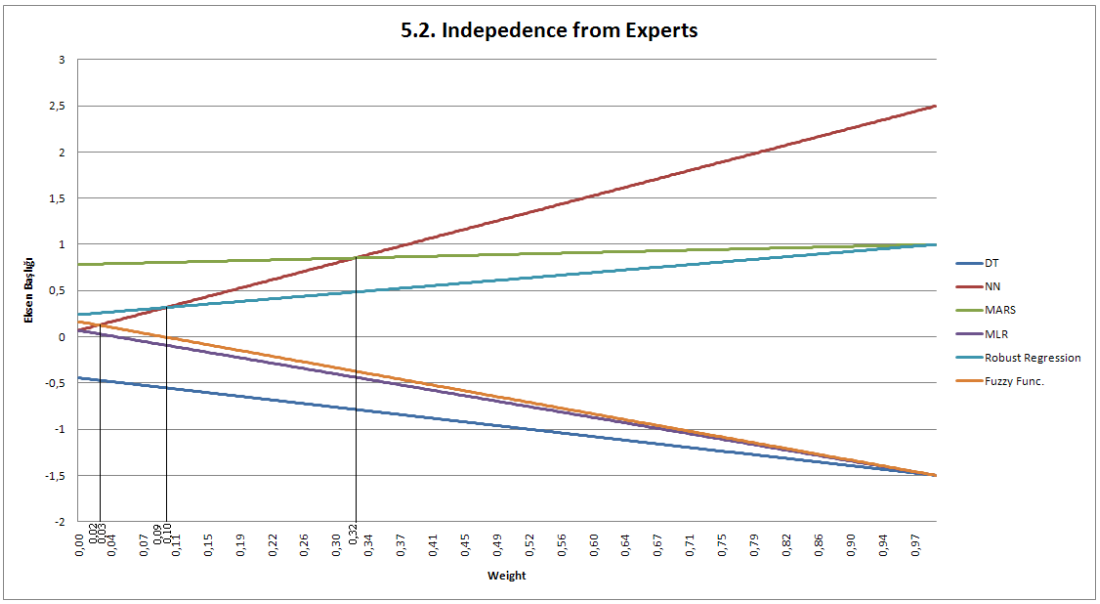



**Figure J. 38 Sensitivity of the net flows with respect to change in the weight of the Independence from experts**

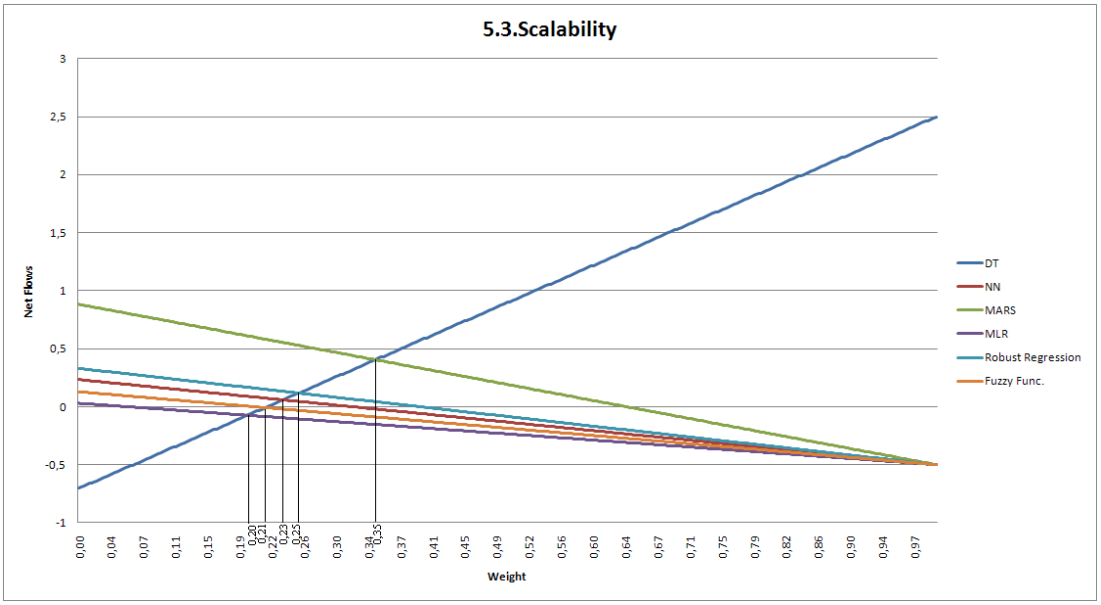



**Figure J. 39 Sensitivity of the net flows with