# WHAT PREDICTS EXAM SCORES IN SCIENCE CLASSES WITH ONLINE PRACTICE SUPPLEMENT AT SECONDARY EDUCATION: A LEARNING ANALYTICS STUDY VIA CHAID

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

BY

CEYLAN NALÇA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER EDUCATION AND INSTRUCTIONAL TECHNOLOGY

FEBRUARY 2021

Approval of the thesis:

**WHAT PREDICTS EXAM SCORES IN SCIENCE CLASSES WITH ONLINE PRACTICE SUPPLEMENT AT SECONDARY EDUCATION: A LEARNING ANALYTICS STUDY VIA CHAID**

submitted by **CEYLAN NALÇA** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Education and Instructional Technology, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Dr. Hasan Karaaslan
Head of the Department, **Computer Education and Instructional Technology** _____

Prof. Dr. Soner Yıldırım
Supervisor, **Computer Education and Instructional Technology, METU** _____

Assoc. Prof. Dr. Sacip Toker
Co-Supervisor, **Information System Engineering, Atılım University** _____

**Examining Committee Members:**

Prof. Dr. Soner Yıldırım
Computer Education and Instructional Technology, METU _____

Assoc. Prof. Dr. Müge Adnan
Computer Education and Instructional Technology, Muğla Sıtkı Koçman University _____

Asst. Prof. Dr. Cengiz Savaş Aşkun
Computer Education and Instructional Technology, METU _____

Date: 15.02.2021

**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 : Ceylan Nalça

Signature :

## ABSTRACT

### WHAT PREDICTS EXAM SCORES IN SCIENCE CLASSES WITH ONLINE PRACTICE SUPPLEMENT AT SECONDARY EDUCATION: A LEARNING ANALYTICS STUDY VIA CHAID

Nalça, Ceylan

Master of Science, Computer Education and Instructional Technology

Supervisor: Prof. Dr. Soner Yıldırım

Co-Supervisor: Assoc. Prof. Dr. Sacip Toker

February 2021, # 87

In this study, within the scope of Learning Analytics; in a determined Learning Management System, the performance of the first semester Science exam of 6491 students studying in the $5^{th}$, $6^{th}$, $7^{th}$ and $8^{th}$ grades among different private schools in Turkey and the variables that affect the exam performance are predicted. For the analysis of the study, Chi-square Automatic Interaction Detector (CHAID) classification technique, which is one of the decision tree methods, offered the best result among the other classification methods used. According to the results of the research, it was estimated that the most important variables related to students' exam performance were the completed tests numbers and question numbers.



Keywords: Learning Analytics, Data Mining, Decision Tree, CHAID

# ÖZ

## ÇEVRİMİÇİ PRATİKLERLE DESTEKLENEN ORTAOKUL ÖĞRENCİLERİNİN FEN BİLİMLERI SINAV SONUÇLARINI ETKİLEYEN FAKTÖRLERİN İNCELENMESİ: CHAID YÖNTEMİ İLE ÖĞRENME ANALİTİĞİ ÇALIŞMASI

Nalça, Ceylan
Yüksek Lisans, Bilgisayar ve Öğretim Teknolojileri Eğitimi
Tez Yöneticisi: Prof. Dr. Soner Yıldırım
Ortak Tez Yöneticisi: Assoc. Prof. Dr. Sacip Toker

Şubat 2021, # 87

Bu çalışmada Öğrenme Analitiği kapsamında; belirlenen bir Öğrenme Yönetim Sistemi'nde Türkiye'nin farklı okullarında; 5., 6., 7. ve 8. sınıf seviyelerinde öğrenim gören 6491 öğrencinin her dönem sonu yapılan özel bir sınavının birinci yarıyıl performansları ve sınav performansını etkileyen değişkenler tahmin edilmektedir. Çalışmanın analizi için karar ağacı yöntemlerinden biri olan Chi-square Automatic Interaction Detector (CHAID) sınıflandırma tekniği, denenen diğer sınıflandırma yöntemleri arasında en iyi sonucu sunmaktadır. Bu sebeple CHAID karar ağacı yöntemi bu araştırmada kullanılmıştır. Araştırma sonuçlarına göre öğrencilerin sınav performanslarıyla ilgili en önemli değişkenlerin tamamlanan test sayısı ve soru sayısı olduğu tahmin edilmiştir.

Anahtar Kelimeler: Öğrenme Analitiği, Veri Madenciliği, Karar Ağaçları, CHAID

To my love and family…

## ACKNOWLEDGMENTS

Initially, I would like to thank my supervisor Prof. Dr. Soner Yıldırım and co-supervisor Assoc. Prof. Dr. Sacip Toker for their guidance, advice, criticism, encouragements and insight throughout the research.

I would also like to thank Assoc. Prof. Dr. Müge Adnan and Assoc. Prof. Dr. Cengiz Savaş Aşkun for their valuable contribution to my study with their evaluations as jury members of my thesis.

My love, Mehmet Berber is one of the biggest supporters on all stages of my life, and he was again the biggest supporter during the hard days of the study. Special thanks to him for his help and love.

The architects of all the achievements I have reached today is my family and especially my dear mother. I am grateful that I have them and their unrequited love. I always owe them for their unconditional support and all the facilities, they have provided to me.

I would like to thank my dear friend Ahmet Açıkgöz for helping me start the process of my thesis by supporting in the early days of my study.

## TABLE OF CONTENTS

## LIST OF TABLES

TABLES

## LIST OF FIGURES

FIGURES

# CHAPTER 1

## INTRODUCTION

Today, information and communication technologies have become the most important sources of information and data oceans have emerged with the development of computerized data collection technologies which are collecting massive data about each person's digital action. There are many different fields which take advantage of using these massive data sets such as in medicine, education, science and engineering etc. These commercial sectors are collecting and analyzing large-scale datasets to understand their market and improve every part of goods for customers (Eynon, 2013). However, that kind of data set cannot be collected without a large-scale technology. This problem enabled the emergence of a technology, that is, Big Data, which enables the accumulation of a large amount of data in a more structured way.

Big Data which is too large for standard database software to operate provides to keep massive data sets easily (Eynon, 2013). That is, BD is a repository that separates the data up to very specific translation levels and stores them longitudinally (Picciano, 2012). It is high volume, fast and diverse information assets for decision making and process automation advanced with low cost and innovative forms of information processing.

On the other hand, using huge datasets in different fields brings some issues which should be considered. These issues are taking into account ethical issues (privacy and protection of data), analyzing data logically according to needs and to guide related fields about inequalities detected via the light of Big Data processing. (Eynon, 2013). Although there are some ethical issues that should be considered, the more important issue is how these data contribute best to researchers, practitioners

and other stakeholders who are working in the field. In order to understand this, these large data sets should first be dug and processed via the Data Mining (DM) method. DM is a way of processing data and analyzing it in detail. In fact, the definition of DM is the process and analysis of huge data sets to find out valuable relationships which create considerable impact for the owner of the data to use in the field (Hand, 1998). DM has been used in banking and financial services, supermarkets in sales strategies, health care which is also one of the first significant fields of applying and developing DM methods and telecommunication companies (Kaya Keleş, 2017).

In addition to those applied fields, DM has started to occupy an important place in education because of an increase in e-learning resources and use of technological devices during the lessons (Koedinger et al., 2008). That is, due to the developments in information and communication technologies, it is possible to process very different and large amounts of data produced by learners during their activities in online learning environments (García & Secades, 2013). All these data sets emerged with the intention of analyzing and exploring how students learn so that quality of managerial decisions would be improved (Romero & Ventura, 2013). Educational Data Mining (EDM) is a type of DM concentrating on specifically educational environments, such as schools, colleges, universities, other academic areas etc. (Romero & Ventura, 2013). EDM is defined as developing methods to reveal meaningful structures and latent patterns from the data obtained from educational environments and using these methods appropriately for the purpose (Siemens & Baker, 2012).

EDM differentiates from DM methods because of the necessity to report "the multi-level hierarchy and non-independence in educational data" (Baker & Yacef, 2009, p.4). It analyzes all huge datasets about learning and education. EDM data sets can contain learners' "demographic data, individual data, collaborating data, administrative data, student affectivity" (Romero & Ventura, 2013, p.12). Educational data sets have typical characteristics. That is, it can be changed according to level of learners, the length of recorded data, the type of lesson etc.

(Romero & Ventura, 2013). According to Romero and Ventura, EDM is affected mainly from the fields of Computer Science, Statistics and Education and their interceptions emerge the relationship of DM and machine learning, computer-based learning and learning analytics (2013).

Many educators' concerns are to make sure that the lessons they apply are effectively implemented, the students learn all the achievements, the interaction during the lesson is sufficient or the needs of lessons to improve for future. Gašević et al. (2015) stated that the data gathered via LMS are rarely used logically to analyze and intervene, however, all this data is very valuable for analyzing and making inferences about each student behavior via integrating a learning support and service system. In different online tools like LMS, many data are gathered, however, the main point is to be able to process gathered data logically. In this step, Learning Analytics (LA) plays a major role in analyzing information and making future predictions.

When these concerns arise, the term of LA has emerged. LA is to collect, measure, analyze and report data on learners and learning contexts to understand and improve the learning process and learning environment (Ferguson, 2012; Siemens & Gasevic, 2012). In LA, the cycle begins with learners, they provide the collection of data via different LMS tools; then, in the light of gathered data, which are processed into metrics, enables to capture the points to be intervened (Clow, 2013). In this case, it eventually provides to determine steps that need to be taken to improve learners' learning (Clow, 2013). LA is a significant technique for education because the major goal of it is to improve learning and teaching practices (Elias, 2011).

In addition, LA has many benefits for those working in the field of education and for learners. LA results enable educators about how to take steps to improve their teaching methods (Analytics, 2018; García & Secades, 2013). That is, teachers can analyze students' grade progress, content data use and assessment results which help them to understand at-risk students and to intervene and develop teaching programs specifically for each of them (Campbell et al., 2007; García & Secades, 2013; Romero & Ventura, 2010). LA results give opportunity to students to see their

progress in courses and take precautions to improve their performance (García & Secades, 2013; Long & Siemens, 2014). Moreover, the analysis results give advice to students to change their habits in a good way (García & Secades, 2013). In addition, administrators can organize their resources according to feedback of LA and improve their curricula and educational programs for teachers (Romero & Ventura, 2010).

Many studies are held with the light of LA. Related studies about LA focuses on mostly predicting final mark of students via participation on learning based forums (Anozie & Junker, 2006; López et al., 2012; Cristóbal Romero et al., 2013), early prediction of academic success of students, predicting test scores (Feng et al., 2006), predicting students' approach and behaviors in online teaching (Akçapınar, 2016; Bruckman, 2007), predicting students' academic performance by using LMS tools in education environments (Chen et al., 2014; Galy et al., 2011; Hu et al., 2014; Huang et al., 2020; Lu et al., 2018; Oladokun et al., 2008; Schalk et al., 2011; Villagrá-Arnedo et al., 2016; Yoo & Kim, 2014), investigating the relationship between learning logs and procrastination (Yamada et al., 2017), investigating students' motivation in the context of a learning analytics (Lonn et al., 2015), analyzing logs in virtual learning environments to improve retention of at risk students (Aguiar et al., 2014; Wolff et al., 2013), applying learning analytics to improve students' engagement (Aguiar et al., 2014; Hussain et al., 2018; Lu et al., 2017). All studies show that LA can be a very effective tool for analyzing and interpreting educational data for all educational environments if issues to be considered are taken into account.

However, it has been stated that the intensive exposure of students to activities without a system can affect their performance negatively (Clark & Mayer, 2016). In this context, the studies did not provide an output on what differences in performance will be provided by variables applied at different frequencies.

The aim of this study is to determine whether there is a relationship between middle school students' platform usage data which are the frequency of logging in, duration on the system, number of tests completed, number of questions completed, and number of contents completed and their achievement in the Science final exam. In addition, analyzing how the frequency of use of these variables in the system by students affect the exam performance.

During the study, the data gathered from a platform which is a collaboration based educational sharing platform in teacher-student, teacher-teacher, student-student interaction. This educational sharing environment enables the effective implementation of Social Learning, Flipped Learning, Peer Interactive Learning, Project Based Learning methodologies.

## 1.1 Purpose of the Study

The E-learning environment has started to gain importance day by day and schools, teachers and students aim to access content in an easy way using such environments. Learning Management Systems (LMS) used for e-learning and other learning tracking systems develop themselves and become not only a content presentation platform, but also able to capture big data on student behavior and performance. While there are platforms that enable such large data to be obtained, it is very valuable for educators to be able to analyze this data correctly. In this context, learning analytics is of great importance and makes a great contribution to the logical interpretation of such data. The aim of this study is to try to predict whether there is a relationship between the results of a Science exam held at the end of the semester and the system usability of middle school students using a specified LMS system.

## 1.2 Research Questions

- How are the independent variables of login count, duration spent, login count with mobile devices, duration spent with mobile devices, answered

questions, completed tests, completed homework and opened contents related to the LMS usage on the Science exam results?

- How are students classified in terms of the science exam performance according to the independent variables of login count, duration spent, login count with mobile devices, duration spent with mobile devices, answered questions, completed tests, completed homework and opened contents regarding LMS usage?

- What is the order of importance of independent variables in classifying students in terms of the Science exam performance?

## 1.3 Significance of the Study

In this period, where data mining has gained such importance, it is understood that high accuracy values can be obtained when the data set and variables suitable for the purpose and quality of the prediction study in the field of education. In this study, it is aimed to predict middle school students' science course exam results levels with the data obtained from a LMS environment and to investigate which variables used in the prediction are effective. It is thought that the results to be obtained from this study will give an idea to education fields workers on which of the interaction data on learning activities in the e-learning environment can be effective in predicting student performance. With the ideas to be obtained here, it is expected to give valuable information for teachers' and administrators about how to intervene in students' learning to increase students' performance.

## 1.4 Limitations of the Study

- The contents in the platform used do not match the curriculum in the education system exactly. Students using this platform can access richer resources in the LMS system.

- It was assumed that the LMS system used in the study was used by students as desired.

- The data about students' course success is limited to the Science course.

- The students were chosen by convenience sampling.

- The research was limited to 6491 secondary school students using a certain LMS system in Turkey.

# CHAPTER 2

# LITERATURE REVIEW

In this section, how Big Data came into life, how these large-scale data started to be processed and the importance of Data Mining in this context will be discussed. In addition, analyzing large-scale data in the field of education and their contribution to related fields will be mentioned. In addition, the difference between Educational Data Mining and Learning Analytics and the current position of Learning Analytics will be explained. Finally, the related researches about the study and practice in e-learning will be touched upon in detail.

## 2.1 Big Data

With the development of technology, the use of technological tools started to increase rapidly. Increase of tools' usage has led to accumulation of data in many fields. That is, information and communication technologies have become the most important sources of information and massive volume of data has been produced from many different fields. Data gathering is not a new issue for the world as these massive data have been gathered since the 1990s (Daniel, 2015). Although the amount of collected data varies among countries, the results show that over the years, its acceleration has always been increasing (Manyika et al., 2011).

Big data refers to data that is too big for a normal software infrastructure to collect, and this size is growing every year. Many companies use big data by processing it. Although many people are concerned about the confidentiality of their data regarding the collection of information, the data collected has been shown to make a major contribution to their fields (Manyika et al., 2011). That is, in a certain manner this cannot be considered as a new concept. Those involved in the commercial business

areas have been gathering and merging big data sets to refine segmentation of products to clients and improve their understanding of the market for a long time (Eynon, 2013).

Big data offers many advantages. Making data accessible to similar work areas reduces processing time and using it enables different fields to discover new needs and improve their performance (Kitchin, 2014). In addition, segmentation of society via analyzing data provides a focus on specific needs for each sample which reduces risk related to human decision results (Manyika et al., 2011).

On the other hand, using huge datasets in different fields bring some issues which should be considered which are ethical issues (privacy and protection of data), analyzing data logically according to needs and to guide related fields about inequalities detected via light of Big Data processing (Eynon, 2013). Therefore, while analyzing the data, the policy should be known well in each field and safeguards should be provided for society's concerns (Rubinstein, 2013). Moreover, the infrastructure should be strong enough with workers having sufficient analytical skills to better analyze the data (Manyika et al., 2011).

Although the size of the big data is very variable, the data needs to be first analyzed in terms of three features which are Volume (amount of information), Velocity (Increasing rate of data) and Variety (Diversity of data) (Laney, 2001). In addition to defining the properties of data, the other critical issue is to apply some stages to reveal the value it provides which are collection, analysis, and visualization (Daniel, 2015).

It is obvious that big data is a tremendous topic for many areas. However, using these techniques should allow to reach meaningful research results for these fields rather than fixing obstacles such as inadequacy in terms of personal or technical source (Eynon, 2013). In this case, the DM plays an important role. It ensures that the large-scale data obtained is segmented and analyzed accurately.

## 2.2 Data Mining

Today, the desire to access ore-like information hidden in the data has increased the popularity of Data Mining (DM). DM provides an analysis of the current situation which eases the decision maker's job to make inferences for the future. In fact, the definition of DM is the process and analysis of huge data sets to find out valuable relationships which create considerable impact for the owner of the data to use in the field (Hand, 1998). DM is also known as Knowledge Discovery Databases (KDD) which enable to discover buried information in enormous volumes of data (García & Secades, 2013). DM has been used in many different fields. Some of the fields are banking and financial services, telecommunication companies, which mostly apply DM techniques to analyze customer behavior to promote their services more effectively, health care and education fields (Kaya Keleş, 2017).

The stages that create the information discovery process in the databases typically involve data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation (Brezany et al., 2004). Methods used in DM are categorized in two broad categories which are predictive methods containing classification, regression, prediction, time series analysis and descriptive methods containing clustering, summarization, association rules, sequential discovery (Daniel, 2015; Gorunescu, 2011).

As in many fields, DM has started to be used in education fields also. DM involves many techniques that allow us to understand the complex relationships in large data sets. Educational data mining (EDM), on the other hand, not only enables us to make sense of these relationships, but also brings important findings about learning and teaching. In this respect, this concept, which diverges from the raw definition of classical DM as its application to education-related data, has led to the emergence of an entirely new research area over time.

## 2.3 Educational Data Mining

With the rise of online learning systems like Massive Open Online Courses (MOOCs) and Learning Management Systems (LMS), a wide range of educational data which administrators should be in charge of has started to accumulate (Analytics, 2018). That is, students who are the touchstone of educational practices at every educational level, provide many kinds of data which are stored in large databases, such as personal information, grades, content interaction, graduate levels and courses that it passes and fails. These data stacks, in which meaningful relationships can be investigated and important information can be obtained, can be used to identify problems that cause issues in education and to improve the quality of education. Many educators' concerns are to make sure that the lessons they apply are effectively implemented, the students learn all the achievements, the interaction during the lesson is sufficient or the needs of lessons to improve for future. The data gathered via LMS are rarely used logically to analyze and intervene (Dawson et al., 2015), however, all this data is very valuable for analyzing and making inferences about each student behavior via integrating a learning support and service system. Main objective of DM in education is to improve learning performance and develop training environments and sources (Romero & Ventura, 2010). EDM is a type of DM technique which specifically focuses on the gathered from educational environments and its analysis (Liñán & Pérez, 2015).

EDM enables students to arrange their needs for sources and classes, distinguish students who use feedback or advice during the study and identify feedback types appropriate for each student group (Romero & Ventura, 2013).

EDM uses the techniques of statistics, machine learning and DM to analyze the data which is collected from educational environments. Romero and Ventura describe the educational data discovery process with 4 steps: educational environment, preprocessing, data mining and interpretation of results (2013). Moreover, the data gathered from educational environments vary in terms of formats, such as video,

audio, text, and image enable to be proceeded of data via analytical methods (Daniel, 2015).

## 2.4 Learning Analytics

The increase of online learning via Learning Management Systems (LMS) and Course Management Systems (CMS) reflects the increase of big data in education (Brown, 2011; Ferguson, 2012). Many countries aim to process big data in education by using tools such as LMS/CMS and to produce intelligent actions in this direction. However, educational communities often fail to analyze and intervene large-scale educational data in an effective and correct way (García & Secades, 2013; Inoue, 2009). Moreover, the online learning systems are not good enough to make an accurate and satisfying deduction with their report results obtained (Elias, 2011).