respect to change in the weight of the Scalability**

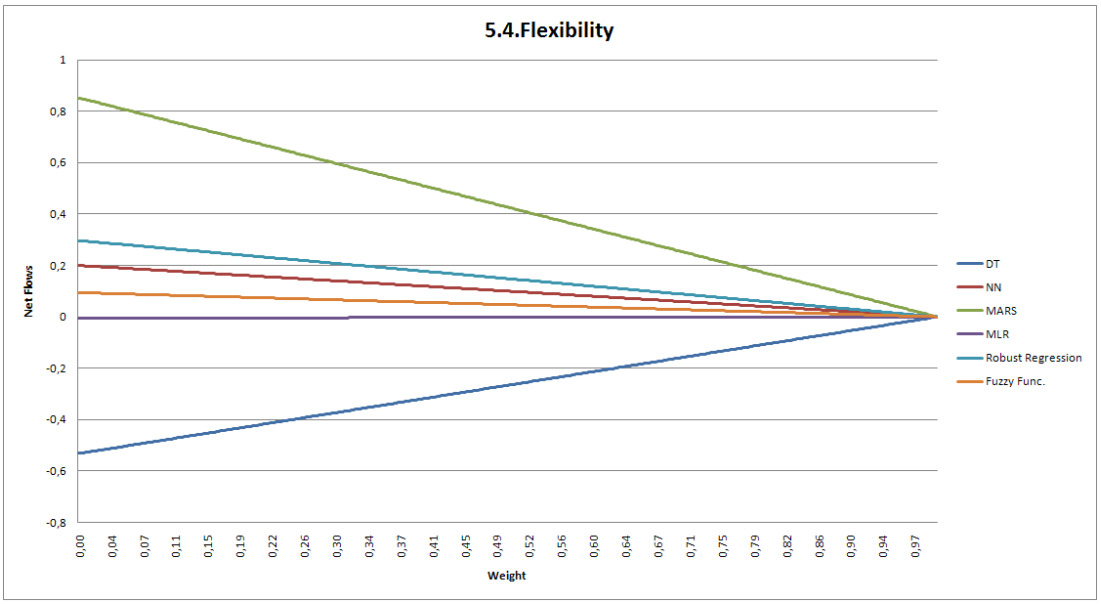



**Figure J. 40 Sensitivity of the net flows with respect to change in the weight of the Flexibility**

## FOR SUB-CRITERIA OF THE CLASSIFICATION METHODS SENSITIVITY OF THE THRESHOLDS

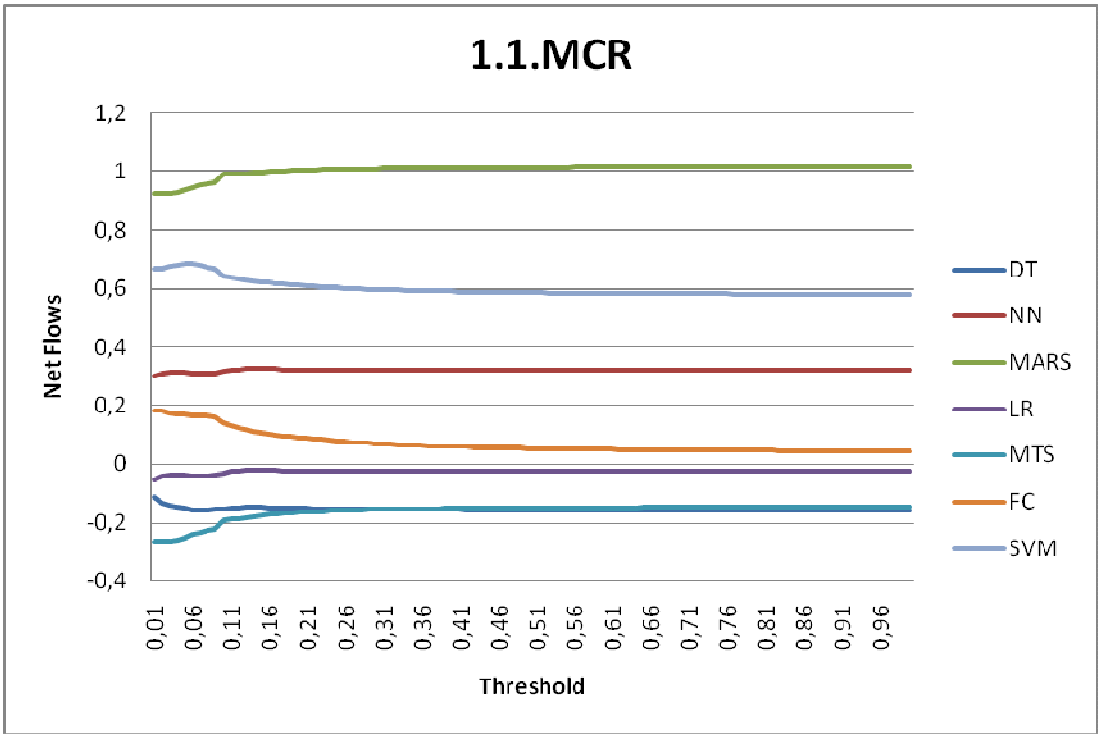



**Figure J. 41 Sensitivity of the net flows with respect to change in the threshold of the MCR**

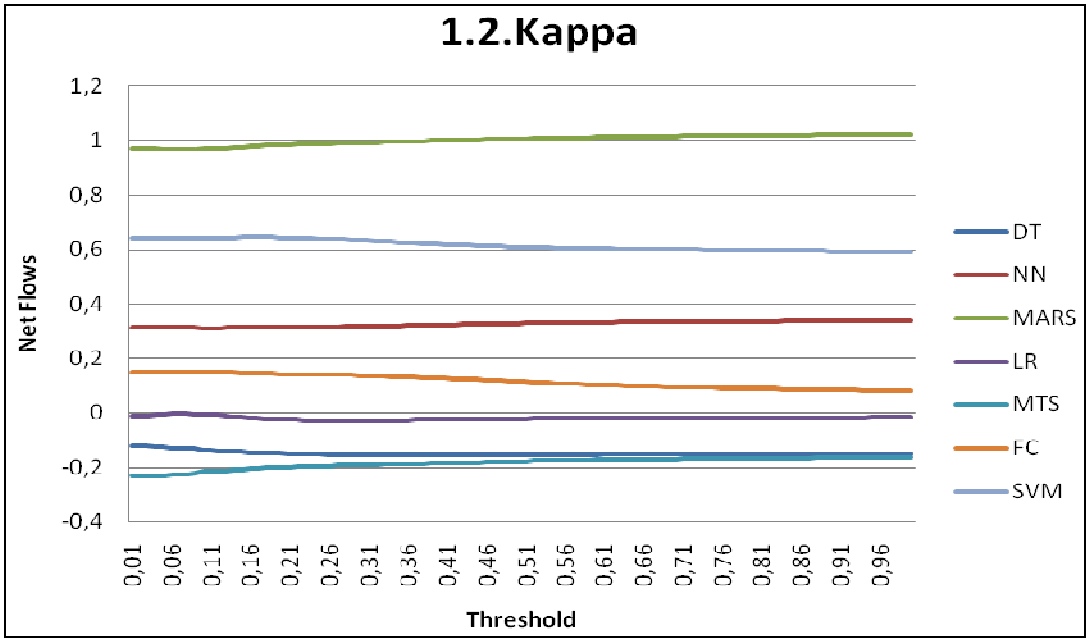



**Figure J. 42 Sensitivity of the net flows