Recently, increasing interest and anxiety about how data can be used to improve teaching and learning has emerged the term of Learning Analytics (LA). LA is to collect, measure, analyze and report data on learners and learning contexts to understand and improve the learning process and learning environments (Ferguson, 2012; Romero et al., 2013; Siemens & Gasevic, 2012). With LA, even very large and unstructured data can be examined in detail and meaningful information can be extracted (Analytics, 2018).

LA cannot be thought of as a completely new field, in fact, it is associated with many fields, such as Personalized Adaptive Learning, Academic Analytics and mainly Educational Data Mining (Almosallam & Ouertani, 2014). Main goal of the LA is to improve learning and teaching practices by getting data results and analyzing them to predict (Bader-Natal & Lotze, 2011).

Clow (2012) describes LA in a four-step cycle. These stages are listed as follows:

- Learners
- Data
- Metrics
- Interventions

LA cycle begins with learners, they provide the collection of data via different LMS tools; then, in the light of gathered data, which are processed into metrics, enables the capture of points to be intervened. In this case, it eventually allows determining steps that need to be taken to improve learners' learning (Clow, 2013).

There are many advantages of using LA in education for administrators, teachers, and learners. LA results show what are the precautions to be taken in order to understand and meet students' needs by capturing all their actions during the use of it via any device (García & Secades, 2013). LA provides instant insight into the analysis and reporting process of feedback, personalized support services, and adjustment of content which prevent the loss of time between the receiving and processing of the data (Elias, 2011; Long & Siemens, 2014; Romero & Ventura, 2010). Thus, instead of the current evaluation methods which analyze the end of semester data to make suggestions for the development of education for the next semester, LA enables to predict and report the steps that need to be taken as a result of analyzing current and past data together.

LA comes about to empower educators to choose how to require steps to make strides on their educational strategies. (Analytics, 2018). That is, according to LA results, they can analyze the effectiveness of the course structure and develop learning models. In addition, by analyzing forum posts, course grades, log-in actions in the management system, it is possible to identify at-risk students who are likely to drop the course (Campbell et al., 2007; Romero & Ventura, 2010) so that teachers have the opportunity to intervene in advance. The analysis results help learners about how to improve their learning habits (Fidalgo-Blanco et al., 2015). In addition, presenting to the students the reports in the analysis results about the subject they are inadequate

also serves as a guide for them to improve their performance and to increase their experience (Long & Siemens, 2014). Administrations in the field of education can also see how they can guide both teachers and students by following the data analyzed with LA. In line with these results, targets can be reviewed and achieved more easily. (Romero & Ventura, 2010).

On the other hand, there are important issues to be considered which are ethical issues, privacy, time scale, competency on interpreting analyzed data (Ferguson et al., 2016). Values that form the basis of ethical practice are rarely revealed clearly or being questioned. However, the values should be considered as universal and applied in the same format and students' data should be protected in terms of privacy and not to cause vulnerability (Prinsloo & Slade, 2016). The output of the data used should be provided to the students and the necessary feedback should be given on time (Greller & Drachsler, 2012). On the other hand, the analyzed data should be interpreted by competent people who are able to decide how to use it (Drachsler & Greller, 2016). Moreover, integrating the experiences gained from the learning sciences, working with a broader data set, interacting with student perspectives are other hardships to be underlined (Ferguson, 2012).

## 2.5 Difference between Learning Analytics and Educational Data Mining

Since 2011, the name LA has also been heard and research conferences are being held about it beside EDM (Baker & Inventado, 2014). Although, there are overlapping points between EDM and LA, LA differentiates with its application methods.

While LA evaluates the system as a whole, EDM is analyzing the data down to small pieces. Furthermore, EDM conducts automatic adaptation, while LA advises teachers and students and empowers them. In addition, it is possible in the LA to make interpretations about the data, while EDM just focused on machine-oriented changes.

Table 2.1 The main difference between EDM and LA

| EDM | LA |
|---|---|
| More focused on machine result data | Including human interpretation of the data |
| Analyzing data down to small pieces | Analyzing entire data |
| Machine-oriented changes | Open to human interpretation |

## 2.6 Related Studies about Learning Analytics

In the literature, the related articles generally focus on some metrics which are demographic data of learners such as family economic status, prior academic history, gender, age, residency, interests, immigration status, school type and racial/ethnic group were used (Arnold & Pistilli, 2012; Choi et al., 2018; Hussain et al., 2018; Lonn et al., 2015; Oladokun et al., 2008; Tsai et al., 2020); using course resources like video or documents including watching, completion, numbers, navigation behaviors like pause, stop, forward and backward) (Agudo-Peregrina et al., 2012; Choi et al., 2018; Dietz-Uhler & Hurn, 2013; Er et al., 2019; Hu et al., 2014; Huang et al., 2020; Lu et al., 2017, 2018; Minaei-Bidgoli et al., 2003; Sun et al., 2019; Tsai et al., 2020); grade of learners including quiz, assessment, exam, course, assignments, homework and GPA (Akçapınar, 2016; Arnold et al., 2012; Choi et al., 2018; Dietz-Uhler & Hurn, 2013; Er et al., 2019; Hu et al., 2014; Hussain et al., 2018; Lonn et al., 2015; Lu et al., 2018; Oladokun et al., 2008; Cristóbal Romero et al., 2013; Tormey et al., 2020; Yamada et al., 2017); group or teachers discussion/posting actions including total spent time, frequency numbers, view and click data (Agudo-Peregrina et al., 2012; Akçapınar, 2016; Er et al., 2019; Fidalgo-Blanco et al., 2015; Huang et al., 2020; Hussain et al., 2018; Lu et al., 2017; Saqr et al., 2017); homework, quiz and assignment actions including completion, delaying, submission, frequency numbers, navigation behaviors like pause, stop, forward and backward (Agudo-Peregrina et al., 2012; Akçapınar, 2016; Choi et al., 2018; Er et al., 2019; Hu et al., 2014; Huang et al., 2020; Hussain et al., 2018; Lu et al., 2018;

Minaei-Bidgoli et al., 2003; Sun et al., 2019; Tempelaar, 2020; Tormey et al., 2020); general actions on LMS features including login, frequency numbers, spent time in the system (Akçapınar, 2016; Anozie & Junker, 2006; Choi et al., 2018; Feng et al., 2006; Hu et al., 2014; Huang et al., 2020; Hussain et al., 2018; Lonn et al., 2015; Minaei-Bidgoli et al., 2003; Saqr et al., 2017; Tempelaar, 2020; Tormey et al., 2020; Villagrá-Arnedo et al., 2016; Yamada et al., 2017; Yoo & Kim, 2014) and number of words, sentences, paragraph, messages, tenses on LMS communication channels (Yoo & Kim, 2014) to investigate the effect of using online education platforms in education environments.

The results of the related studies in the literature show some conclusions to make inferences about associated topics. One of them is that using online learning tools allows educators to gather both academic and behavioral data of students in lessons very quickly. This information can be very helpful in measuring students' performance in lessons. With these performance results, educators can predict the success of each student in the lessons, which may include content base performance, project-based performance, their course grades, or exam grades. Moreover, they can identify at-risk students and intervene in advance not to allow them to drop the course. Besides, using online learning system data enables educators to prepare a personalized study path for students and give them timely feedback. Furthermore, the analysis of students' performance data can be displayed to each via dashboards so that students have the opportunity to interpret their progress during the semester. However, these analyzes should be done very carefully by educators as it can cause misinterpretation by students about what steps to take.

To conclude, the vast majority of studies focused on similar metric types and investigated the effects of these metrics on students' performance. On the other hand, although some studies find some effects of integrating online learning platforms into educational environments, there were no related studies on how the frequency of students' exposure to practice through those systems would have different results on exam performance.

## 2.7 Practice in E-learning

With the technology age we are in, the education field has started to be integrated with digital environments. It is an undeniable opportunity to bring very fast and diverse educational environments to learners with technological support. Learners can easily access tests, animated, video-based, interactive contents and communication channels where they can get information about the course, they want to learn through such digital education environments. However, access to such channels in education is not the only criterion for better learning, these are only a small part of the puzzle and there are much more critical points in education to consider.

During the design of distance education environments, because of mostly focusing on latest technology features, the role of learners is ignored. While designing education environment, first how the human mind learns should be understand. Learning behavior in each field is different and without taking this into account, learning cannot be achieved completely (Clark & Mayer, 2016). This situation also affects how and how often the practices integrated into the created learning environments, the forms of feedback and communication channels should be.

It is always emphasized that more practice will provide better success. That is why, more tests are always prepared to reinforce learning. Practicing a lot reduces the time to complete (Rosenbaum et al., 2001), however, the learning speed of the learner continues at a decreasing rate in the same type of practices. The long-term benefit from practices is related to the frequency and sequence in which they are given. The frequency and way of practices are given for each field should be designed differently because of the conditions for being an expert in each field is different. The frequency of practice given in an artistic field is different and the way it is given is more challenging and effortful to learn according to more repetitive job types. Therefore, it is necessary to adjust how often practice should be given and they need to be designed to fully reflect the profession of each field.

In addition, practice is a necessary but not sufficient tool to achieve a good performance result (Ericsson, 2012). Learners should have more interactive educational environments (Clark et al., 2018). The environments in which learners are exposed to only direct subjects and tests and see their results as a result of their studies cannot take them to the next level. For this reason, the contents and practices should be arranged in a way that allows more interaction with the learner.

Furthermore, feedback should be given correctly, as well as content should be varied, and practices should be provided at the right frequency and level. Direct result feedback of the practices does not make a positive contribution to the development of the learners. Instead, it is necessary to provide explanatory feedback to learners and follow their progress accordingly (Van der Kleij et al., 2015).

In summary, being exposed to too much practice does not directly and continuously increase the success, besides, there are points that need attention. Distance education environments should be designed by paying attention to the fact that the education learners receive is created in accordance with the field they want to specialize in, as well as the variety, order, level and frequency of the content and tests provided, and the clarity of the feedback received enough to guide the learners about their deficiencies.

In this study, some standard metrics, which are mobile/web login numbers and duration, tests completion-number, spent time for tests, completed homework number and rate of completion, content opened number and the exam success level is used.

# CHAPTER 3

# METHOD

In this study, the students' data on their interactions and performances with the content, tests, homework and Kazanım Değerlendirme Uygulamaları (KDU) exams which is a type of exam offered to private schools by the used LMS for research in the first semester of 2018- 2019 academic year, were collected. The Learning Management System (LMS) is a management software that allows learners to select and enroll in the course, present the contents, make measurement and evaluation, and monitor and report user information in asynchronous (different time) or blended education. The LMS system which is used for the study is a supportive learning management system additional to the curriculum of the Ministry of Education offered to private schools as optional. Operations were performed to predict the science exam success levels of students based on these data and to determine which and how variables used in this prediction analysis were effective as a result of the prediction.

## 3.1 Data Collection Environment

The K12 level LMS platform, which is a product of a private company, is optionally sold to private schools in Turkey. Schools use this system to support lessons and also apply some of the exams provided by the system. System usage data of secondary school students in a certain time period was taken with the permission of the company.

The LMS contains different types of educational content within the scope of Social Studies, Science, Turkish, Mathematics, English lessons and different lessons for private schools.

The LMS also includes a different number of questions and small-scale essays on each subject. When the students solve these tests, correct, incorrect, blank numbers and success percentages in that test are recorded. Each course performance of the student is presented to teachers, students and parents with separate reports. Furthermore, through the group pages, students can write ideas, share messages and participate in voting with their classmates and teachers.

The research was carried out with the last data set obtained by pre-processing the anonymized system usage and the Kazanım Değerlendirme Uygulamaları (KDU) exam performance data of the study group in the first term of the 2018-2019 academic year. The data used includes system usage data of secondary level students in 186 different private schools. Kazanım Değerlendirme Uygulamaları (KDU) is a type of exam offered to private schools by the used LMS. The exam is administered to $5^{th}$, $6^{th}$, $7^{th}$ and $8^{th}$ grades twice in one academic year and it consists of the tests of the Mathematics, Science and Social Studies (Turkish Republic History of Revolution and Kemalism in the 8th grade) and English lessons. Since it is an optional exam offered to private schools, not all students can take this exam. In the study, the science lesson, which is one of the common and basic lessons of all grade levels of the students, was chosen and the performance and usage data of this lesson were drawn.

## 3.2 Participants

The research group of the study consists of 6491 students studying in 5th, 6th, 7th and 8th grades of different private secondary schools in Turkey using the selected LMS.

Table 3.1 Number of students by class levels

| Level | Total students | % |
|---|---|---|
| 5th Grade | 3985 | 61.40% |
| 6th Grade | 841 | 12.95% |
| 7th Grade | 1038 | 15.99% |
| 8th Grade | 627 | 9.65% |

## 3.3 Data Collection Procedures

System usage and exam performance data of students using the LMS were obtained from the database of the system. The necessary SQL queries were created, and the raw data was drawn in ".xlsx" format with Microsoft Excel tool. As a result of the related studies, it was determined that the metric is generally used to understand whether there is a relationship between LMS usage data and exam performance data in terms of learning. That is why, the gathered data from the system was related with researched study metrics which are KDU theta numbers, KDU levels, frequency of logging in, duration on the system, number of tests completed, number of questions completed, number of homework completed, progress on homework, number of contents completed and school id numbers.

## 3.4 Measures

All variables that can be used were renamed in the excel document and the document has been rearranged so that the analysis tool can easily understand, select and read the variables.

**All variables in raw data and their meanings are explained below:**

- **Dependent variable:** These variables were aimed to use as dependent in the study.
    - **KDU Theta:** It refers to the range of values (-5 to +5) determined for the KDU exam performance.
    - **KDU Level:** There are 4 developmental levels determined based on the level of KDU theta. These levels are called beginner, intermediate, upper-intermediate and advanced. These results show the ranges of values that determine students' competencies among their own grade levels in Turkish, Mathematics, Social and Science courses.
- **Independent Variables:**
    - **Login count:** It shows how many times students have logged into the system.
    - **Duration spent:** It shows how long students spent time in the system.
    - **Login count with mobile devices**: It shows how many times students have logged into the system via mobile devices.
    - **Duration spent with mobile devices:** It shows how much time students spent in the system via mobile devices.
    - **Answered questions:** It shows how many questions the students have answered in the system.
    - **Completed tests:** It shows how many tests students have answered in the system.
    - **Completed homework:** It shows how many homework students have completed in the system.
    - **School id:** It shows the id numbers of the schools using the system.
    - **Opened contents:** It shows the number of opened contents by the students in the system.

The data were taken separately for two academic terms. Since the KDU exam is held twice a year, it was aimed to compare each KDU theta value and /or KDU level with other data. Therefore, in the table below (Table 3.2), besides the first and second semester KDU theta values and levels, there are also students' 1st and 2nd semester LMS usage data.

Table 3.2 All variables and the distribution of their numbers by class level

| **Variables** | **Type** |
|---|---|
| Number of "KDU Theta 1st " | Continues |
| Number of "KDU Theta 2nd " | Continues |
| Number of "KDU Level 1st " | Categorical |
| Number of "KDU Level 2nd " | Categorical |
| Number of "Login count 1st " | Continues |
| Number of "Duration spent 1st " | Discrete |
| Number of "Login count with mobile devices 1st " | Continues |
| Number of "Duration spent with mobile devices 1st " | Discrete |
| Number of "Login count 2nd " | Continues |
| Number of "Duration spent 2nd " | Discrete |
| Number of "Login count with mobile devices 2nd " | Continues |
| Number of "Duration spent with mobile devices 2nd " | Discrete |
| Number of "Answered questions 1st" | Continues |
| Number of "Completed tests 1st" | Continues |
| Number of "Completed tests 2nd" | Continues |
| Number of "Answered questions 2nd" | Continues |
| Number of "Completed homework 1st" | Continues |
| Number of "Completed homework 2nd" | Continues |
| Number of "School id" | Continues |
| Number of "Opened contents 1st" | Continues |
| Number of "Opened contents 2nd" | Continues |

### 3.4.1 Data Pre-Processing

Not all data obtained from the LMS system were used in the study. First, the variable names of the raw data in the excel file are arranged and the missing values are determined. All data that would be useful for analysis were evaluated. The "School id" and "Class level" are eliminated from the data.

In addition, KDU exam outputs are presented in two ways; these are theta value (KDU Theta) which ranges from -5 to +5, and based on these theta values, the KDU levels (KDU Level) range from level 1, level 2, level 3 and level 4. The KDU Level variable has been omitted from the data set because the two variables have the same meaning, and it would be easier to interpret a continuous variable.

The variables that will provide the most reliable comparison between theta values of the KDU exam at the end of the first semester of the school were decided to be the values obtained for the same term. Therefore, the LMS usage data (shown in Table 3.3.) of the first academic term and the KDU Theta value of the first term (Dependent Variable) were determined as variables of this study.

Table 3.3 Dependent and independent variables used in the study

| **Variables Name** | **Type** | **Role** |
|---|---|---|
| KDU Theta | Continuous | Dependent |
| Login count | Continuous | Independent |
| Duration spent | Discrete | Independent |
| Login count with mobile devices | Continuous | Independent |
| Duration spent with mobile devices | Discrete | Independent |
| Answered questions | Continuous | Independent |
| Completed tests | Continuous | Independent |
| Completed homework | Continuous | Independent |
| Opened contents | Continuous | Independent |

### 3.4.2 Kazanım Değerlendirme Uygulamaları (KDU) Exam

Kazanım Değerlendirme Uygulamaları (KDU) is a type of exam offered to private schools by the used LMS. The exam is administered to $5^{th}$, $6^{th}$, $7^{th}$ and $8^{th}$ grades twice in one academic year and it consists of the tests of the Mathematics, Science and Social Studies (Turkish Republic History of Revolution and Kemalism in the 8th grade) and English lessons.

With the KDU, students' knowledge and skills are evaluated within the framework of levels, instead of numerical values such as application results, proficiency level definitions corresponding to these numerical values are given.

By determining the competency levels of the students, the learning-teaching process can be arranged according to the needs of the students and effective educational processes can be designed. In this exam, it is possible to monitor the performance and academic development of students over time by establishing a structure that enables a connection between applications. With questions representing cognitive levels such as practice at different difficulty levels, having knowledge and reasoning, it is ensured that the learning level of both each student and a class consisting of students with different proficiency levels is correctly defined. One model of "Item Response Theory" is used in the analysis of the exam results which is One-Parameter Logistic Model (OPLM). IRT is a modeling technique that tries to define the relationship between the test performance of an examinee and the hidden feature underlying the performance (Umobong, 2017). IRT assumes a very dominant variable that affects performance, and this variable value is called theta. The OPLM is a simplest model of IRT and compares the parameter (theta) that defines the latent reaction of the person to the situations encountered with the other parameters that he has difficulty with (Verhelst & Glas, 1995). With this model, the total number of questions to be used in the assessment and evaluation study is increased without increasing the number of questions that students will answer. KDU exam performance is evaluated via theta values and their level outputs. KDU theta values range from -5 to +5. According to KDU theta values, the level of students in each

lesson are decided. The KDU level ranges and their meaning are described in the Table 3.4.

Table 3.4 KDU exam level table

| KDU Level | Meaning |
|---|---|
| Level 1- Level 2 | Beginner |
| Level 2- Level 3 | Intermediate |
| Level 3- Level 4 | Upper-Intermediate |
| Level 4- Level 5 | Advanced |

## 3.5 Data Analysis

In this study, data mining decision tree algorithms CHAID analysis was used. Decision trees and CHAID analysis in this part of the study will be explained in detail.

### 3.5.1 Decision Tree Algorithm

Decision trees are a classification method that creates a model in the form of a tree structure consisting of decision nodes and leaf nodes by feature and target. Decision tree algorithm is developed by breaking the data set into small pieces. A decision node can contain one or more branches. The first node is called the root node. A decision tree can consist of both categorical and numerical data.

While input (predictor) variables (categorical or numerical) are independent variables, the output (target) variable (categorical) is dependent (Milanović & Stamenković, 2017). Algorithms developed for creating decision trees include CHAID, Exhaustive CHAID, C&R, QUEST, C5.0, SLIQ, SPRINT.

### 3.5.2 Deciding on the Model

The most popular method for predicting students' performance is classification (Shahiri et al., 2015). One of them is decision tree methods are mostly used techniques that can be applied to categorical and continuous variables and allows working with complex data sets.