with respect to change in the threshold of the Kappa**

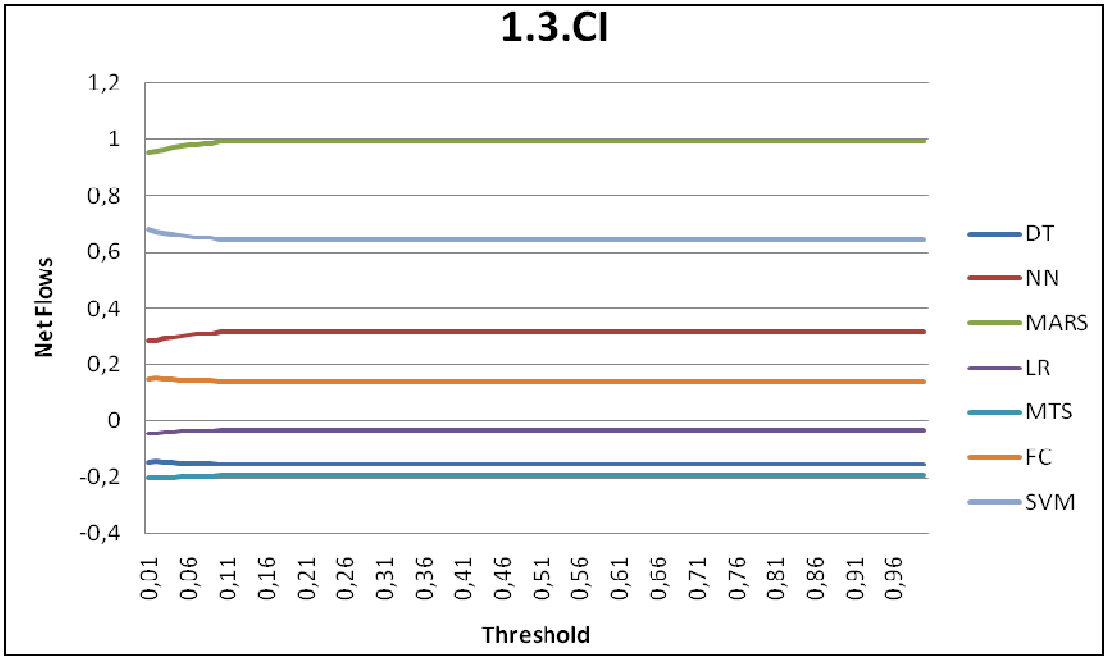



**Figure J. 43 Sensitivity of the net flows with respect to change in the threshold of the CI**

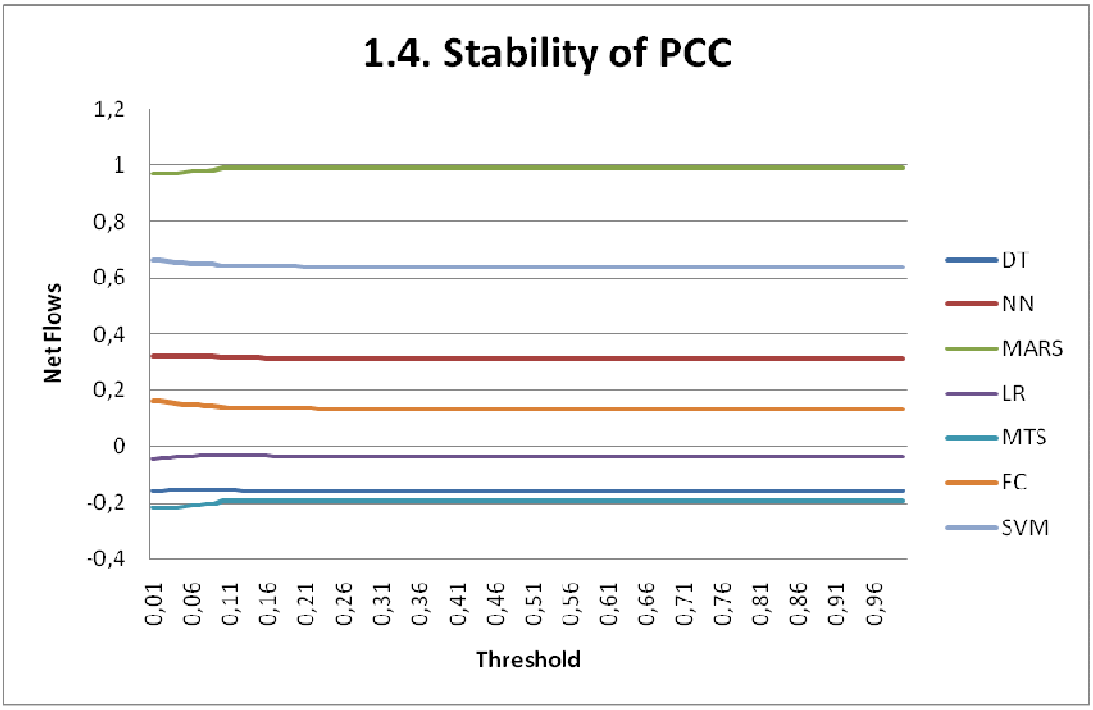