In this study, IBM SPSS Modeler was preferred as the analysis tool. In order to understand which decision tree model is more appropriate to use in the study, the Auto Numeric model in Modeling section was run. Auto Numeric provides models that can give the most logical analysis results of the available data and their output. Before the data were used, the appropriate variables were selected for analysis and filtered.

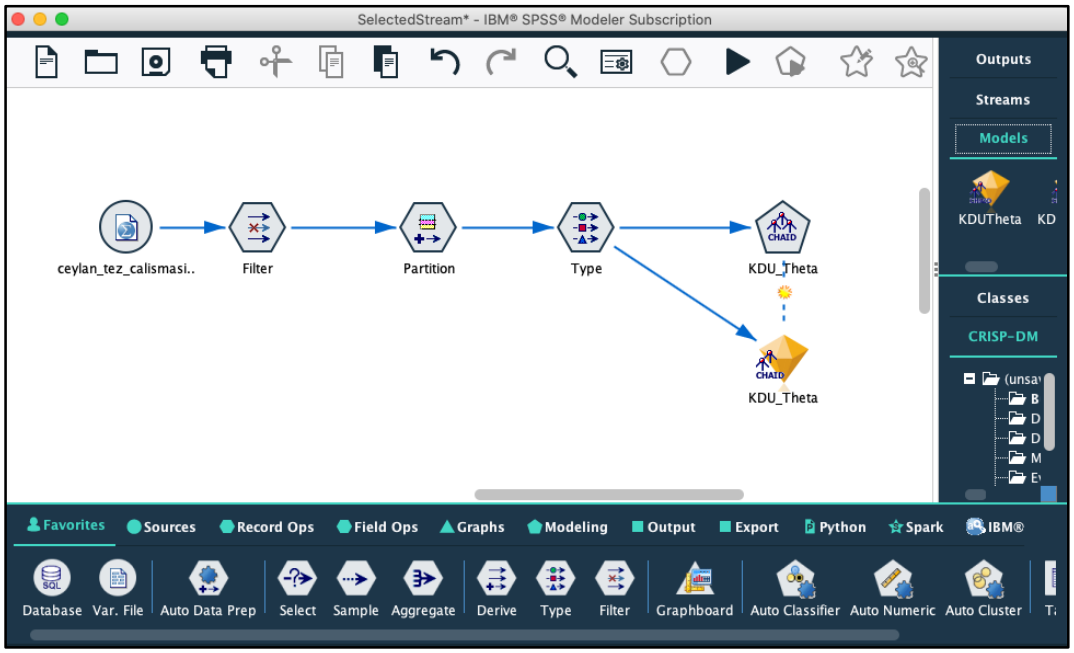


Figure 3.1 Running Auto Numeric in IBM SPSS Modeler

Auto Numeric has been run many times with different variables and the models that appear each time in the result table are displayed the most suitable models as CHAID, C&R tree, Random Trees, XGBooost Tree and Generalized Linear.

| Graph | Model | Build Time (mins) | Correlation | No. Fields Used | Relative Error |
|---|---|---|---|---|---|
|  | CHAID 1 | 2 | 0.274 | 8 | 0.935 |
|  | XGBoost Tree 1 | 2 | 0.288 | 9 | 0.925 |
|  | Random Trees 1 | < 1 | 0.295 | 9 | 0.923 |
|  | C&R Tree 1 | 2 | 0.300 | 8 | 0.913 |
|  | Generalized Linear 1 | 1 | 1.0 | 9 | ∞ |

Figure 3.2. Auto Numeric Node Estimation Model Results

### 3.5.2.1 XGBoost Model

XGBoost is a progressed execution of a slope boosting algorithm with a linear model as the beginning model. Boosting algorithm automatically learn powerless categorized and after that include them to a last solid categorized (Ji et al., 2019). That is, it is not a valid model to generate a non-binary tree structure like CHAID. When the model was run separately, it could not provide an output.

### 3.5.2.2 Random Trees Model

Random Trees is a classification and estimation system based on the methodology of the Classification and Regression Tree. As with C&R Tree, to break training records into parts with related output field values, this estimation approach utilizes

recursive partitioning (Pal, 2005). However, this model also was run separately, and it could not provide a tree model. The output was just predictor importance table.

#### 3.5.2.3 The Classification and Regression (C&R) Model

The analysis technique is evaluated according to the dependent variable handled. If the dependent variable is categorical, it is defined as the classification tree; if continuous, it is defined as regression tree (Ture et al., 2009). The C&R Tree node begins by examining the input fields to find the best split and is measured by the decrease in the pollution index caused by the split (Ture et al., 2009). The division determine two subgroups, then those are divided into two more subgroups. This continues until one of the halt criteria is activated. This model can be generated for binary splits, that is, they can be divided to just two subgroups. The difference from the CHAID model is that, by using CHAID model, the data can be divided into more than two subgroups in the decision tree. That is, CHAID can produce nonbinary trees.

#### 3.5.2.4 Chi-Squared Automatic Interaction Detector (CHAID) Model

CHAID (Chi-Squared Automatic Interaction Detector) algorithm is one of the most widely used decision tree algorithms. CHAID is an extension of AID analysis designed for categorical dependent variables (Wilkinson et al., 1992). The CHAID analysis method, which played a major role in the development of decision trees, was presented by Kass in the 1970s (Milanović & Stamenković, 2017). CHAID analysis is used when there is more complex relationship than linear relationship between variables. That is, in this analysis, the aim is to divide data into more homogeneous subgroups.

CHAID algorithm uses chi-square test if dependent variable is categorical and F test if dependent variable is continuous in branching criterion (Rokach & Maimon, 2008).

Based on the dependent variable, the algorithm gathers the values that can be considered statistically homogeneous and accepts the remaining values as heterogeneous. Afterwards, the most appropriate predictive variable is determined according to the first branch structure in the decision tree. Each node forms a group of homogeneous values belonging to the specified variable. This process continues until the divisions are over. Pruning is done automatically in order to prevent unnecessary branching in the tree created with the CHAID algorithm (Rokach & Maimon, 2008).

CHAID algorithm has become a highly preferred algorithm because it can work with all variable types categorically and continuously and it can divide each node in the tree into more than two subgroups. Variables that are continuous in the CHAID algorithm are automatically categorized in accordance with the purpose of the analysis. CHAID classifies the groups with differences according to the level of relationship using the chi-square test. Therefore, the leaves of the tree are not just binary, but as many as the number of different structures in the data.

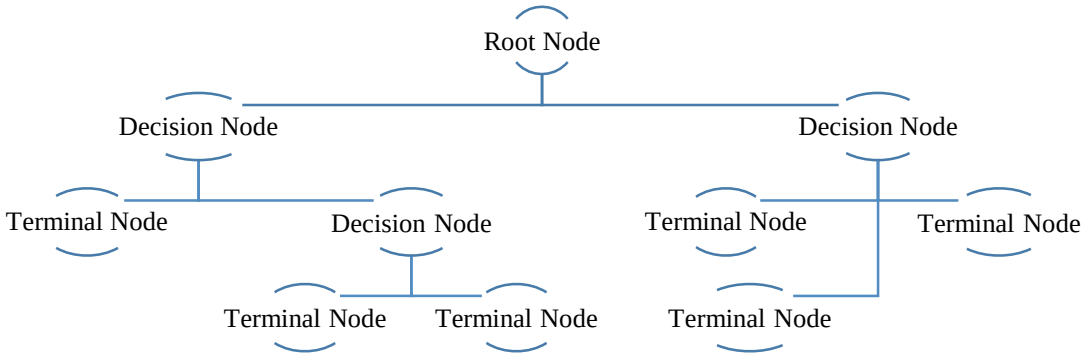


Figure 3.3. An example of a CHAID tree diagram

Missing values in the data set do not affect CHAID analysis (Rokach & Maimon, 2008). When the target is missing, the value can be discarded. When the predictor is missed, most similar values can also be grouped and generates a new category on the tree (McCormick & Salcedo, 2017).

### 3.5.3 Running the Decided Model - CHAID Model

In the study, each suitable model that came out with Auto Numeric node was rerun separately and their outputs were examined. CHAID, which is a model that allows wide-scale analysis and interpretation with its tree structure that can be divided into multiple splits, was identified as the most suitable model for this study among other models.

The KDU Level and KDU Theta values, which were thought to be dependent, were run with all variables separately and using different types of models (CHAID, C&R, Random Trees) to create a logical interpretation model. Since theta value of the two

dependent variables feeds the level variable, that is, theta value range determines the levels of KDU exam, it would be more meaningful to go above only one. In this context, when the results were compared for two dependent variables, it was determined that a continuous variable instead of a category was more meaningful to interpret the data. At the end, KDU Theta variable as a dependent variable and 2018-2019 first academic term LMS usage data as independent variables were used.

The data were divided into 3 groups with partition node in IBM SPSS Modeler program.

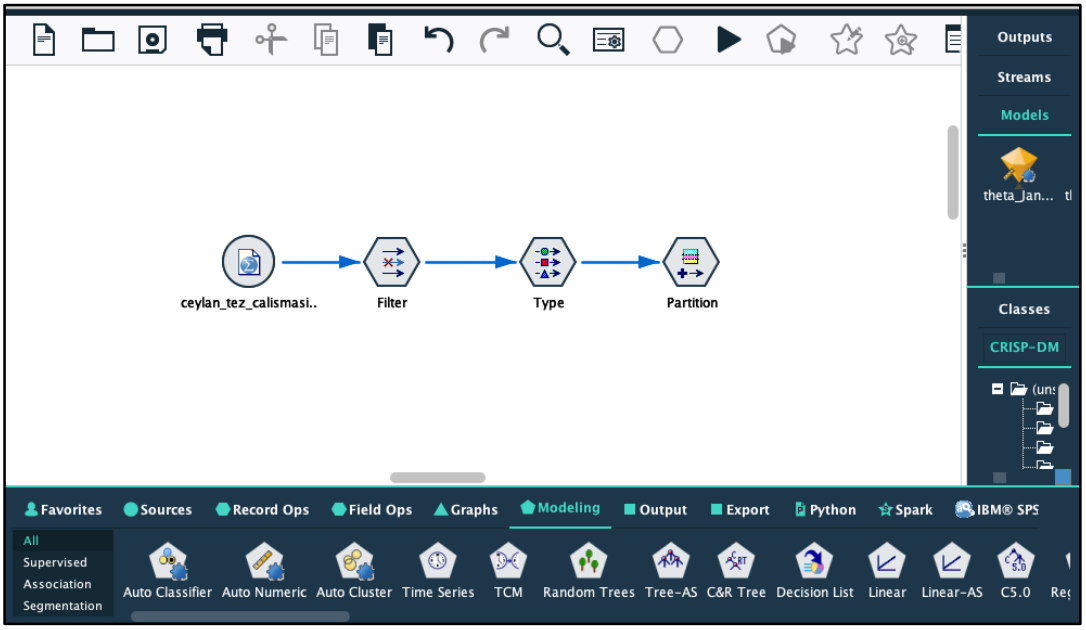


Figure 3.4. Running partition node in IBM SPSS Modeler

These are distributed as 60% Training, 20% Testing and 20% Validation.

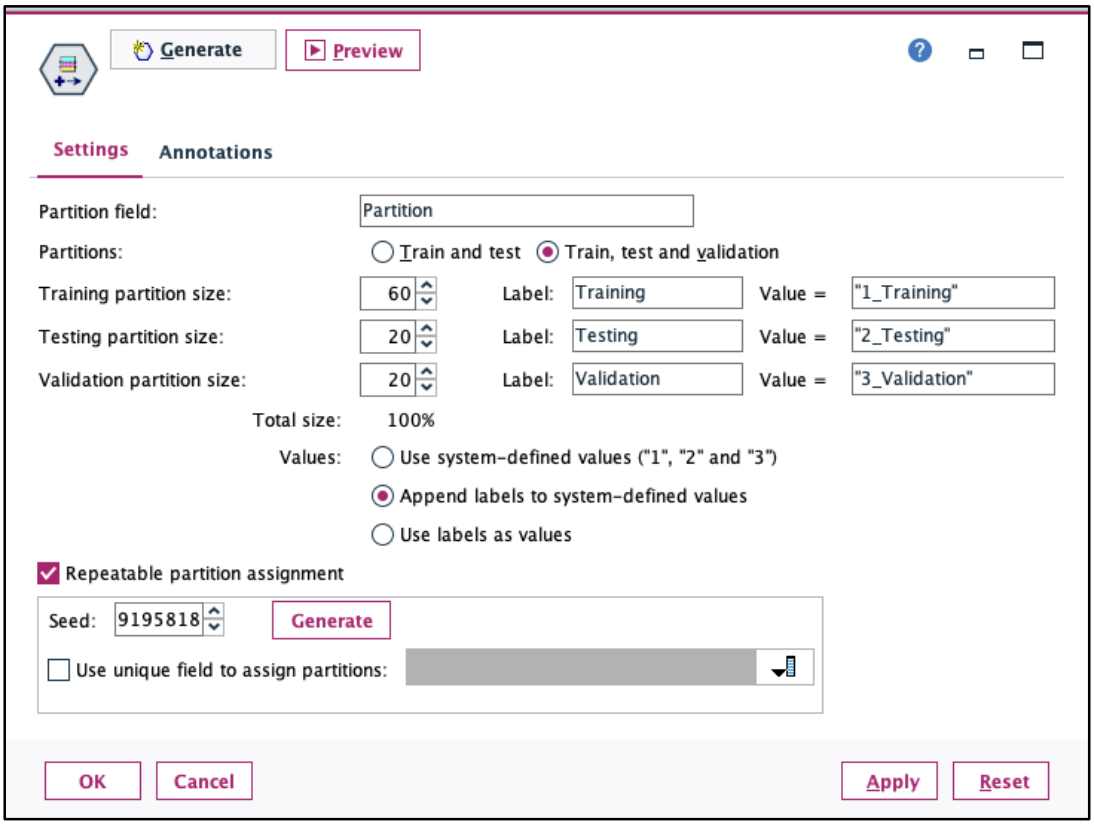


Figure 3.5. Dividing data into groups via Partition node

The reason to use the partition node structure is to evaluate the validity and accuracy of the data, which is a significant step to be able to have rational interpreted classification tree (Shafique & Hato, 2015) First, the data is taught to the system with the Training group, that is, the system is fed with the logic of data distribution. With the formulated rules learned from the Training phase, the data goes through the Testing phase again. Finally, the data that has passed the Training and Testing phases is finally verified with Validation part. Validation shows the generalizability of the established model to the population. When the test data is passed through the model processor, the accuracy value drops slightly, if the accuracy value decreases widely, it may be an indication that the model is overfitting in the training data. If the accuracy drops only slightly, this is proof that the model will work well in the future (Pasteur & Koch, 1941).

The results showed consistent validity among the 3 separated groups which are Training, Testing and Validation. As it is seen in the Table 3.5, standard deviation of the groups is so close each other.

Table 3.5 Results for comparing $ R-KDU Theta with KDU Theta

| **Partition** | **1_Training** | **2_Testing** | **3_Validation** |
|---|---|---|---|
| Minimum Error | -1.013 | -0.975 | -1.057 |
| Maximum Error | 1.596 | 1.696 | 1.716 |
| Mean Error | -0.001 | 0.007 | 0.031 |
| Mean Absolute Error | 0.269 | 0.276 | 0.275 |
| Standard Deviation | 0.346 | 0.359 | 0.359 |
| Linear Correlation | 0.365 | 0.262 | 0.29 |
| Occurrences | 3,932 | 1,282 | 1,277 |

The testing, training and validation results were also shown on a gain chart, it can be seen that how the samples are distributed according to the all-sample distribution. In the Figure 3.6., with the gain chart, the consistency of the 3 groups with $ R-KDU Theta and KDU Theta value can be seen separately. The gain chart result showed that in testing, training and validation distributions, KDU Theta values were distributed closely similar with $ R-KDU Theta.

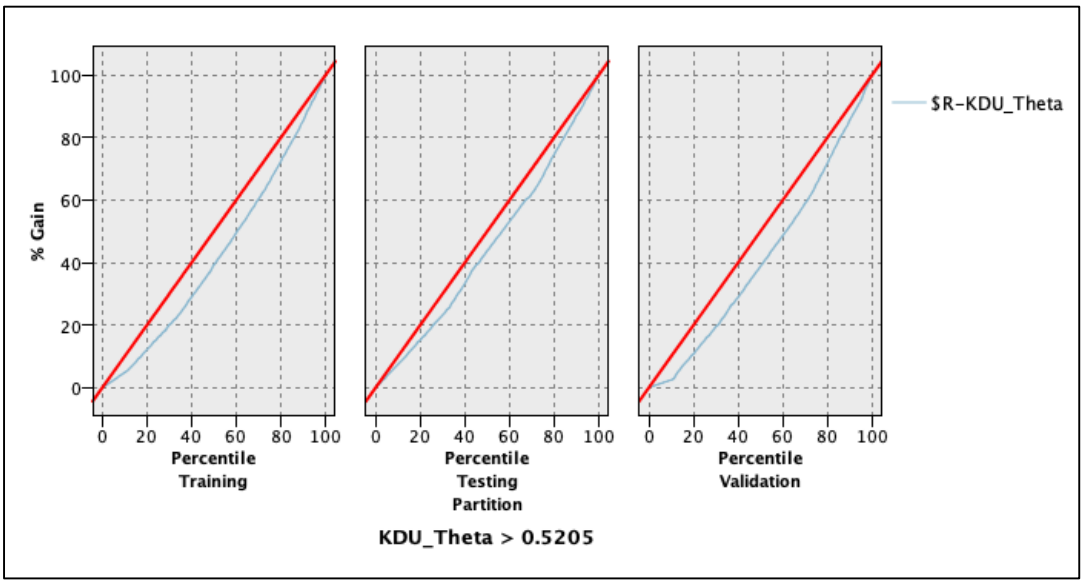


Figure 3.6. Gain chart with $ R-KDU Theta and KDU Theta

# CHAPTER 4

# RESULTS

The dependent and independent data set specified in the method section was run with the CHAID model. The variables showing distribution in the decision tree are listed in the “Predictor Importance” graph (Figure 4.1.) below according to their density. As it can be understood from the ranking in the table, "Completed tests" takes the first place, while the least effective variable is the number of entries to the system.

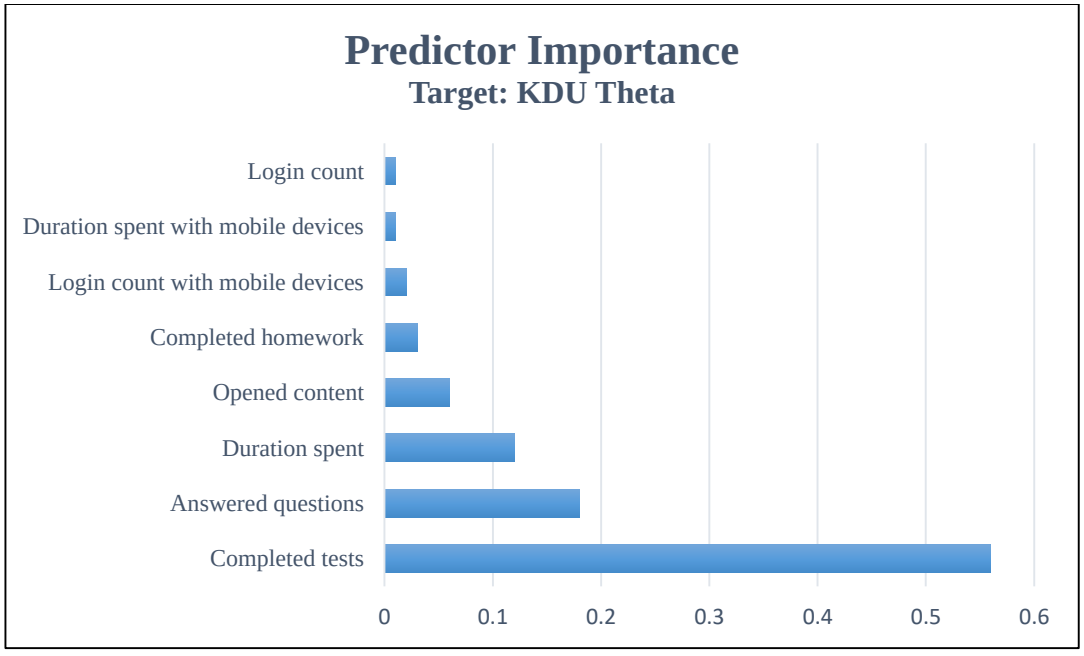


Figure 4.1. Predictor Importance Graph

After applying CHAID analysis, the relationship between relevant variables is in the form of nodes was classified. The diagram obtained is shown in Figure 4.2. As seen in the diagram, a total of 46 nodes have been formed. As a result of this analysis, among the independent variables whose effects on the dependent variable (KDU

Theta) were found to be statistically significant, the variable with the highest F value takes the first place in the CHAID diagram ($p>0.05$).

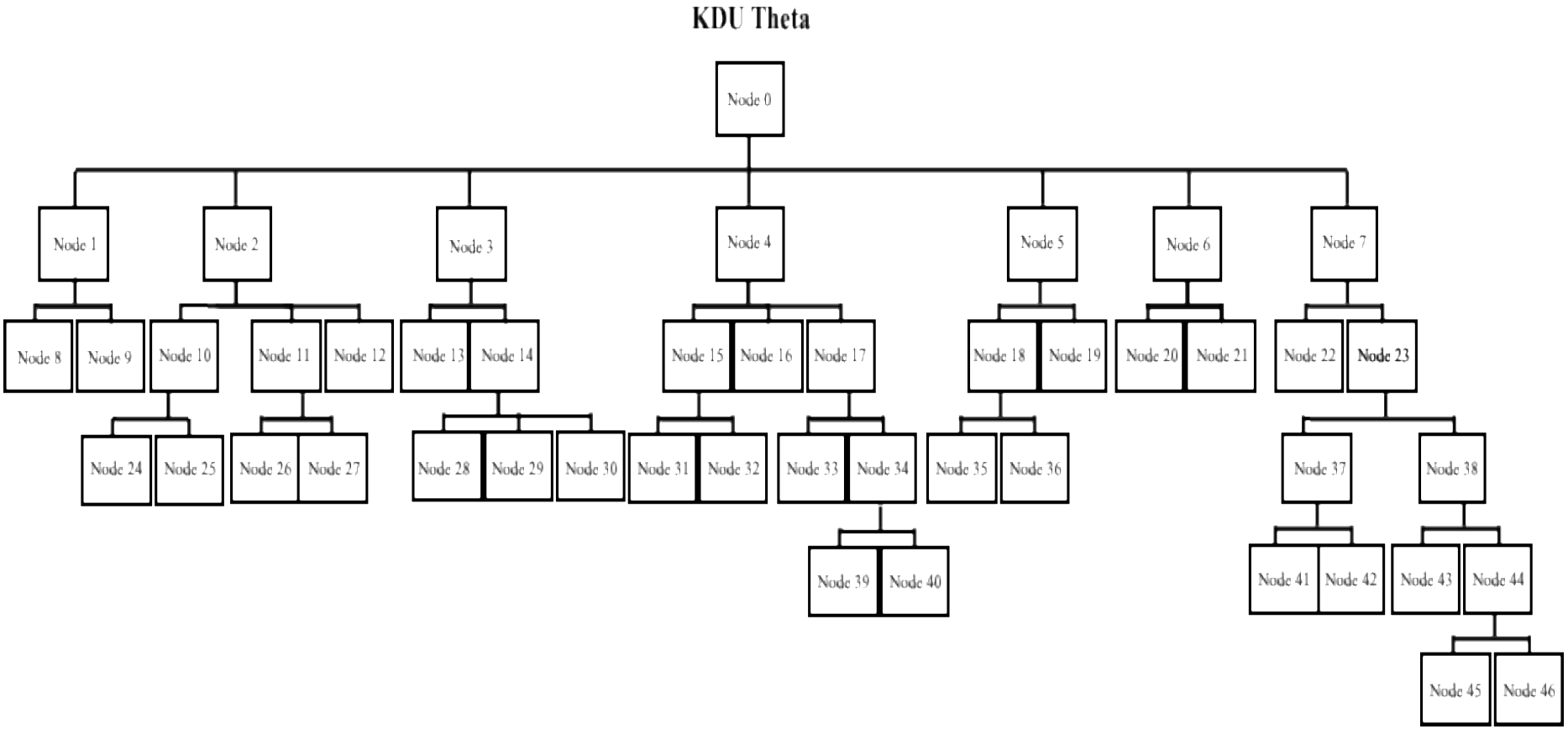


Figure 4.2. Sample Diagram of the Study Result

## 4.1 Node 0

It was found that the independent variable that affected the dependent variable the most was the Completed tests ($F_{(df1=6,\ df2=4579)}$=64.009, p=0.000) with 7 subgroups (Node 1, Node 2, Node 3, Node 4, Node 5, Node 6 and Node 7). For this reason, this variable has taken the first place in the CHAID analysis stage. Parallel to this, a total of 39 subbranches were determined in the decision tree.

Nodes were divided into branches by comparing "Predicted" values. "Predicted" value is evaluated according to the theta value range (-5, +5) taken by the students who have taken the KDU exam.

In other words, it was determined that among the dependent variable values ranging from +5 to -5, variables with predicted values close to +5 provide a positive impact on achievement of students, and variables with values close to -5 had a negative impact.

"Node X" is one of the values in each node to indicate the rank of the nodes. "n" is the total number of students in the relevant node. The "%" value indicates the percentage distribution of the number of students. "Predicted" value indicates the result of the determined dependent variable in the relevant node.

In Node 0, the whole number of students is seen and depending on the effect of the variable that appears in each sub-node, the distribution of the number of students and the effect of the KDU exam result can be observed. Except for Node 1 (P=0.363) among the first 7 nodes, Node 2 (P=0.002), Node 3 (P=0.072), Node 4 (P=0.132), Node 5 (P=0.182), Node 6 (P=0.239) and Node 7 (P=0.285) predicted values progressed to the right.

The Nodes were explained in detail from top to bottom, in a vertical manner, taking into account every branch within them.

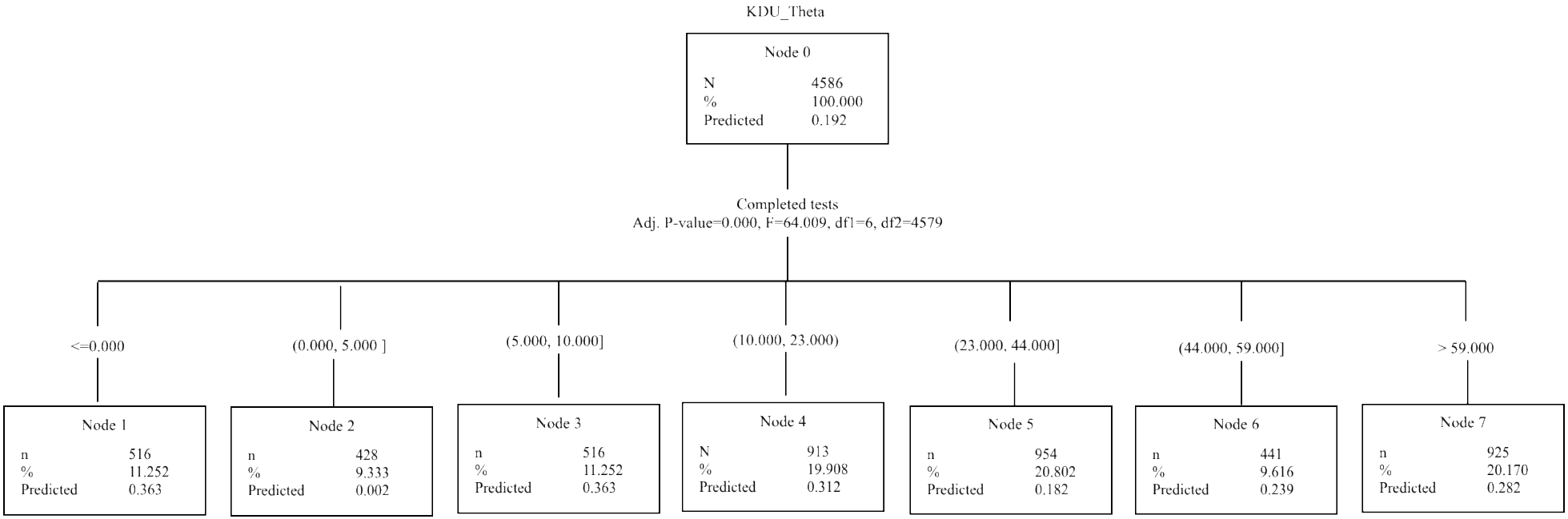


Figure 4.3. Diagram of Node 0

### 4.1.1 Node 1

The independent variable with the highest values of theta result is Node 1 (Predicted=0.363) which consists of 516 students who have never answered the test. When the F value ($F_{(df1=1,\ df2=514)}$=6.293, p=0.050) was examined, it was determined that the independent variable that best explains the cluster formed by students who take zero tests, which had a statistically significant and highest level of relationship with the dependent variable, was the duration in the system (Duration spent).

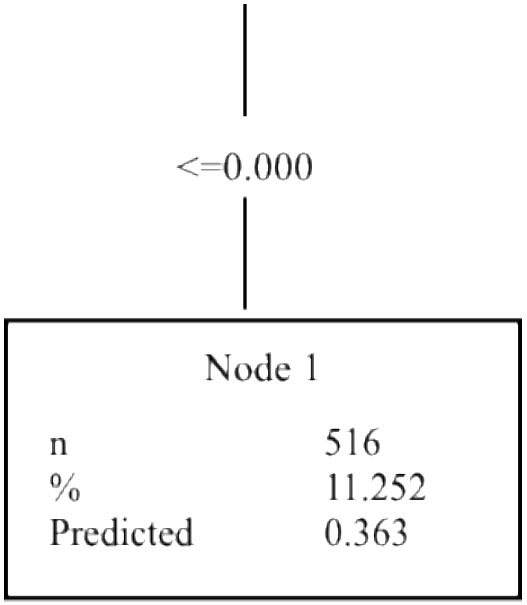


Figure 4.4. Diagram of Node 1

#### 4.1.1.1 Node 8 and 9

According to Node 1, the cut-off point for the duration of stay in the system was determined as 3 minutes, and it was observed that students (n=447, Predicted=0.345) who stayed in the system for 3 minutes or less than 3 minutes had worse theta results

(Node 8) than students (n=69, Predicted=0.486) who stayed more than 3 minutes (Node 9). At this stage, Node 1 branching has ended.

As a result, students on the 1st Node have the highest theta results even though they never took any tests. This shows that these students can be successful without integrating this kind of systems into their education.

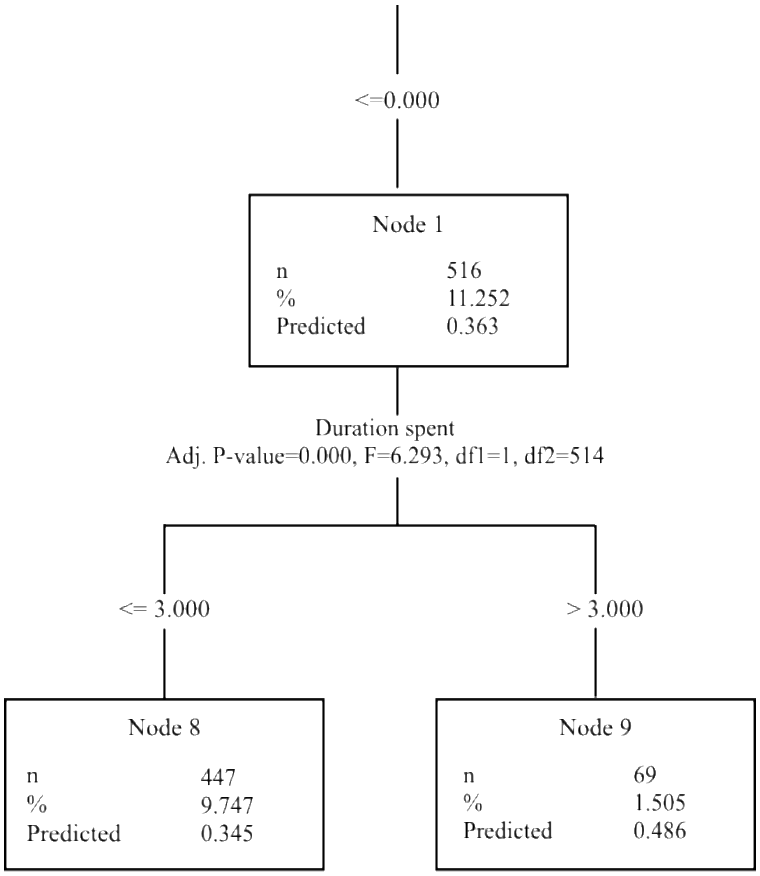


Figure 4.5. Diagram of Node 8 and Node 9

### 4.1.2 Node 2

For the Node 2 (Predicted=0.002), 428 students with completion test number up to 5 came together ($F_{(df1=2,\ df2=425)}$=16.610, p=0.000). The variable affecting the 2nd

Node was determined as content completion number of students (Opened contents) and this node has 3 different branches (Node 10, Node 11, Node 12) which are students who have completed no content (n = 216, Predicted=-0.080), students who have completed at most 8 contents (n = 117, Predicted=0.036) and students who have completed more than 8 contents (n = 95, Predicted=0.148).

As a result of the first branches of the 2nd Node, it was observed that as the number of content completion increased, the theta value also increased. Unlike Node 1, there are 2 different sub-branches for Node 2. In these sub nodes, the duration in the system via mobile devices (Duration spent with mobile devices) and login numbers (Login count) in the system were determinant.

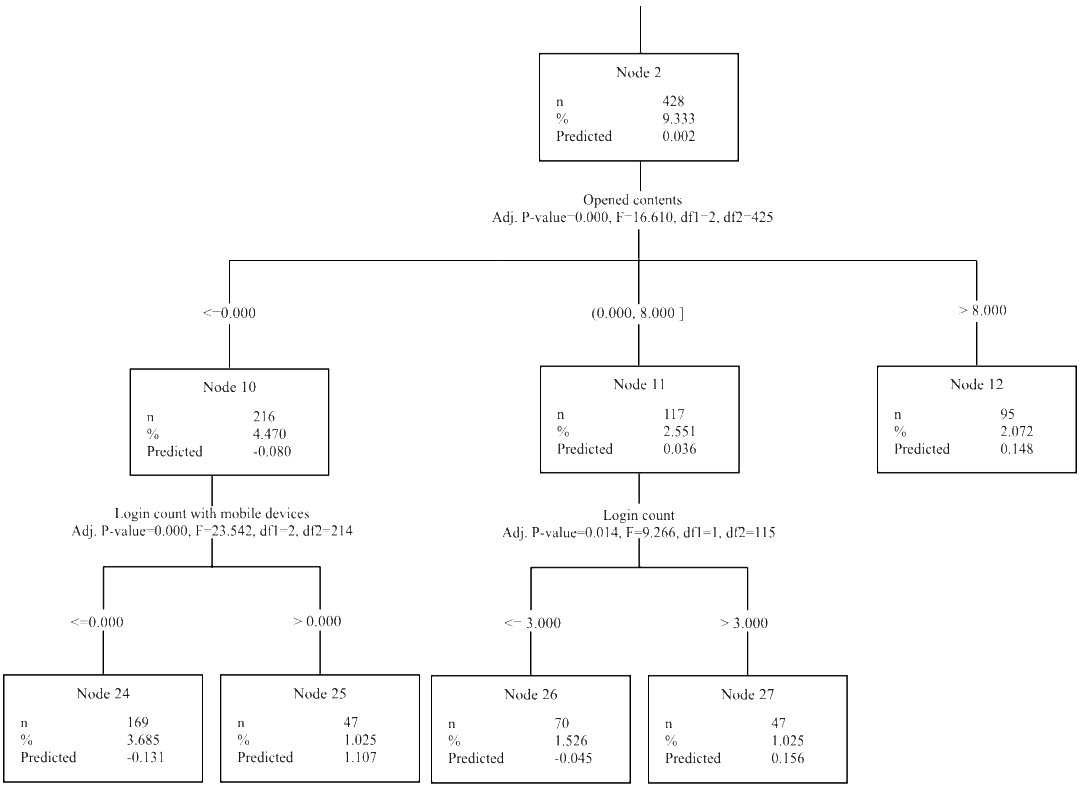


Figure 4.6. Diagram of Node 2

#### 4.1.2.1 Node 10

It was determined that among the students on the 10th node ($F_{(df1=1, df2=214)=}$ 23.542, p=0.000) who (n=216, Predicted=-0.080) did not complete any content, the students (n=169, Predicted=-0,131) who stayed in the system for at least 1 minute had higher theta results than the students (n=47, Predicted=0.107) who did not stay.

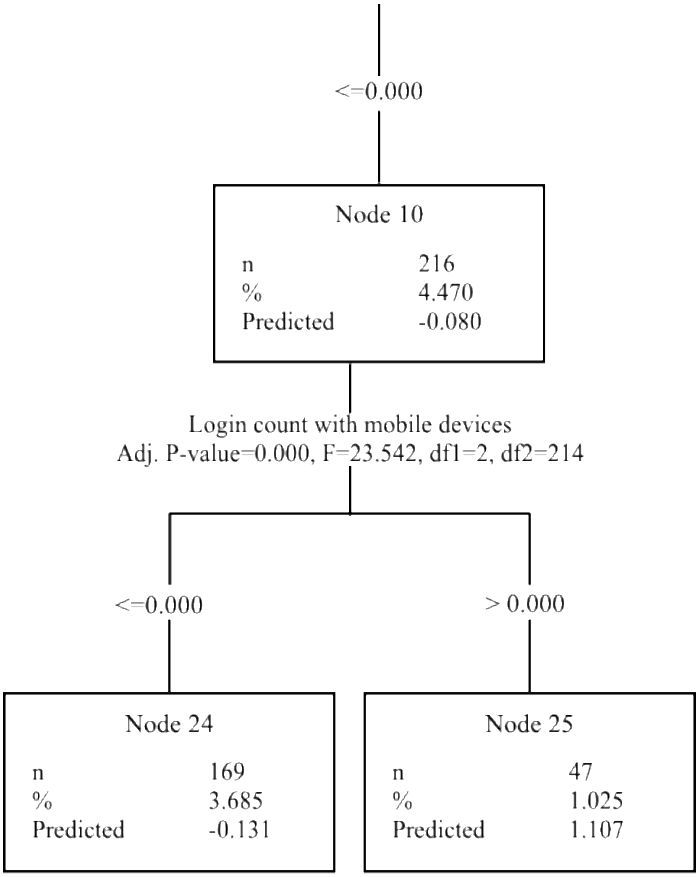


Figure 4.7. Diagram of Node 10

#### 4.1.2.2 Node 11

Moreover, in the 11th node, it was determined that among the students (n=117, Predicted=0.036) who completed at most 8 contents, the students (n=70, Predicted=-0.045) who logged into the system more than 3 times had higher theta results than the students (n=47, Predicted=0,156) who logged in at least 3 times.

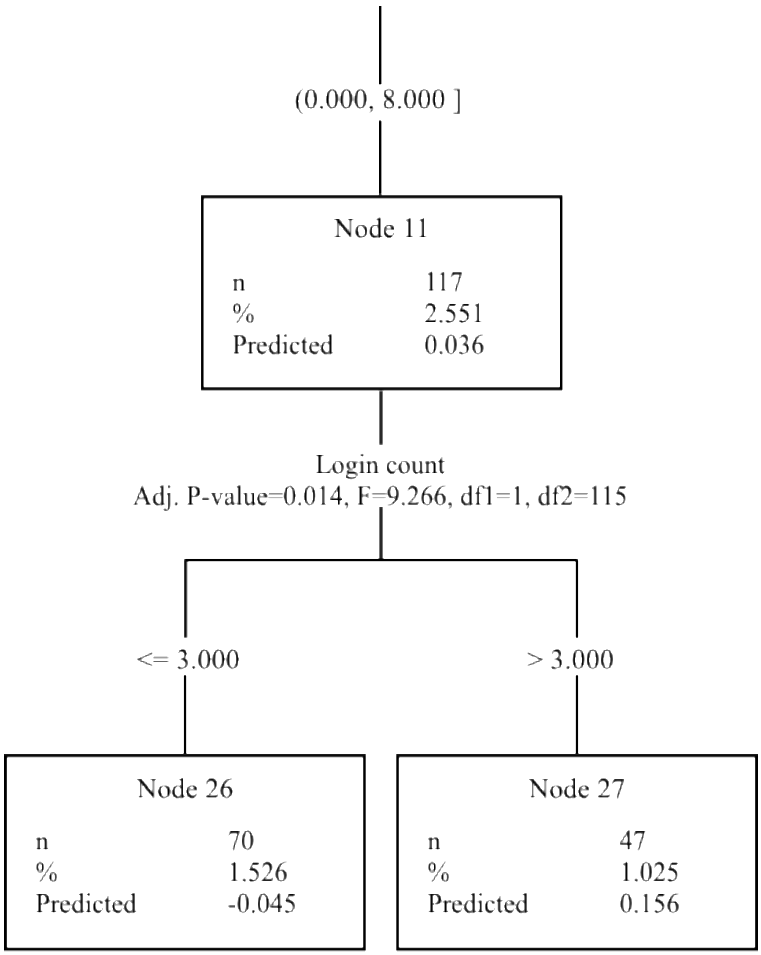


Figure 4.8. Diagram of Node 11

#### 4.1.2.3 Node 12

On Node 12, the strongest node of node 2 due to its predictor value (Predicted=0.148), students (n=95) who answered more than 8 contents had the highest theta results.