**Figure J. 44 Sensitivity of the net flows with respect to change in the threshold of the Stability**

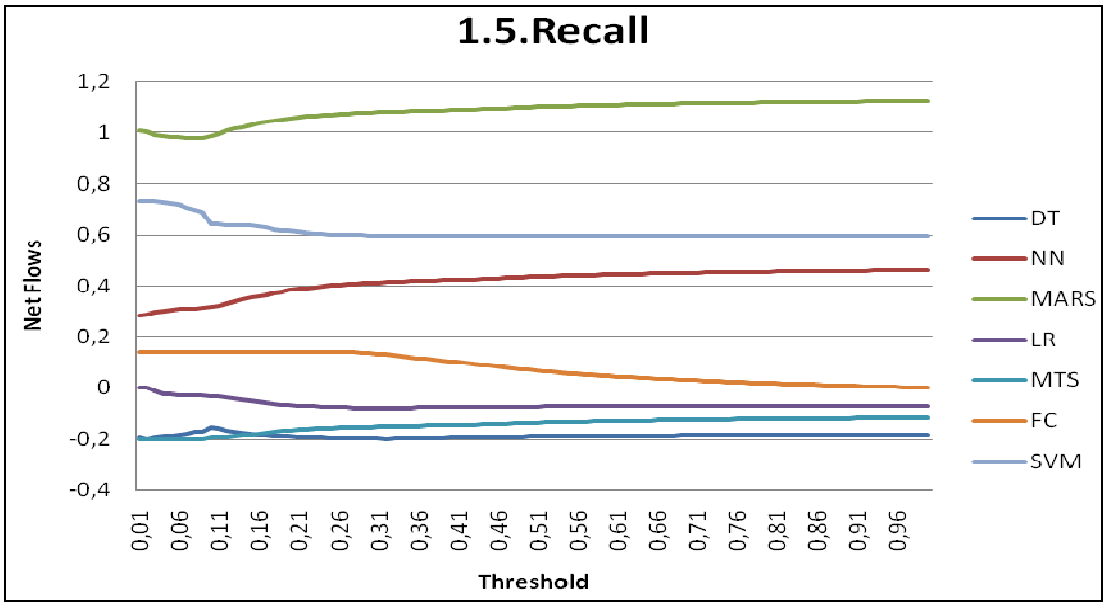



**Figure J. 45 Sensitivity of the net flows with respect to change in the threshold of the Recall**

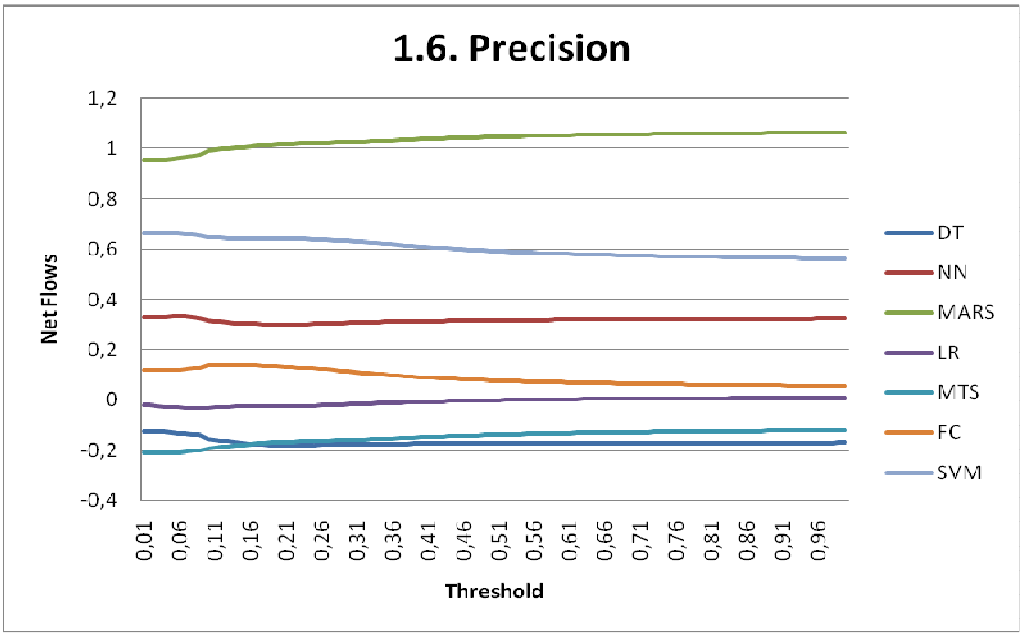