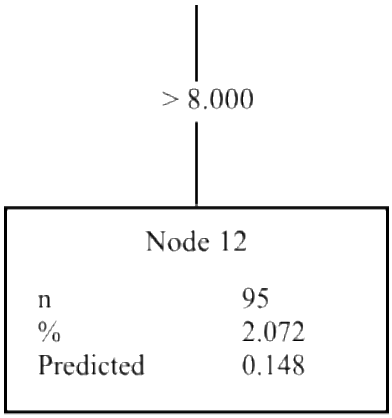


Figure 4.9. Diagram of Node 12

As a result of Node 2, of those students who completed between 0 and 6 tests, those who completed more than a certain number of contents perform better in theta results. At the same time, those who completed less content had better performances in theta results when they logged into the system more with a mobile device or from the web environment.

### 4.1.3 Node 3

Node 3 consists of students (n=409, Predicted=0.072) who have completed more than 5 and at most 10 tests ($F_{(df1=1, df2=407)}$=20.662, p=0.000). The variable affecting the 3rd Node was determined as login numbers (Login count) in the system. This node has 2 different branches (Node 13, Node 14) which are students (n=179, Predicted=-0,011) who log in to the system at most once and the students (n=230, Predicted=0.136) who log in more than once. This first division showed that students who logged in to the system more than 1 time had better theta results than those who logged in to the system once.

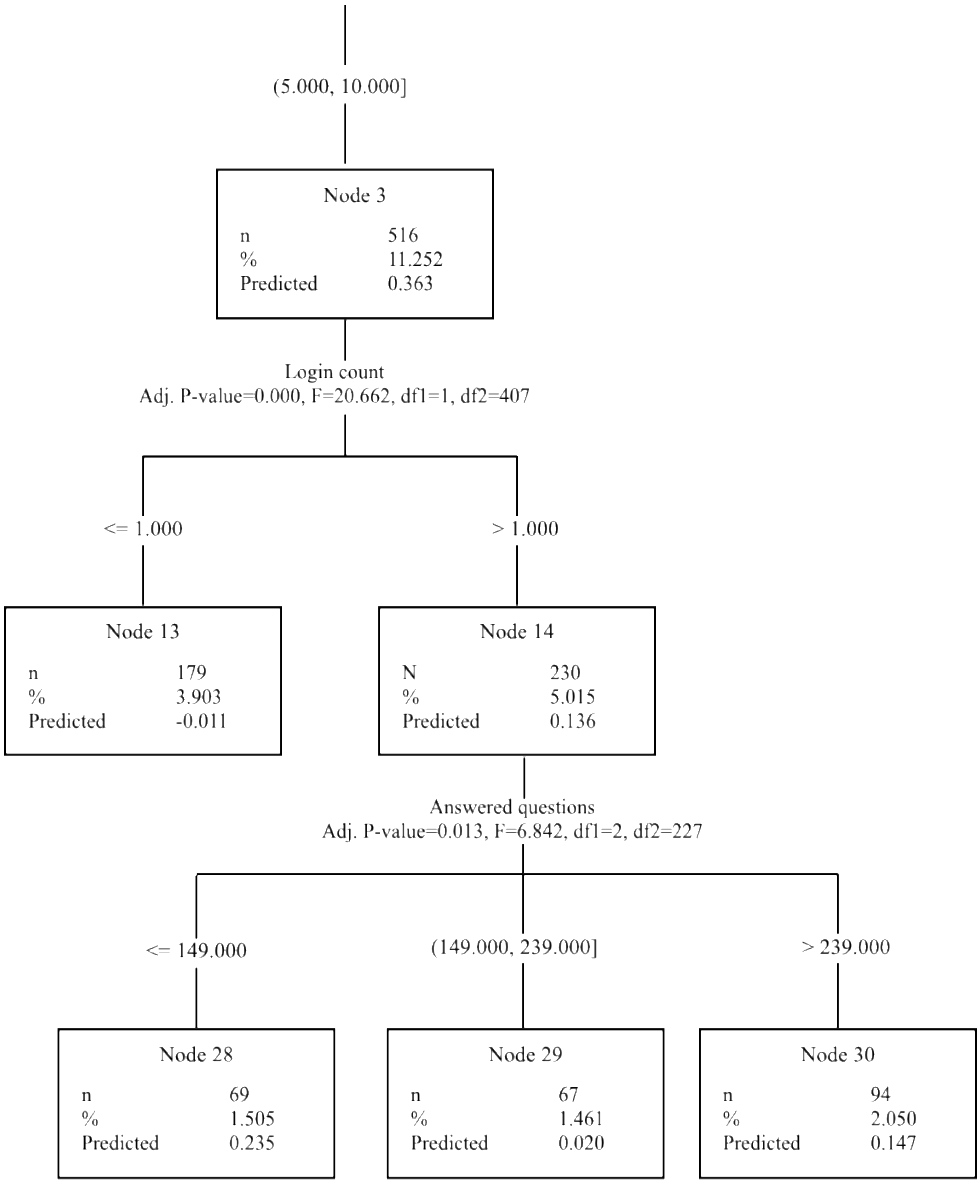


Figure 4.10. Diagram of Node 13

#### 4.1.3.1 Node 14

Node 14 was divided into 3 branches, the variable affecting its separation is the total number of questions (Answered questions) that students (n=230) have answered ($F_{(df1=2, df2=227)}$=6.842, p=0.013). In the Node 14, when these 3 branches (Node 28, Node 29 and Node 30) were compared, it was determined that students (n=69, Predicted=0.235) who answered less than 150 questions had the best theta results,

while students (n=94, Predicted=0.020) who answered more than 239 questions achieved a success in the second place. However, theta results of the students (n=67, Predicted=0.147) who answered more than 149 questions and less than 240 questions were found to be the lowest.

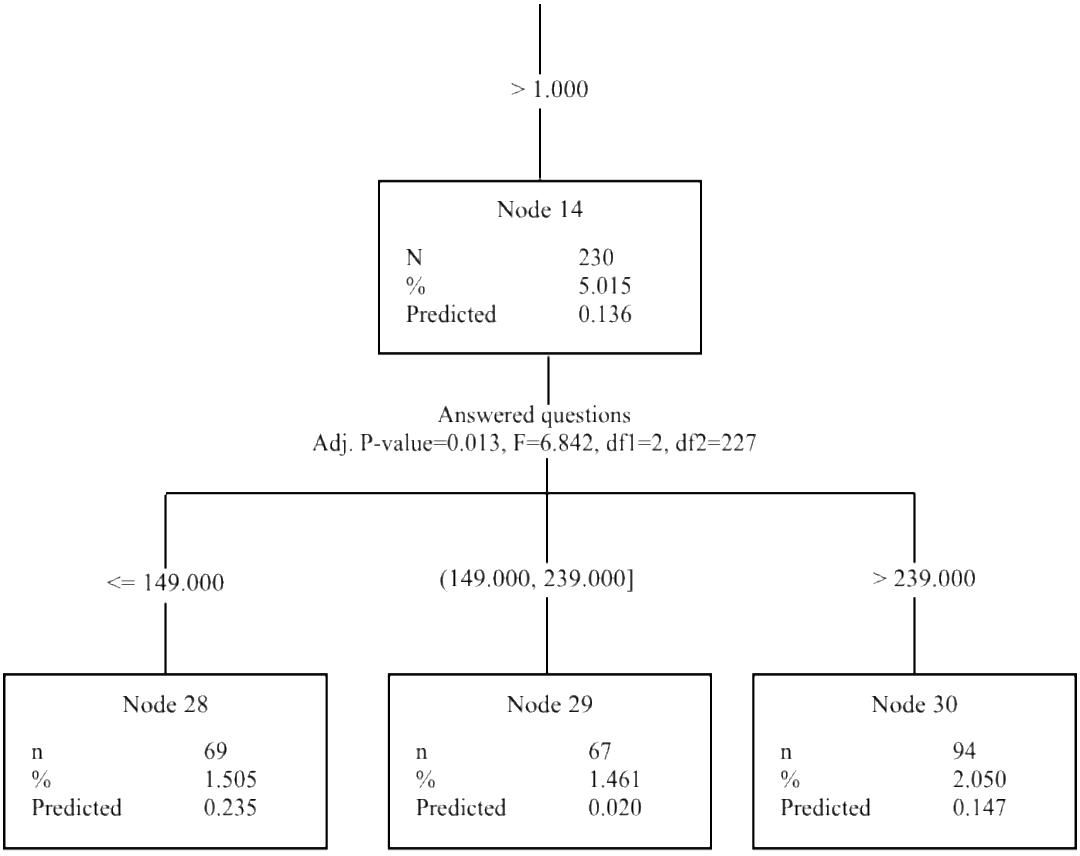


Figure 4.11. Diagram of Node14

The Node 3 results show that students who logged in to the system more than 1 time among students who completed between 6 and 10 tests achieved a better theta result, but it was determined that the increase in the total number of questions answered by students who logged in to the system more than 1 time negatively affected the theta result. That is, in addition to the positive effect of increasing the number of logins to the system by students who took tests between 5 and 11 on theta results, the reduction of the number of questions completed by those who entered the system provided even better results.

### 4.1.4 Node 4

Node 4 is a cluster of students (n=913, Predicted=0.132) who completed more than 10 tests and less than 24 tests ($F_{(df1=2,\ df2=910=)}$=23.797, p=0.000). The variable affecting the 4$^{th}$ Node is determined as answered question numbers (Completed tests). Node 4 was divided into three branches (Node 15, Node 16 and Node17) in the initial breakdown which are students (n=240, Predicted=0.250) answering less than 240 questions, students (n=347, Predicted=0.066) answering between 239 and 399 questions, and students (n=326, Predicted=0.115) answering more than 398 questions. In the Node 4, when these 3 branches were compared, it was determined that students who answered less than 240 questions had the best theta results, while students who answered more than 398 questions achieved a success in the second place. However, theta results of the students who answered more than 239 questions and less than 399 questions were found to be the lowest.

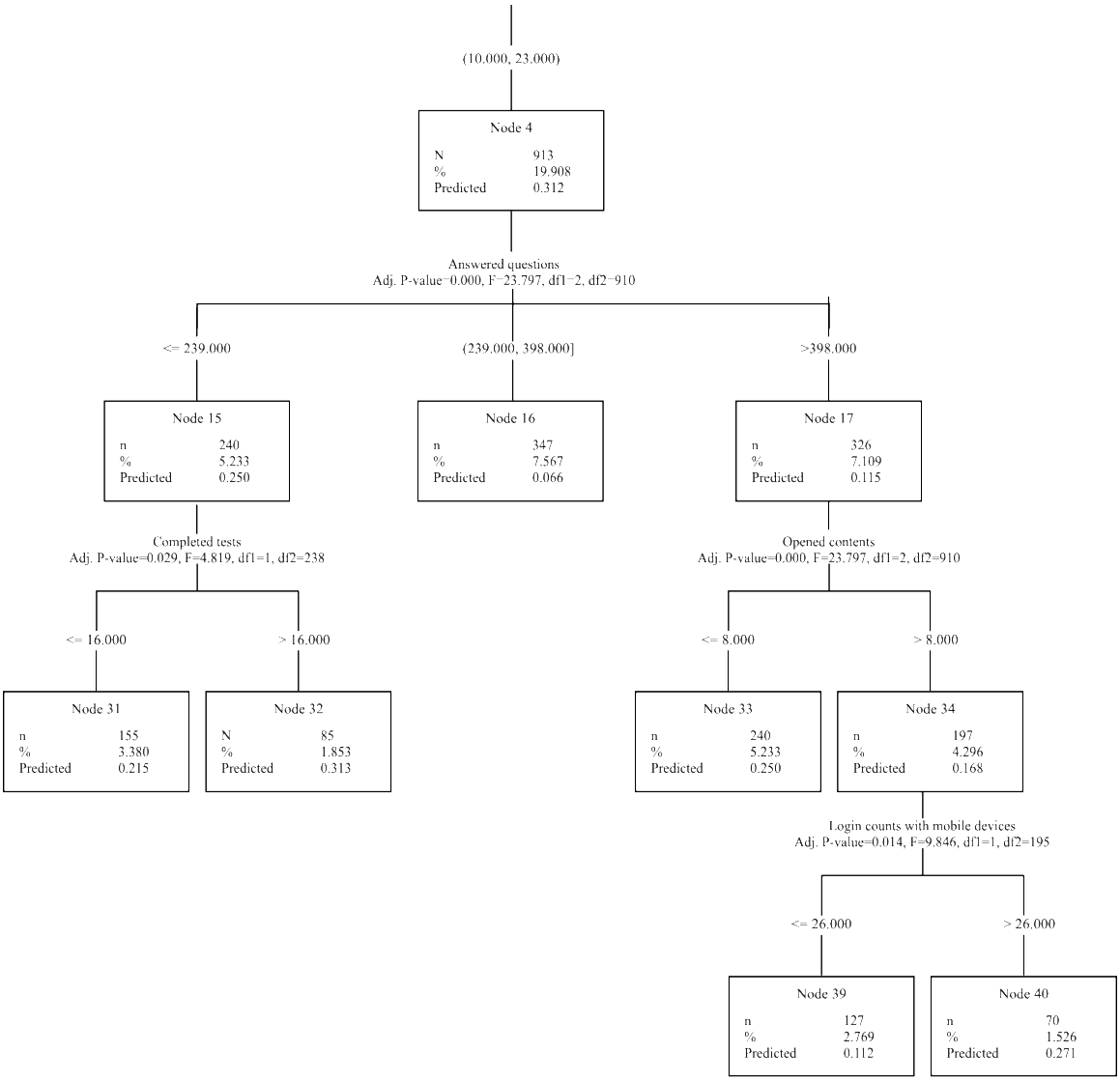


Figure 4.12. Diagram of Node 4

#### 4.1.4.1 Node 15

Node 15 ($F_{(df1=1, df2=238)}$=4.819, p=0.029) was divided into 2 subbranches (Node 31 and Node32) and the affecting variable was determined as answered test numbers (Completed tests) in the system. As a result of Node 15, students (n=85, Predicted=0.313) who took more than 16 tests had higher theta results than students (n=155, Predicted=0.215) who took less than 17 tests.

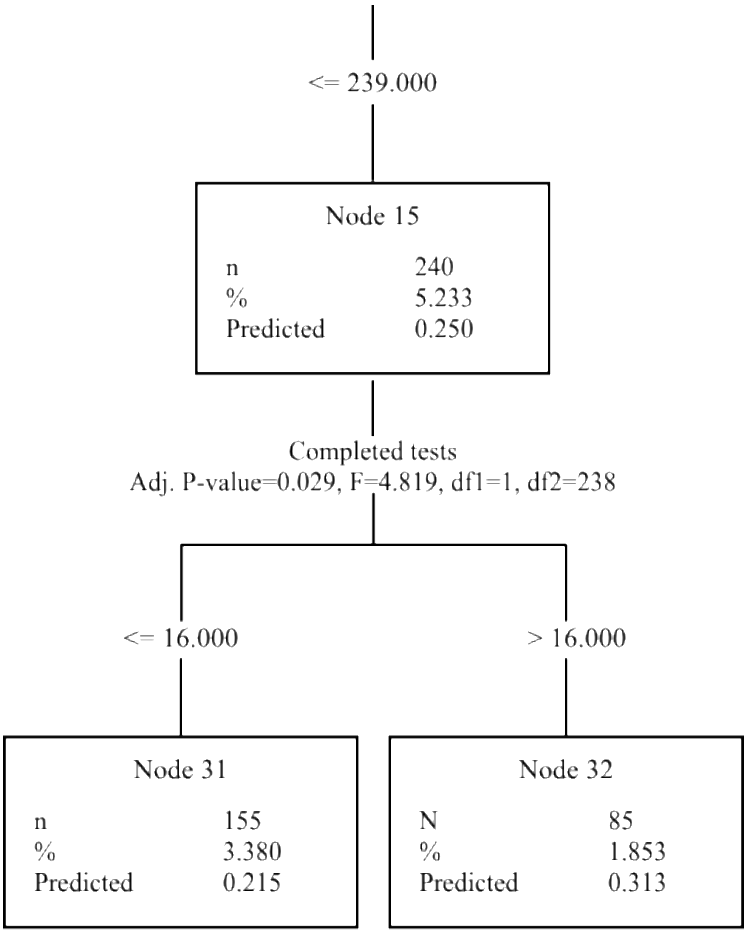


Figure 4.13. Diagram of Node 15

#### 4.1.4.2 Node 17

The variable affecting the 17th Node ($F_{(df1=1,\ df2=324)}$=13.950, p=0.002) was determined as content completion number of students (Opened contents), and this node had 2 subbranches (Node 33 and Node 34) which are students (n=129, Predicted=0.033) who have completed at most 8 contents and students who have completed more than 8 contents (n=197, Predicted=0.168). As a result of the 17th node, it was observed that as the number of content completion increased, theta value also increased.

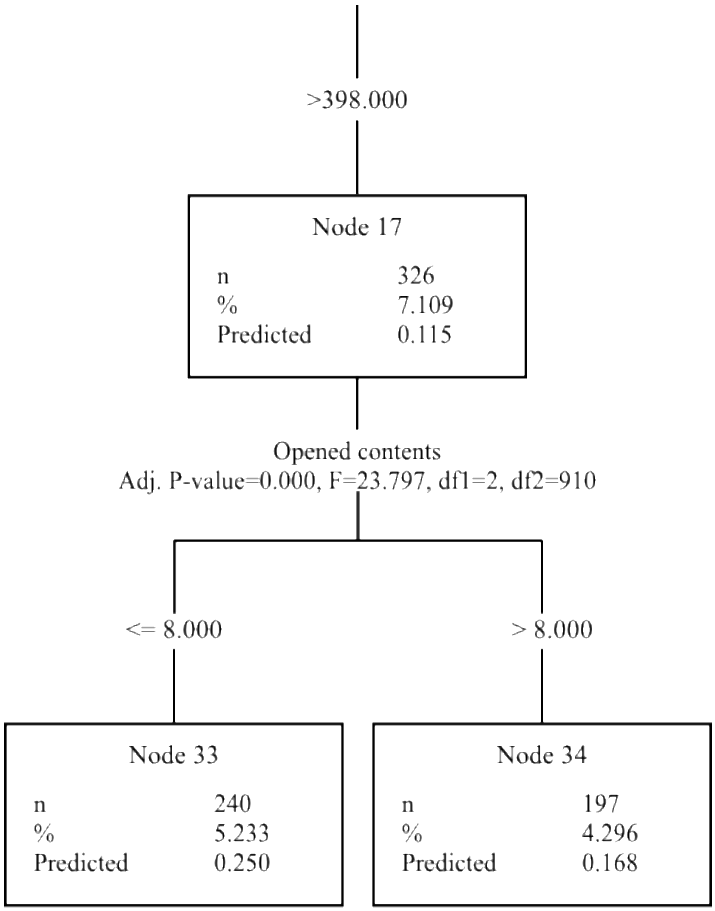


Figure 4.14. Diagram of Node 17

#### 4.1.4.2.1 Node 34

Node 34 ($F_{(df1=1, df2=195)}$=9.846, p=0.014) was split into 2 sub-branches (Node 39 and Node 40) and the login number via mobile devices (Login count with mobile devices) became the determining variable in the split. The result of the 34$^{th}$ Node shows that the theta results of the students (n=127, Predicted=0.112) who have logged into the system more than 26 times via a mobile device are higher than the theta results of the students (n=70, Predicted=0.271) who have logged into the system less than 27 times.