**Figure J. 46 Sensitivity of the net flows with respect to change in the threshold of the Precision**

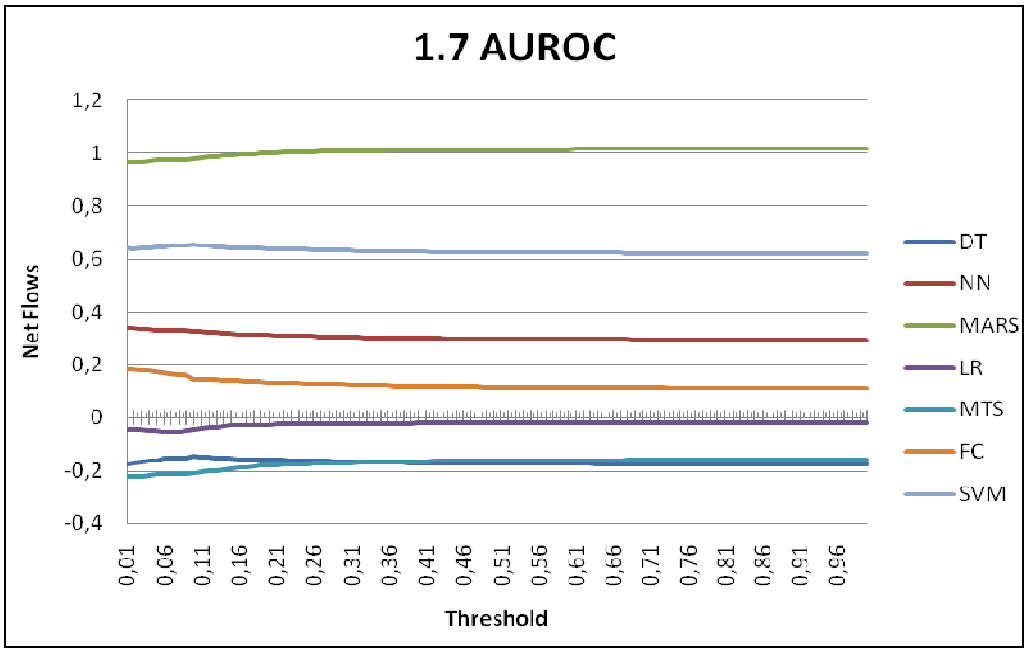



**Figure J. 47Sensitivity of the net flows with respect to change in the threshold of the AUROC**

## FOR SUB-CRITERIA OF THE PREDICTION METHODS

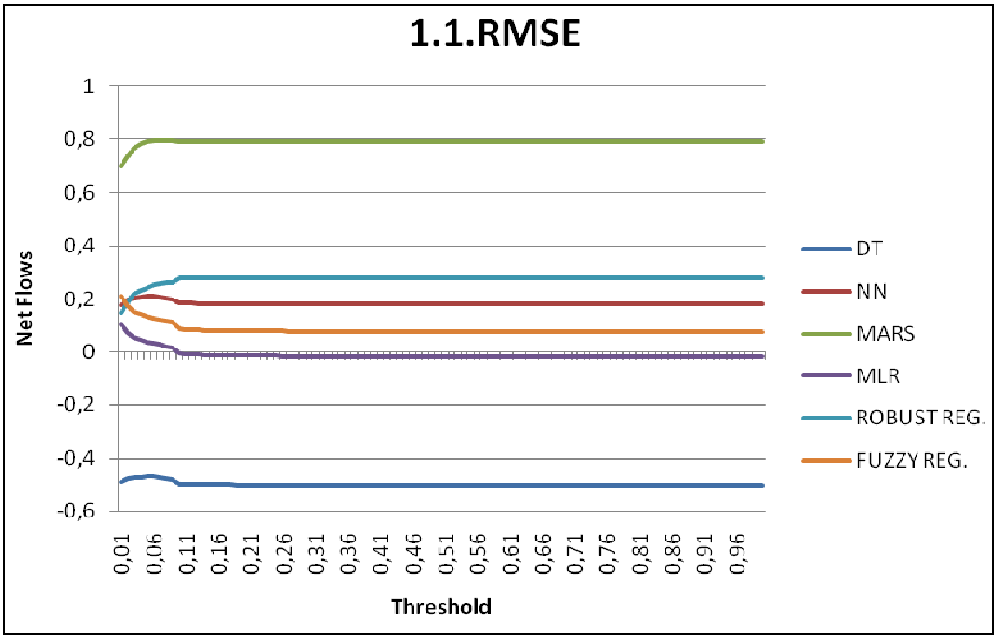



**Figure J. 48 Sensitivity of the net flows with respect to change in the threshold of the RMSE**

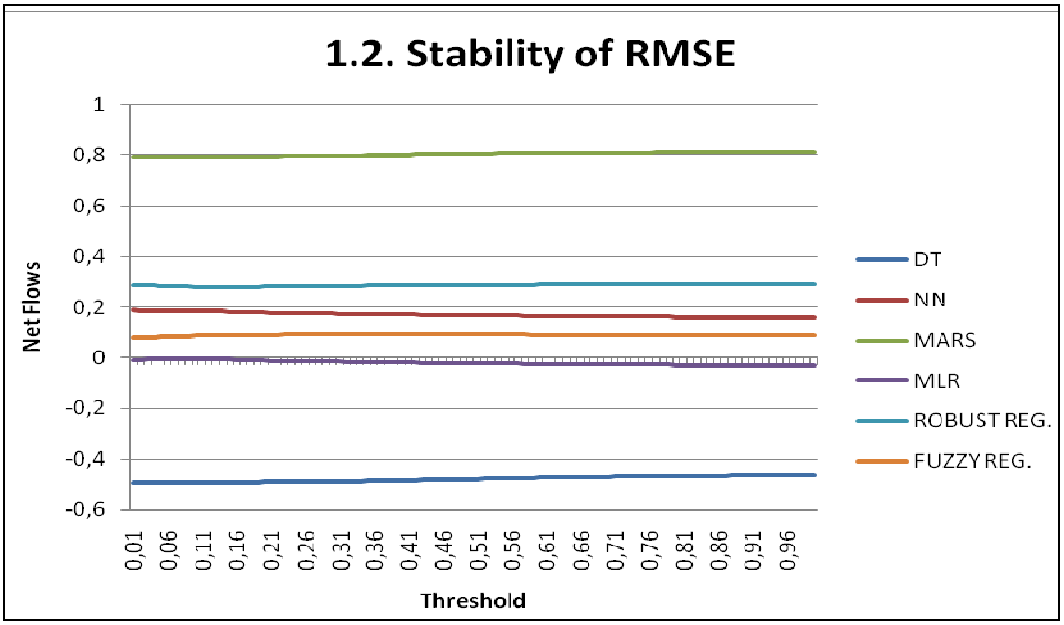