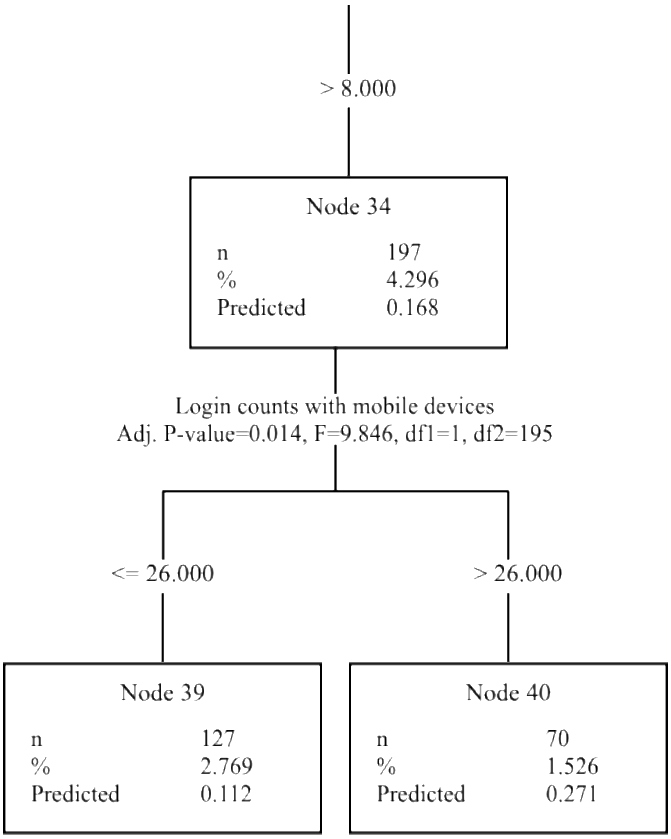


Figure 4.15. Diagram of Node 34

As a result, there is a correlation between the theta results and the number of questions answered for students distributed in Node 4. But it's not a decent relationship. In general, it can be said that answering fewer questions affects theta result better. In addition, as the total number of tests completed by students who answered fewer questions increases, theta result is positively affected. Although answering more questions negatively affects theta result, the total number of contents completed by students in this cluster more than 8 provides a positive effect in terms of theta result. Accordingly, the number of logins with mobile devices is positively affects those who complete contents more than 26.

### 4.1.5 Node 5

Node 5 consists of students (n=954, Predicted=0.182) who have completed more than 23 and at most 44 tests ($F_{(df1=1, df2=952)}$=11.176, p=0.006). The variable affecting the 5$^{th}$ Node was determined as answered question numbers (Completed tests). Node 5 was divided into two branches (Node 18 and Node 19) in the initial breakdown which are students (n= 529, Predicted=0.215) answering less than 491 questions, students (n=425, Predicted=0.141) answering more than 490 questions. With the division in the first branches, it was determined that students who completed more than 490 questions had lower theta results.

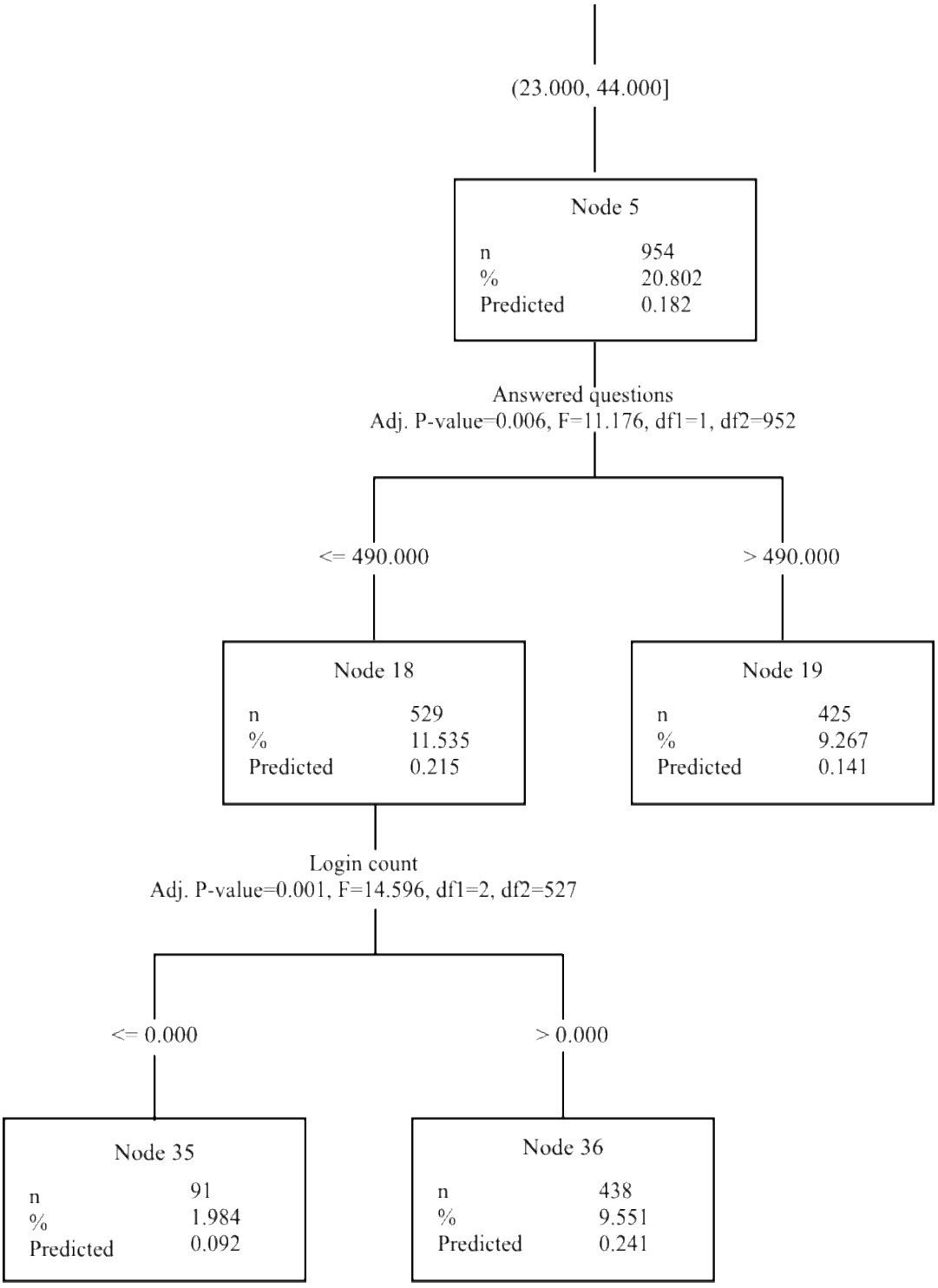


Figure 4.16. Diagram of Node 5

#### 4.1.5.1 Node 18

Node 18 ($F_{(df1=1, df2=527)}$=14.596, p=0.001) was divided into 2 branches (Node 35 and Node 36), which are students (n=91, Predicted=0.092) who had never logged into the system (Login count) and the students (n=438, Predicted=0.241) who had logged in at least once. It was determined in the 18th node that the students who had logged in at least once had higher theta results than students who had never logged into.

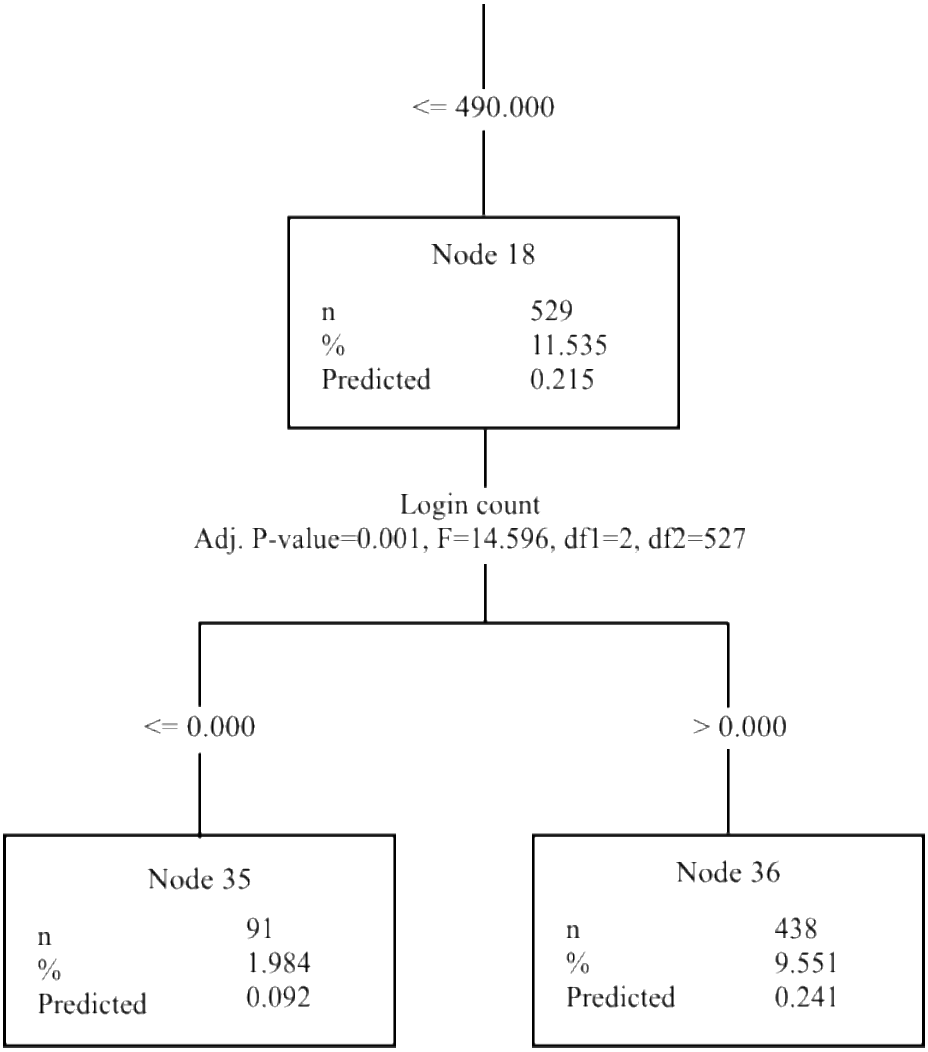


Figure 4.17. Diagram of Node 18

As a result of Node 5, the fact that the students who completed the test between 23 and 45 answered more than a certain number of questions negatively affected the theta results, on the contrary, the students who answered fewer questions in these range and logged into the system more, affected theta results even more positively.

### 4.1.6 Node 6

Node 6 consists of students (n=441, Predicted=0.239) who have completed more than 44 and at most 59 tests ($F_{(df1=1, df2=439)} = 12.899$, $p=0.003$). The variable affecting the $6^{th}$ Node is determined as completed homework numbers (Completed homework). This node has 2 different branches (Node 20, Node 21) which are students (n= 248, Predicted=0.183) who completed at most 38 homework and students (n=193, Predicted=0.310) who completed more than 38 homework. According to the Node 6 result, it was determined that among students who completed between 44 and 60 tests, students who completed more than 38 homework had better theta scores than those who completed less than 39 homework. That is, in addition to completing a certain number of tests, completing homework also contributed positively to theta results.

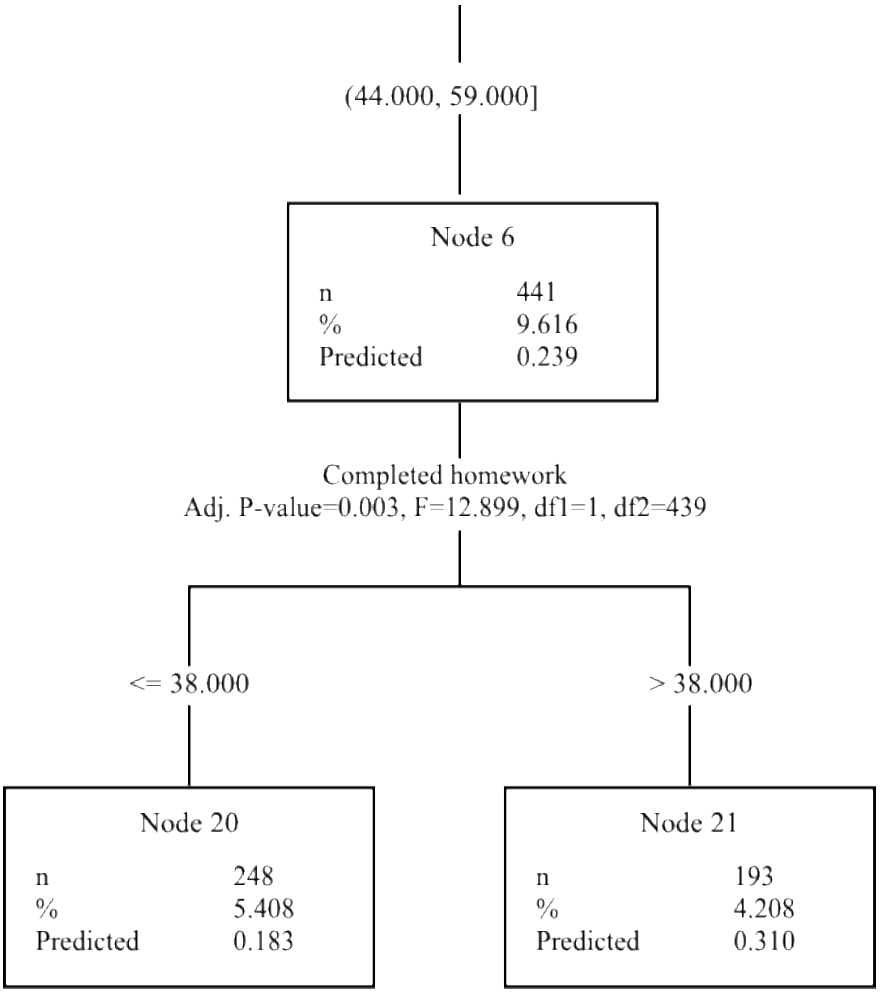


Figure 4.18. Diagram of Node 6

### 4.1.7 Node 7

Node 7 is a cluster of students (n=925, Predicted=0.285) who completed more than 59 tests ($F_{(df1=1,\ df2=923)}$=9.761, p=0.015). The variable affecting the 7$^{th}$ Node was determined as login numbers (Login count) in the system. This node has 2 different branches (Node 22 and Node 23) which are students (n=110, Predicted=0.186) who logged in to the system at most 3 times and the students (n=815, Predicted=0.299) who logged in more than three times. These branches show that students who had logged in more than 3 times (Node 23) had better theta result.

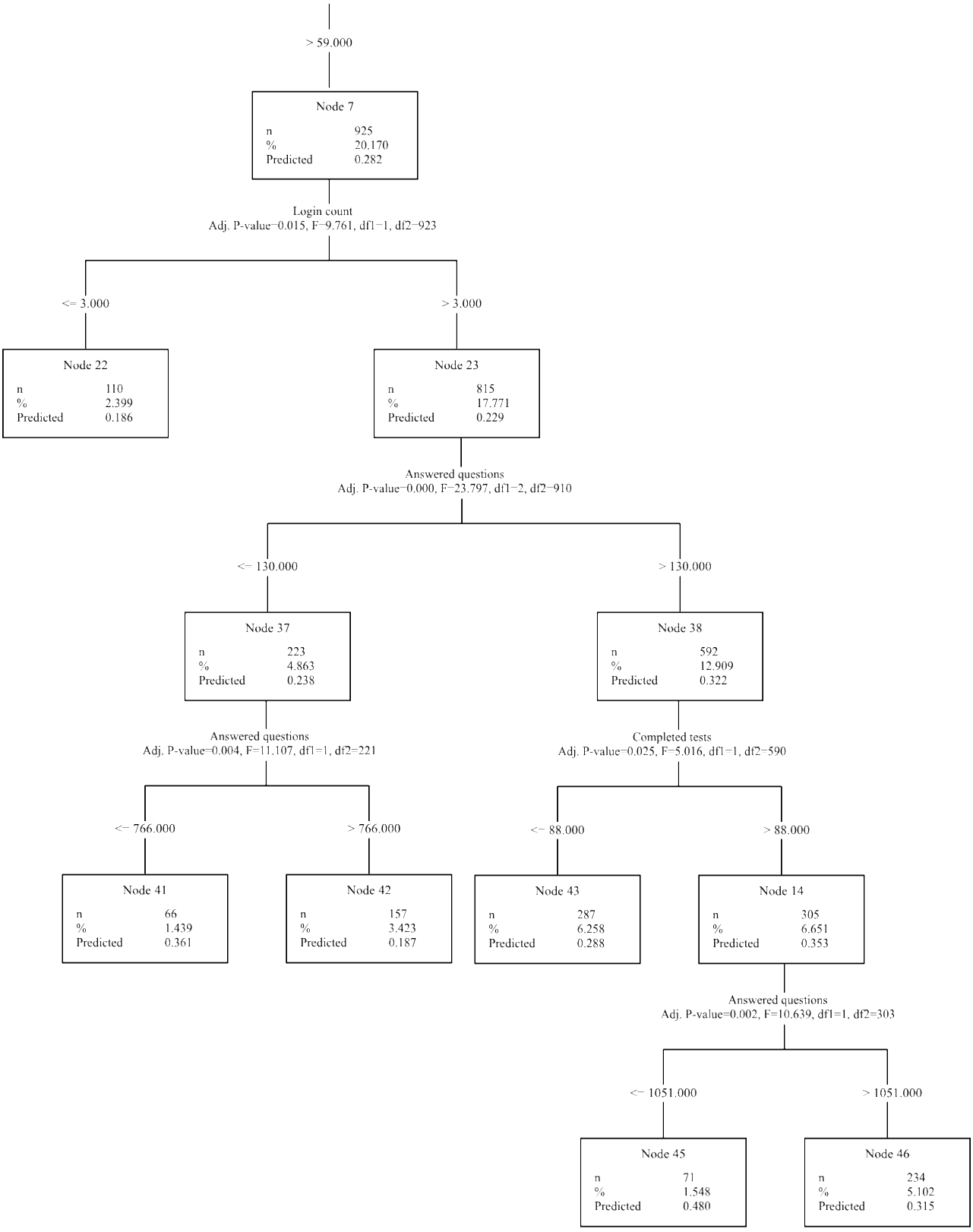


Figure 4.19. Diagram of Node 7

#### 4.1.7.1 Node 23

Node 23 ($F_{(df1=1, df2=815)}$=8.653, p=0,024) also was divided into 2 branches (Node 37 and Node 38) in terms of the duration in the system via mobile devices (Duration spent with mobile devices). Accordingly, the cut-off point for the duration of stay in the system via mobile devices was determined as 130 minutes, and it was observed that students (n=223, Predicted=0.238) who stayed in the system for 130 minutes or less than 130 minutes had lower theta results than students (n=592, Predicted=0.322) who stayed more than 130 minutes.