**Figure J. 49 Sensitivity of the net flows with respect to change in the threshold of the Stability of RMSE**

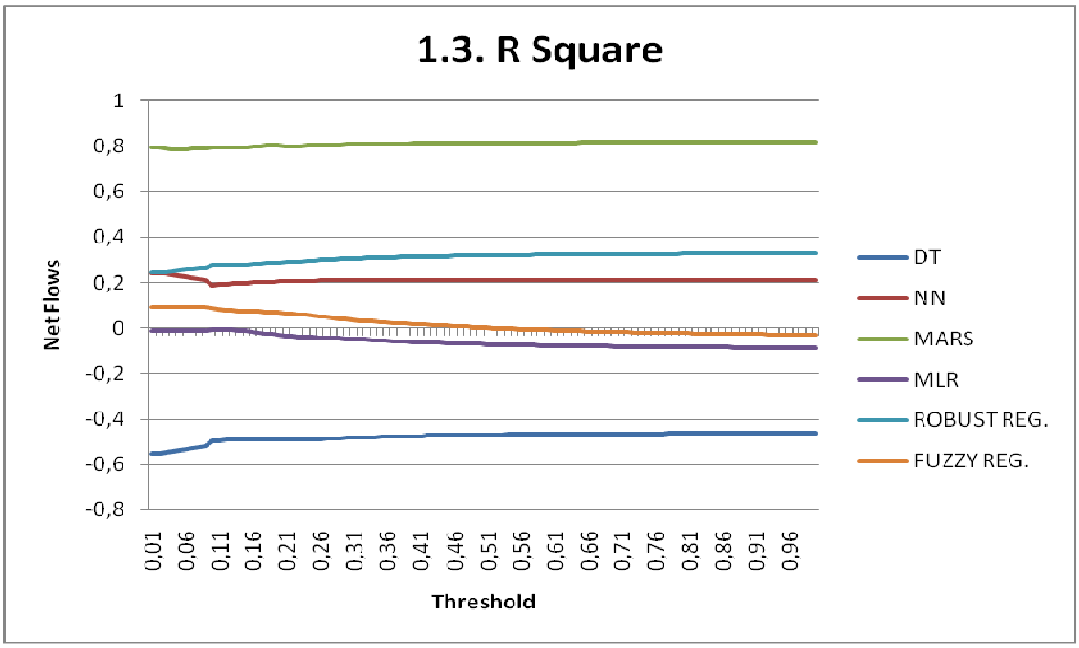



**Figure J. 50 Sensitivity of the net flows with respect to change in the threshold of the R Square**