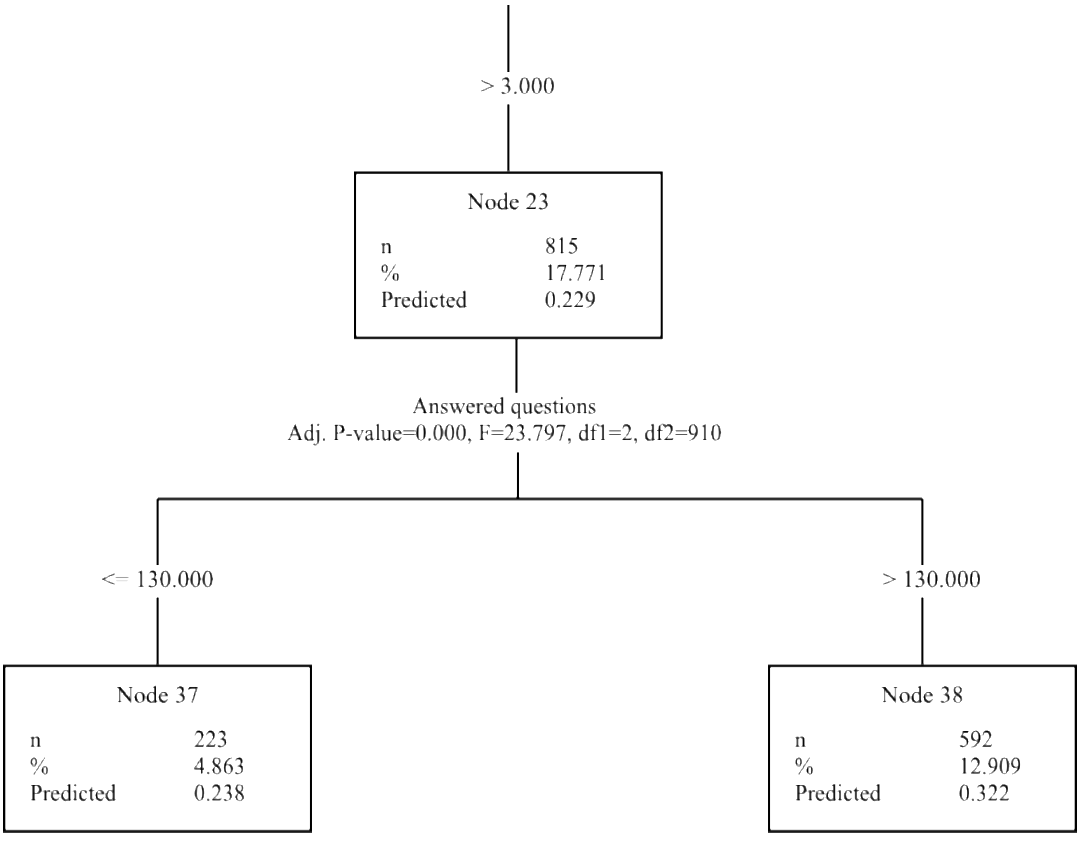


Figure 4.20. Diagram of Node 23

#### 4.1.7.2 Node 37

In addition, Node 37 ($F_{(df1=1, df2=221)}$=11.107, p=0.004) was divided into 2 branches (Node 41 and Node 42) and affecting variable of the division is total answered question number (Answered questions). According to Node 37, students (n=66, Predicted=0.361) who answered less than 767 questions had better theta results than students (n=157, Predicted=0.187) who answered more than 766 questions. Moreover, Node 38 ($F_{(df1=1, df2=590)}$=5.016, p=0.025) was divided into 2 branches (Node 43 and Node 44) and affecting variable of the division is total completed test number (Completed tests).

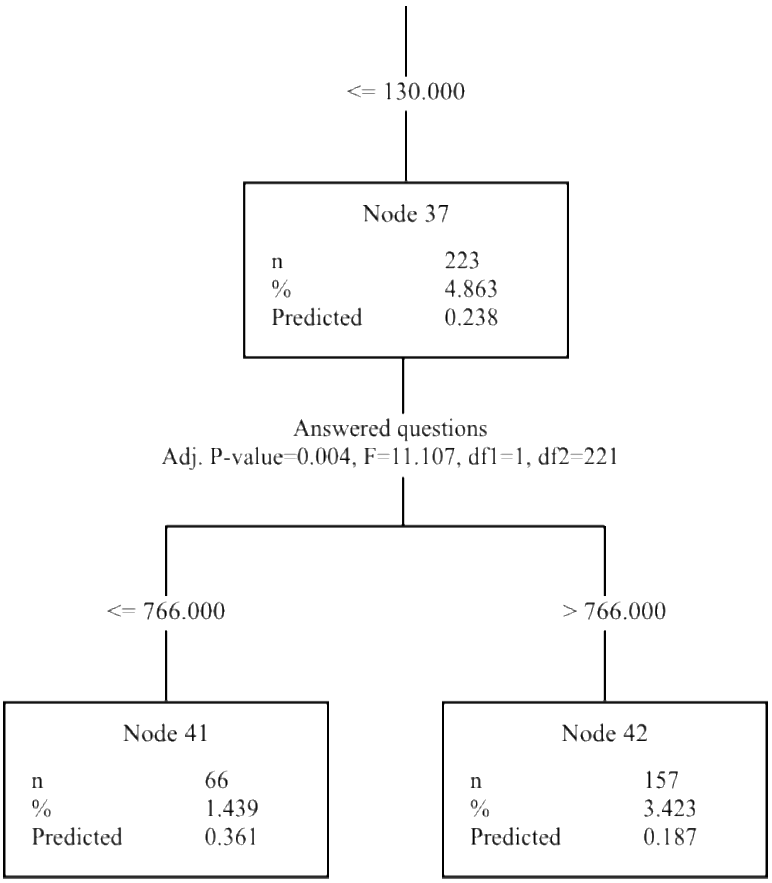


Figure 4.21. Diagram of Node 37

#### 4.1.7.3 Node 38

According to Node 38, students (n=305, Predicted=0.353) who completed more than 88 tests had better theta results than students (n=287, Predicted=0.288) who completed less than 89 tests.

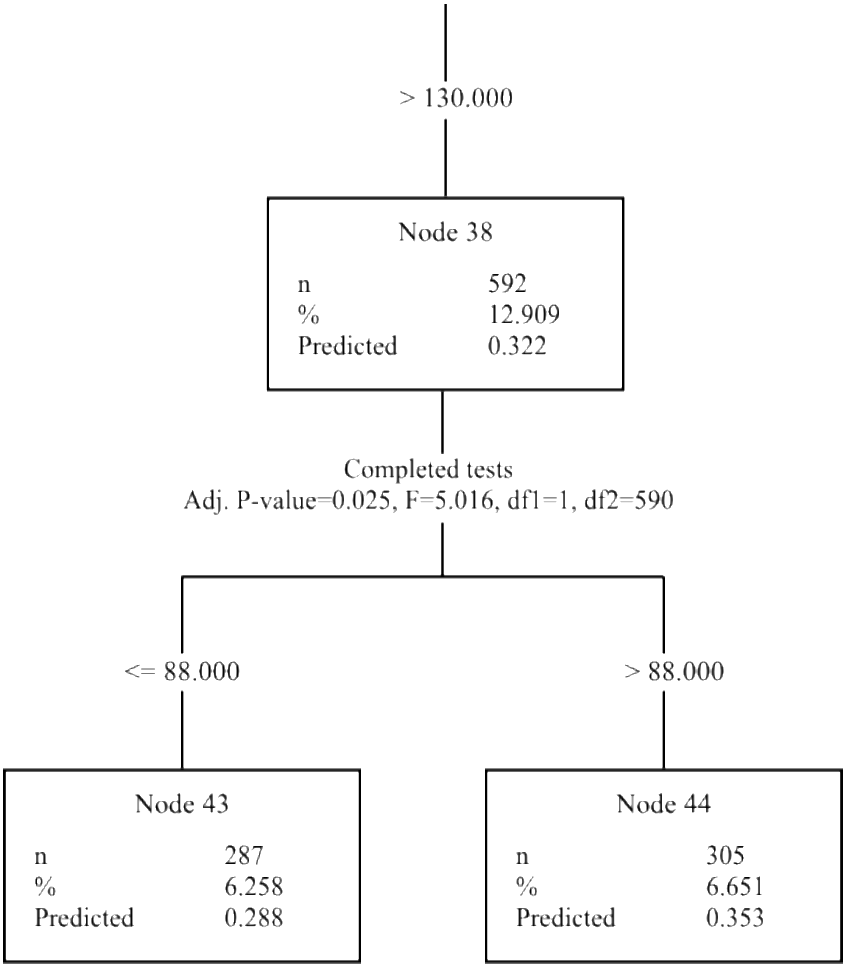


Figure 4.22. Diagram of Node 38

#### 4.1.7.3.1 Node 44

Lastly, Node 44 ($F_{(df1=1, df2=303)}$=10.639, p=0.002) was divided into 2 branches (Node 45 and Node 46) and affecting variable of the division was total answered question number (Answered questions). According to Node 44, students (n=71, Predicted=0.480) who answered less than 1052 questions had better theta results than students (n=234, Predicted=0.315) who answered more than 1051 questions.

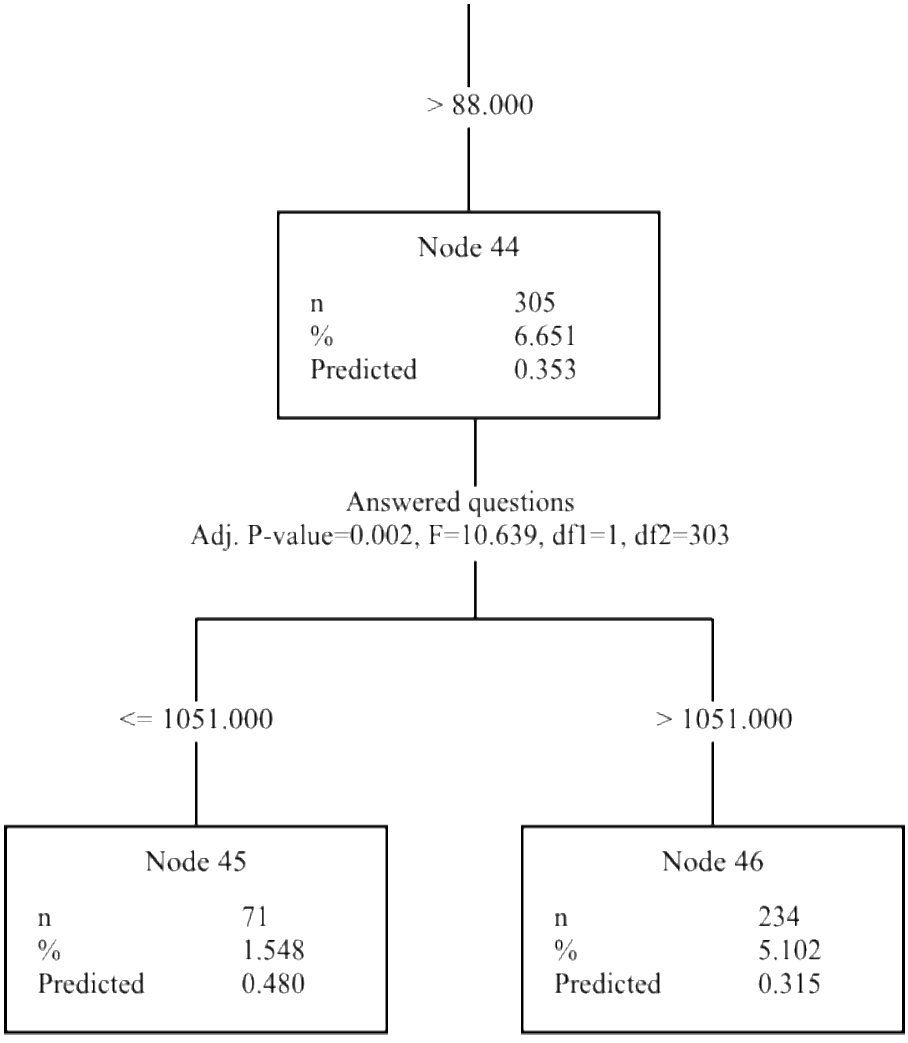


Figure 4.23. Diagram of Node 44

As a result, there is a positive correlation between theta results and the number of logging into the system of the students distributed in Node 7. Among the students who logged in to the system more, theta results of the students who logged in to the system with mobile devices are affected more positively. However, when the mentioned relationship is evaluated from the number of questions answered, the increase in the number of questions causes a negative effect; when evaluated in terms of the number of tests, the increase in the number of tests provides a positive effect in terms of theta results.

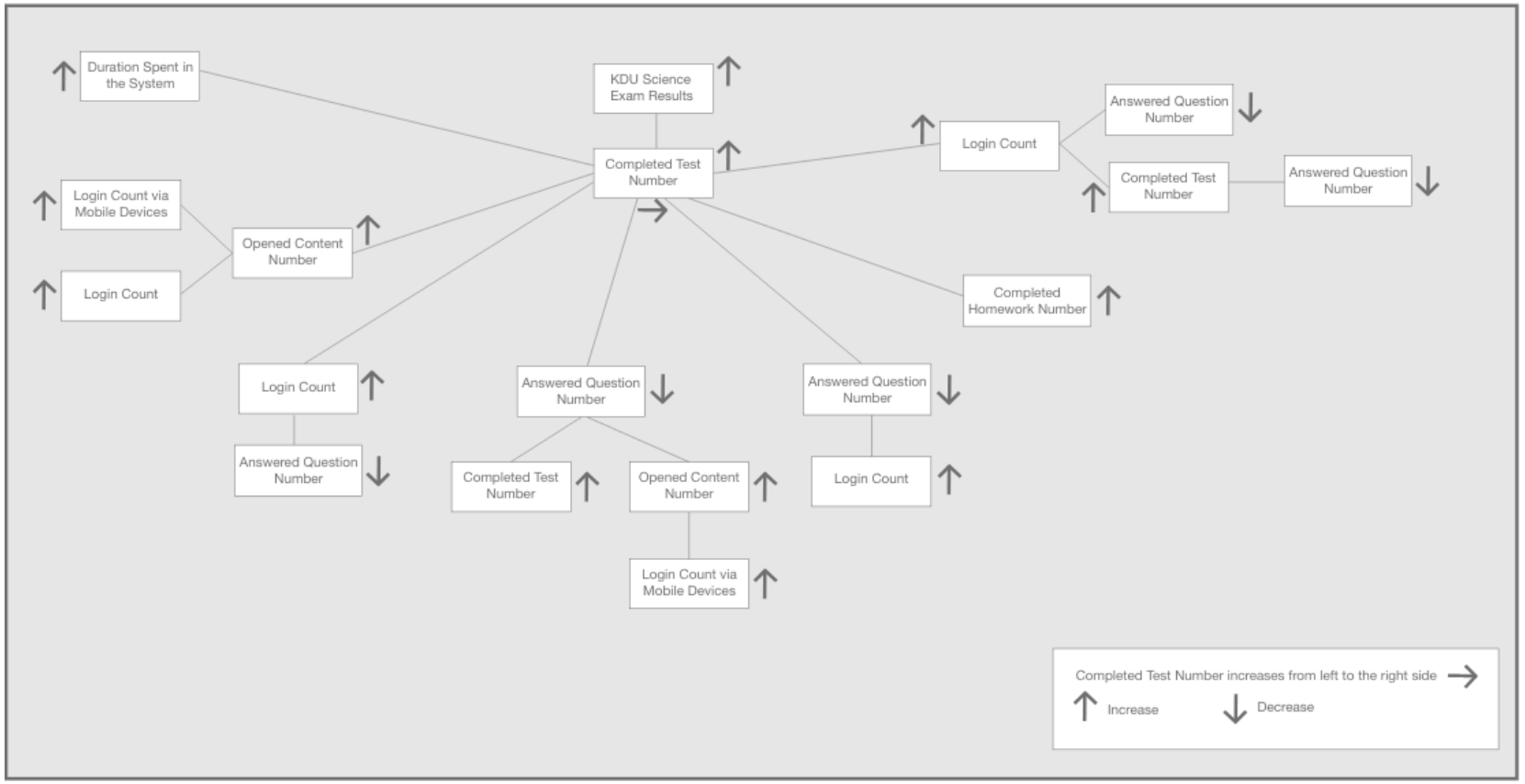


Figure 4.24. Sample Diagram of Overall Results

# CHAPTER 5

## DISCUSSION AND CONCLUSION

In this study, by using the data obtained from the learning activities on a LMS used by middle school students, it was investigated to predict the results of an exam they took at the end of each semester and which of the variables used during this study were effective on these results. Although there are many studies in the literature to predict academic achievement, most of them could not identify any definite variable as the result of success and learning. However, the results of the studies can give clues about the metrics which should be paid attention and studied more. The variables observed on the LMS system are the number of tests they have completed, the number of questions, the number of logins to the system, the time spent in the system and the number of contents they completed.

As a result of the study, as the number of all variables except one variable increased, a positive relationship was observed in the exam results of the students. In the Predictor Importance table, it was determined that the most linked relationship among KDU results and LMS data usage in the study was the total number of tests completed and secondly answered question numbers. There was a high correlation between the number of tests students solved individually on the LMS and the result they got from the exam, and the variable that most positively affected the success of the exam was determined as Completed tests. In a study conducted by Minaei-Bidgoli et al. (2003) also, it was observed that the "Completed tests" variable used by students regarding their performance in tests showed a supportive result for this outcome. In addition, the results showed that the opposite of the number of tests, the increase in the number of questions has a negative impact on the exam results. As the total number of answered questions (Answered questions) increased, a decrease was observed in the exam results of the students. Although the students solve many

tests, the more the total number of questions in the solved tests, the more it affects this relationship negatively. In a study on cigar wrapping, it has been observed that there has always been an increase when the performance of employees over cigar wrapping for 4 years is examined (Crossman, 1959). This example can only change the speed and practicality of doing a repetitive job for employees. However, learning does not take place in this way for students. Working frequency of each lesson and variety of practices provided affect learning.

Rosenbaum et al., (2001) emphasize in their book that the amount of practice given to students should be adjusted according to the difficulty of the tasks and the performance of the students. In other words, the fact that if students are exposed to too many same types of questions in the tests without any measure of students' performance and given tasks negatively affects the learning speed. Diversifying the number of tests and keeping the questions in it low can provide a better result.

All other independent variables used during the analysis established a significant relationship with the dependent variable. It was observed that the positive progression of other variables except the number of answered questions showed a positive correlation between the exam results. When which variables have a positive impact on the exam result is examined, it is seen that they are the duration spent in the system, the number of opened contents, the number of completed homework, the duration spent with mobile devices, number of logging in with mobile devices, and finally the general number of logins to the system which are the subgroups of the number of completed test variable.

As a result of the study conducted by Güldal and Çakıcı (2017) in estimating the academic success of the students, according to the findings; in the data set "Completed homework" feature in the first place has been seen. One of the outputs of the study results was that as the completed number of homework increases, their exam performance increases. That is, providing students with homework related to the subject matter makes an additional positive contribution to their exam performance.

One of the other variables that affects the exam performance of students is the number of contents opened and it has been observed that it has a positive effect on the exam performance as it increases. In addition to the test completion numbers of the students, their studies on the relevant content may support their knowledge of the subject and contribute positively to the exam performance.

Another independent variable that has an effect on most of the subbranches in the results is the frequency of logging into the system. The more students visit the system, the more this affects their exam performance positively. In a study conducted by Akçapınar (2016), it was determined that the number of students entering the system was an effective variable to determine different students' profiles in the online learning environment by clustering method. In further studies also, to classify students' approaches to learning based, students who entered the system more can be analyzed specifically.

Furthermore, the results showed that, in addition to the frequency of logging into the system, students' logging into the system with mobile devices and their stay in the system contributed to the performance of the exam. In other words, it can be said that mobile devices support the frequency of logging into the system in terms of providing an easy access.

On the other hand, the total sample was 6491 students, and 11.25% of these students achieved the highest test score even though they never took the test. These students are exclusively students who can study with their own motivation. In fact, the results showed that if these students spend additional time in the system, then their exam performance would increase even more. That is, if they are also supported with the external guide, their performance can increase more.

Generally, the practice types students are subjected to in such LMS environments are multiple choice, determining whether the definition is true or false or answering with a short text. The impact of such practices on students' end-of-year achievements is related to how such environments design practices. It has been observed that more

positive results are obtained in terms of success when students are provided with more powerfully designed features in such environments (Bernard et al., 2009).

The practices of the LMS system, which was also examined in this study, were in the form of multiple choice, giving short answers or determining whether the definition was correct or incorrect. At the point where the test completion variable is determined as the most positively related variable with the test performance of the results, the LMS environment in which the practices are provided can be increased with practices where there is more interaction. Thus, students' learning and comprehension skills can be supported more with practice interactions (Moreno & Mayer, 1999).

The CHAID decision tree was used in the analysis of the study and it was determined as the most appropriate decision tree structure for interpreting the data. In regression analysis, the data set is expected as just one sample group and extreme values cannot be taken into account. This leads to the conclusion that some estimates may actually be biased. Classification methods have recently been used frequently in many fields due to the opportunity to interpret data more easily (Linoff & Berry, 2013). There are many classification techniques to be used. However, some tree structures (C&R) may not provide a completely homogeneous branching due to their binary structure. In this sense, CHAID analysis has the advantage of its algorithm and analysis findings. It is preferred because of its features such as presenting in the form of a diagram (tree) that everyone can easily interpret and dividing down values to the lowest parts in a homogeneous distribution. In the analysis, while the variables related to the dependent variable in the CHAID analysis were ranked in a hierarchical order, each independent variable that could be a sample at each stage was also reduced to homogeneous subgroups. Considering the advantages of CHAID analysis, it can be thought that its use in research in educational fields can provide effective results and can produce more meaningful and relevant results in interpretation of data.

The results show that two important issues need to be emphasized. First, the learning is a spaced repetition. If the content is repeated on a certain type of frame such as frequency of completing tests, homework or contents about the topic, students are going to learn better. Secondly, commitment makes learning easier for the individuals and login count is an indication of the commitment.

However, it is seen that keeping conducting learning analytic studies from the LMS parameters will not give any conclusive result about how students learn better in online environments. The study of bibliometric analysis of current literature in learning analysis conducted by Phillips and Ozogul (2020) informed that most of the studies on learning analytics have been done on predicting students' success and failure. However, there was no significant link between learning design and learning analytics in those studies. According to the systematic literature review of empirical evidence on learning analytics for learning design conducted by Mangaroska and Giannakos (2019), it is very important to base learning analytics theoretically. In other words, the selection of the methods and analysis used in the studies should be driven by an educational theory and findings from studies should be used to enlighten educational theory and learning design. That is why, first, researchers need to develop some metrics for the learning management systems. Those metrics need to be set up with some sort of parameters before the data is collected about learning. After setting the parameters, learners need to be allowed to experience their online environments. Based on these data sets, researchers should come up with conclusions about how people are learning online. The results finally should be linked to the field of learning. Only in this way educational theories can be associated with the use of online environments in terms of learning.

## REFERENCES


Agudo-Peregrina, A. F., Hernandez-Garcia, A., & Iglesias-Pradas, S. (2012). Predicting academic performance with learning analytics in virtual learning environments: A comparative study of three interaction classifications. 2012 International Symposium on Computers in Education, SIIE 2012.

Aguiar, E., Chawla, N. V., Brockman, J., Ambrose, G. A., & Goodrich, V. (2014). Engagement vs performance: Using electronic portfolios to predict first semester engineering student retention. ACM International Conference Proceeding Series, 103–112. https://doi.org/10.1145/2567574.2567583

Akçapınar, G. (2016). Predicting Students' Approaches To Learning Based on Moodle Logs. EDULEARN16 Proceedings, 1(July), 2347–2352. https://doi.org/10.21125/edulearn.2016.1473

Almosallam, E. A., & Ouertani, H. C. (2014). Learning analytics: Definitions, applications and related fields. Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), 721–730.

Analytics, L. (2018). Special Issue on Educational Big Data and Learning Analytics. Interactive Learning Environments, 26(5), 712–715. https://doi.org/10.1080/10494820.2018.1472430

Anozie, N. O., & Junker, B. W. (2006). Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. AAAI Workshop - Technical Report, WS-06-05, 1–6.

Arnold, K. E., Hall, Y., Street, S. G., Lafayette, W., Pistilli, M. D., Hall, Y., Street, S. G., & Lafayette, W. (2012). course signals in Pudu_using learning analytics to enhance leanring .pdf. May, 267–270.

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. ACM International Conference Proceeding Series, April 2012, 267–270. https://doi.org/10.1145/2330601.2330666

Bader-Natal, A., & Lotze, T. (2011). Evolving a learning analytics platform. ACM International Conference Proceeding Series, 180–185. https://doi.org/10.1145/2090116.2090146

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. Learning Analytics: From Research to Practice, 61–75. https://doi.org/10.1007/978-1-4614-3305-7_4

Baker, R. S. J. D., & Yacef, K. (2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. Journal of Educational Data Mining, 1(1), 3–16.

Bernard, R. M., Abrami, P. C., Borokhovski, E., Wade, C. A., Tamim, R. M., Surkes, M. A., & Bethel, E. C. (2009). A meta-analysis of three types of interaction treatments in distance education. Review of Educational Research, 79(3), 1243–1289. https://doi.org/10.3102/0034654309333844

Brezany, P., Janciak, I., Woehrer, A., & Tjoa, A. M. (2004). GridMiner : A Framework for Knowledge Discovery on the Grid - from a Vision to Design and Implementation. Cracow Grid Workshop, 12–15.

Brown, M. (2011). The Coming Third Wave. EDUCAUSE Learning Initiative, April, 1–4.

Bruckman, A. (2007). Analysis of Log File Data to Understand Behavior and Learning in an Online Community. The International Handbook of Virtual Learning Environments, 1449–1465. https://doi.org/10.1007/978-1-4020-3803-7_58

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic Analytics: A New Tool for a New Era. Educause Review, 42(August 2007), 40–57.

http://net.educause.edu/ir/library/pdf/ERM0742.pdf%5Cnhttp://net.educause.edu/ir/library/pdf/erm0742.pdf%5Cnhttp://eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ769402%5Cnhttp://net.educause.edu/ir/library/pdf/ERM0742.pdf%5Cnhttp://net.educause.edu/ir/li

Chen, G. D., Liu, C. C., Ou, K. L., Liu, B. J., Hamsa, H., Indiradevi, S., Kizhakkethottam, J. J., Schalk, P. D., Wick, D. P., Turner, P. R., Ramsdell, M. W., Ley, T., Kump, B., Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., Punch, W. F., Wolff, A., Zdrahal, Z., … Tedre, M. (2014). Analysis of Log File Data to Understand Behavior and Learning in an Online Community. Interactive Learning Environments, 1(1), 1–6. https://doi.org/10.1109/FIE.2011.6143086

Choi, S. P. M., Lam, S. S., Li, K. C., & Wong, B. T. M. (2018). Learning analytics at low cost: At-risk student prediction with clicker data and systematic proactive interventions. Educational Technology and Society, 21(2), 273–290.

Clark, R. C., & Mayer, R. E. (2016). E-learning and the Science of Instruction important: Fourth Edition. Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Clark, R. C., Mayer, R. E., Bernard, R. M., Abrami, P. C., Borokhovski, E., Wade, C. A., Tamim, R. M., Surkes, M. A., Bethel, E. C., Van der Kleij, F. M., Feskens, R. C. W., Eggen, T. J. H. M., Moreno, R., Mayer, R. E., Alexander, P. A., Duschl, R., & Hamilton, R. (2018). Handbook of Research on Learning and Instruction Learning Science Publication details. Review of Educational Research, 19(3), 475–511. https://doi.org/10.1007/s10648-007-9047-2

Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. ACM International Conference Proceeding Series, 134–138. https://doi.org/10.1145/2330601.2330636

Clow, D. (2013). An overview of learning analytics. Teaching in Higher Education, 18(6), 683–695. https://doi.org/10.1080/13562517.2013.827653

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. British Journal of Educational Technology, 46(5), 904–920. https://doi.org/10.1111/bjet.12230

Dawson, S., Mirriahi, N., & Gasevic, D. (2015). Importance of Theory in Learning Analytics in Formal and Workplace Settings. Journal of Learning Analytics, 2(2), 1–4. https://doi.org/10.18608/jla.2015.22.1

Dietz-Uhler, B., & Hurn, J. E. (2013). Using learning analytics to predict (and improve) student success: A faculty perspective. Journal of Interactive Online Learning, 12(1), 17–26.

Drachsler, H., & Greller, W. (2016). Privacy and analytics - it's a DELICATE issue a checklist for trusted learning analytics. ACM International Conference Proceeding Series, 25-29-Apri, 89–98. https://doi.org/10.1145/2883851.2883893

Elias, T. (2011). Learning analytics: {Definitions}, processes and potential. {Retrieved} {February} 10, 2012. 1–22.

Er, E., Gómez-Sánchez, E., Dimitriadis, Y., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., & Álvarez-Álvarez, S. (2019). Aligning learning design and learning analytics through instructor involvement: a MOOC case study. Interactive Learning Environments, 27(5–6), 685–698. https://doi.org/10.1080/10494820.2019.1610455

Er, E., Gómez-Sánchez, E., Dimitriadis, Y., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., Álvarez-Álvarez, S., Lu, O. H. T. H. T., Huang, J. C. H. C. H., Huang, A. Y. Q. Y. Q., Yang, S. J. H. J. H., Huang, J. C. H. C. H., Lin, A. J. Q., Ogata, H., Yang, S. J. H. J. H., Yamada, M., Oi, M., Konomi, S., Yoo, J., Kim, J., … Zaïane, O. R. (2014). Analysis of Log File Data to Understand Behavior and Learning in an Online Community. Interactive Learning Environments, 1(1), 1–6. https://doi.org/10.1109/FIE.2011.6143086

Ericsson, K. A. (2012). The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance. The Cambridge Handbook of

Expertise and Expert Performance, 683–704. https://doi.org/10.1017/cbo9780511816796.038

Eynon, R. (2013). The rise of Big Data: What does it mean for education, technology, and media research? Learning, Media and Technology, 38(3), 237–240. https://doi.org/10.1080/17439884.2013.771783

Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). Predicting state test scores better with intelligent tutoring systems: Developing metrics to measure assistance required. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4053 LNCS, 31–40. https://doi.org/10.1007/11774303_4

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5–6), 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Ferguson, R., Hoel, T., Scheffel, M., & Drachsler, H. (2016). Guest Editorial: Ethics and Privacy in Learning Analytics. Journal of Learning Analytics, 3(1), 5–15. https://doi.org/10.18608/jla.2016.31.2

Fidalgo-Blanco, Á., Sein-Echaluce, M. L., García-Peñalvo, F. J., & Conde, M. Á. (2015). Using Learning Analytics to improve teamwork assessment. Computers in Human Behavior, 47, 149–156. https://doi.org/10.1016/j.chb.2014.11.050

Galy, E., Downey, C., & Johnson, J. (2011). The effect of using e-learning tools in online and campus-based classrooms on student performance. Journal of Information Technology Education:Research, 10(1), 209–230. https://doi.org/10.28945/1503

García, O. A., & Secades, V. A. (2013). Big data and learning analytics: A potential way to optimize elearning technological tools. Proceedings of the International Conference E-Learning 2013, 313–317.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. TechTrends, 59(1). https://doi.org/10.1007/s11528-014-0822-x

Gordon S. Linoff, & Michael J. A. Berry. (2013). Data Mining Techniques. Third Edition For Marketing, Sales, and Customer Relationship Management. In Journal of Chemical Information and Modeling (Vol. 53, Issue 9).

Gorunescu, F. (2011). Data mining. Concepts, models and techniques. Extended and updated translation from the 2006 Romanian orginal (Vol. 12). https://doi.org/10.1007/978-3-642-19721-5

Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. Educational Technology and Society, 15(3), 42–57.

Güldal, H., & Çakıcı, Y. (2017). Ders Yönetim Sistemi Yazılımı Kullanıcı Etkileşimlerinin Sınıflandırma Algoritmaları ile Analizi (*) Hakan GÜLDAL (**) Yılmaz ÇAKICI (***) Analysis of Course Management System Software Users' Interactions Using Classification Algorithms. 21(4), 1355–1367.

Hand, D. J. (1998). Data mining: Statistics and more? American Statistician, 52(2), 112–118. https://doi.org/10.1080/00031305.1998.10480549

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. Computers in Human Behavior, 36, 469–478. https://doi.org/10.1016/j.chb.2014.04.002

Huang, A. Y. Q., Lu, O. H. T., Huang, J. C. H., Yin, C. J., & Yang, S. J. H. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. Interactive Learning Environments, 28(2), 206–230. https://doi.org/10.1080/10494820.2019.1636086

Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. R. (2018). Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores. Computational Intelligence and Neuroscience, 2018. https://doi.org/10.1155/2018/6347186

Inoue, Y. (2009). Cases on Online and Blended Learning Technologies in Higher Education: Concepts and Practices: Concepts and Practices. IGI Global.

Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An application of a three-stage XGboost-based model to sales forecasting of a cross-border e-commerce enterprise. Mathematical Problems in Engineering, 2019. https://doi.org/10.1155/2019/8503252

Kaya Keleş, M. (2017). An overview: the impact of data mining applications on various sectors. Tehnički Glasnik, 11(3), 128–132.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data and Society, 1(1), 1–12. https://doi.org/10.1177/2053951714528481

Koedinger, K. R., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An open repository and analysis tools for fine-grained, longitudinal learner data. Educational Data Mining 2008 - 1st International Conference on Educational Data Mining, Proceedings, May 2014, 157–166.

Laney, D. (2001). {3D} Data Management: Controlling Data Volume, Velocity, and Variety. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Liñán, L. C., & Pérez, Á. A. J. (2015). Minería de datos educativos y análisis de datos sobre aprendizaje: Diferencias, parecidos y evolución en el tiempo. RUSC Universities and Knowledge Society Journal, 12(3), 98–112. https://doi.org/10.7238/rusc.v12i3.2515

Long, P. D., & Siemens, G. (2014). Penetrare la nebbia: tecniche di analisi per l'apprendimento. Italian Journal of Educational Technology, 22(3), 132–137. https://ijet.itd.cnr.it/article/view/195

Lonn, S., Aguilar, S. J., & Teasley, S. D. (2015). Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. Computers in Human Behavior, 47, 90–97. https://doi.org/10.1016/j.chb.2014.07.013

López, M. I., Luna, J. M., Romero, C., & Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. Proceedings of the 5th International Conference on Educational Data Mining, EDM 2012, January.

Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Lin, A. J. Q., Ogata, H., & Yang, S. J. H. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. Educational Technology and Society, 21(2), 220–232.

Lu, O. H. T., Huang, J. C. H., Huang, A. Y. Q., & Yang, S. J. H. (2017). Applying learning analytics for improving students engagement and learning outcomes in an MOOCs enabled collaborative programming course. Interactive Learning Environments, 25(2), 220–234. https://doi.org/10.1080/10494820.2016.1278391

Mangaroska, K., & Giannakos, M. (2019). Learning Analytics for Learning Design: A Systematic Literature Review of Analytics-Driven Design to Enhance Learning. IEEE Transactions on Learning Technologies, 12(4), 516–534. https://doi.org/10.1109/TLT.2018.2868673

Manyika, J., Chui Brown, M., B. J., B., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition and productivity. McKinsey Global Institute, June, 156. https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf

McCormick, K., & Salcedo, J. (2017). SPSS statistics for data analysis and visualization. John Wiley & Sons.

Milanović, M., & Stamenković, M. (2017). CHAID Decision Tree: Methodological Frame and Application. Economic Themes, 54(4), 563–586. https://doi.org/10.1515/ethemes-2016-0029

Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: An application of data mining methods with an educational web-based system. Proceedings - Frontiers in Education Conference, 1, 1–6.

Moreno, R., & Mayer, R. E. (1999). Multimedia-Supported Metaphors for Meaning Making in Mathematics. Cognition and Instruction, 17(3), 215–248. https://doi.org/10.1207/S1532690XCI1703_1

Oladokun, V., Adebanjo, A., & Charles-Owaba, O. (2008). Predicting students academic performance using artificial neural network: a case study of an engineering course. 1, 68–72.

Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217–222. https://doi.org/10.1080/01431160412331269698

Pasteur, L., & Koch, R. (1941). 1. Introduction 1. Introduction. 74(1934), 535–546.

Phillips, T., & Ozogul, G. (2020). Learning Analytics Research in Relation to Educational Technology: Capturing Learning Analytics Contributions with Bibliometric Analysis. TechTrends, 64(6), 878–886. https://doi.org/10.1007/s11528-020-00519-y

Picciano, A. G. (2012). The evolution of big data and learning analytics in american higher education. Journal of Asynchronous Learning Network, 16(3), 9–20. https://doi.org/10.24059/olj.v16i3.267

Prinsloo, P., & Slade, S. (2016). Student Vulnerability, Agency and Learning Analytics: An Exploration. Journal of Learning Analytics, 3(1), 159–182. https://doi.org/10.18608/jla.2016.31.10

Rokach, L., & Maimon, O. Z. (2008). Data mining with decision trees: theory and applications (Vol. 69). World scientific.

Romero, Cristbal, & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 40(6), 601–618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, Cristóbal, López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. Computers and Education, 68(October), 458–472. https://doi.org/10.1016/j.compedu.2013.06.009

Romero, Cristobal, & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12–27. https://doi.org/10.1002/widm.1075

Rosenbaum, D. A., Carlson, R. A., & Gilmore, R. O. (2001). Acquisition of intellectual and perceptual-motor skills. Annual Review of Psychology, 52(1), 453–470.

Rubinstein, I. S. (2013). Big data: The end of privacy or a new beginning? International Data Privacy Law, 3(2), 74–87. https://doi.org/10.1093/idpl/ips036

Saqr, M., Fors, U., & Tedre, M. (2017). How learning analytics can early predict under-achieving students in a blended medical education course. Medical Teacher, 39(7), 757–767. https://doi.org/10.1080/0142159X.2017.1309376

Schalk, P. D., Wick, D. P., Turner, P. R., & Ramsdell, M. W. (2011). Predictive assessment of student performance for early strategic guidance. Proceedings - Frontiers in Education Conference, FIE, October. https://doi.org/10.1109/FIE.2011.6143086

Shafique, M. A., & Hato, E. (2015). Formation of Training and Testing Datasets, for Transportation Mode Identification. Journal of Traffic and Logistics Engineering, 3(1). https://doi.org/10.12720/jtle.3.1.77-80

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science, 72, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Siemens, G., & Baker, R. S. J. D. (2012). Learning analytics and educational data mining: Towards communication and collaboration. ACM International Conference Proceeding Series, 252–254. https://doi.org/10.1145/2330601.2330661

Siemens, G., & Gasevic, D. (2012). Learning analytics special issue. Journal of Educational Technology & Society, 15(3), 1–2. http://www.ifets.info/journals/15_3/ets_15_3.pdf

Sun, F. R., Hu, H. Z., Wan, R. G., Fu, X., & Wu, S. J. (2019). A learning analytics approach to investigating pre-service teachers' change of concept of engagement in the flipped classroom. Interactive Learning Environments, 0(0), 1–17. https://doi.org/10.1080/10494820.2019.1660996

Tempelaar, D. (2020). Supporting the less-adaptive student: the role of learning analytics, formative assessment and blended learning. Assessment and Evaluation in Higher Education, 45(4), 579–593. https://doi.org/10.1080/02602938.2019.1677855

Tormey, R., Hardebolle, C., Pinto, F., & Jermann, P. (2020). Designing for impact: a conceptual framework for learning analytics as self-assessment tools. Assessment and Evaluation in Higher Education, 45(6), 901–911. https://doi.org/10.1080/02602938.2019.1680952

Tsai, Y. S., Perrotta, C., & Gašević, D. (2020). Empowering learners with personalised learning approaches? Agency, equity and transparency in the context of learning analytics. Assessment and Evaluation in Higher Education, 45(4), 554–567. https://doi.org/10.1080/02602938.2019.1676396

Ture, M., Tokatli, F., & Kurt, I. (2009). Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients. Expert

Systems with Applications, 36(2 PART 1), 2017–2026. https://doi.org/10.1016/j.eswa.2007.12.002

Umobong, M. E. (2017). The One-Parameter Logistic Model ( 1PLM ) And Its Application In Test Development. 4(24), 126–137.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. Review of Educational Research, 85(4), 475–511. https://doi.org/10.3102/0034654314564881

Verhelst, N. D., & Glas, C. A. W. (1995). The One Parameter Logistic Model. Rasch Models, 215–237. https://doi.org/10.1007/978-1-4612-4230-7_12

Villagrá-Arnedo, C., Gallego-Duran, F. J., Compan-Rosique, P., Llorens-Largo, F., & Molina-Carmona, R. (2016). Predicting academic performance from Behavioural and learning data. International Journal of Design and Nature and Ecodynamics, 11(3), 239–249. https://doi.org/10.2495/DNE-V11-N3-239-249

Wilkinson, L., Partitioning, K. W. R., & Trees, R. (1992). Tree Structured Data Analysis: {AID}, {CHAID} and {CART}. Paper presented at the 1992. University of Ostrava, January 1992.

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment. ACM International Conference Proceeding Series, April, 145–149. https://doi.org/10.1145/2460296.2460324

Yamada, M., Oi, M., & Konomi, S. (2017). Are learning logs related to procrastination? From the viewpoint of self-regulated learning. 14th International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2017, Celda, 3–10.

Yoo, J., & Kim, J. (2014). Can online discussion participation predict group project performance? Investigating the roles of linguistic features and participation

patterns. International Journal of Artificial Intelligence in Education, 24(1), 8–32. https://doi.org/10.1007/s40593-013-0010-8