IMPLEMENTATION OF DATA MINING ALGORITHMS ON ROCK MECHANICS TEST DATA FOR KNOWLEDGE DISCOVERY

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

BY

CENGIZ KAYDIM

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

AUGUST 2021

Approval of the thesis:

IMPLEMENTATION OF DATA MINING ALGORITHMS ON ROCK MECHANICS TEST DATA FOR KNOWLEDGE DISCOVERY

submitted by **CENGİZ KAYDIM** in partial fulfillment of the requirements for the degree of **Master of Science in Mining Engineering, Middle East Technical University** by,

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

Prof. Dr. Naci Emre Altun
Head of the Department, **Mining Engineering, METU** ________________

Assist. Prof. Dr. Mustafa Erkayaoğlu
Supervisor, **Mining Engineering, METU** ________________

Examining Committee Members:

Assist. Prof. Dr. Muhammet Mustafa Kahraman
Mining Engineering, GÜ ________________

Assist. Prof. Dr. Mustafa Erkayaoğlu
Mining Engineering, METU ________________

Assist. Prof. Dr. Ahmet Güneş Yardımcı
Mining Engineering, METU ________________

Date: 31.08.2021

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 : Cengiz Kaydım

Signature :

ABSTRACT

IMPLEMENTATION OF DATA MINING ALGORITHMS ON ROCK MECHANICS TEST DATA FOR KNOWLEDGE DISCOVERY

Kaydım, Cengiz
Master of Science, Mining Engineering
Supervisor : Asst. Prof. Dr. Mustafa Erkayaoğlu

August 2021, 82 pages

Rock mechanics is a fundamental research field of engineering as the mechanical properties of rocks are crucial in mining and civil engineering applications. These properties control main production processes like excavation, drilling, and blasting in addition to geotechnical studies, such as slope stability for surface mining. Experimental studies performed conforming to suggested methods provide essential results representing the mechanical properties of rock material. Within the scope of this thesis study, a database was created containing a total of 9,967 test results, including 284 different projects carried out in the METU Mining Engineering Rock Mechanics Laboratory since the year 2000. After the raw experiment data was prepared by data cleaning operations, it was transferred to the database. OLAP cubes with multidimensional query features were developed to allow advanced analysis by collecting the data in a data warehouse. It is aimed to investigate the potential knowledge discovery of the rock mechanics-related test data by data mining algorithms with the support of the developed data warehouse. A case study was conducted to demonstrate the potential knowledge discovery capability of the data warehouse. In this study, rock types were classified to back fill the missing rock type information using decision tree and random forest algorithms trained. The validation results revealed that the random forest model performed approximately 43 % better than the decision tree model.

Keywords: Data warehouse, Rock mechanics, Relational database, OLAP, Data analysis

ÖZ

BİLGİ KEŞFİ AMACIYLA VERİ MADENCİLİĞİ ALGORİTMALARININ KAYA MEKANİĞİ DENEY SONUÇLARI VERİ AMBARI ÜZERİNDE UYGULANMASI

Kaydım, Cengiz
Yüksek Lisans, Maden Mühendisliği
Tez Yöneticisi : Dr. Öğr. Üyesi Mustafa Erkayaoğlu

Ağustos 2021, 82 sayfa

Kaya mekaniği, kaya malzemesinin mekanik özelliklerinin madencilik ve inşaat uygulamalarında önemli bir rol oynaması nedeniyle temel bir mühendislik araştırma alanıdır. Bu özellikler kazı, delme, patlatma, açık ocak madenciliği için şev stabilitesi gibi jeoteknik çalışmalardaetkili olmaktadır. Kaya malzemesinin mekanik özellikleri uluslararası önerilen metodlara uygun şekilde yürütülen deney sonuçları ile temsil edilmektedir. Bu çalışma kapsamında 2000 yılından günümüze kadar ODTÜ Maden Mühendisliği Kaya Mekaniği Laboratuvarı'nda gerçekleştirilmiş 284 farklı projeye ait toplam 9.967 adet deney sonucunu içeren bir veri tabanı kurulmuştur. Ham deney verisinin temizlenmesi ile hazırlık aşamasının ardından veri tabanına aktarımı sonrasında veriler bir veri ambarında toplanarak ileri düzey analizlere olanak sağlaması için çok boyutlu sorgu özelliği bulunan OLAP küpleri geliştirilmiştir. OLAP küpleri ve veri ambarı ile veri madenciliği algoritmaları kullanılarak bilgi keşfi amacıyla incelenmesi amaçlanmıştır. Bu çalışma kapsamında veri ambarı altyapısının bilgi keşfi potansiyelinin irdelenmesi amacıyla örnek bir vaka çalışması yürütülmüştür. Bu çalışmada karar ağacı ve rastgele orman algoritmaları aracılığıyla eksik kaya tipi bilgisinin tamamlanması için kaya türleri

sınıflandırılmıştır. Doğrulama sonuçları rastgele orman modelinin karar ağacı modeline kıyasla yaklaşık % 43 daha iyi performans gösterdiğini ortaya koymuştur.

Anahtar Kelimeler: Veri ambarı, Kaya mekaniği, İlişkisel veri tabanı, OLAP, Veri analizi

To my family

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to my advisor Assist. Prof. Dr. Mustafa Erkayaoğlu for his assistance, encouragement, and supervision. I am also grateful to my committee members Assist. Prof. Dr. Ahmet Güneş Yardımcı and Assist. Prof. Dr. Muhammet Mustafa Kahraman.

I want to thank my friends Ahmet Göktekin, Mert Bostankolu, Görkem Bulut, and Serkan Dedeoğlu for their friendship and help during my study. I also want to thank my colleagues Alper Kırmacı, Ceren Karataş Batan, Kutay Karadeniz, Doğukan Güner, and Enver Yılmaz for their assistance and friendship. I would like to extend my appreciation to all of the administrative, academic, and technical staff in the Department of Mining Engineering, METU, for their support of my education.

Finally, I am immensely grateful to my fiancee Hatice Esin Çaydaş, my mother Serpil Kaydım, to my father Turan Kaydım, my brother Erdem Kaydım, my aunt Tiffany Bruckert Kaydım, my niece Ela Grace Kaydım and my nephew Elliot Demir Kaydım and my cousin Samet Kaydım for their love and support. This thesis study would not have been completed without their encouragement and faith. I dedicate this thesis study to them.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Problem Statement

The mechanical properties of rock material are determined by different tests conducted in accordance with suggested methods, such as those of the International Society for Rock Mechanics (ISRM). Some of the most commonly performed tests are static deformability, uniaxial/triaxial compressive strength, direct/indirect tensile strength, shear test, and density and porosity determination. The storage, data management, and analysis of these test results can become a complicated task as the number of experiments increases, with an inconsistent presentation of data that makes it more difficult to perform comparative analysis. Test results should instead be kept in a standard form to better enable advanced analysis and to more effectively and securely share data among researchers and engineers. This ideal data standardization and collective analysis can only be achieved using data management tools, such as a database and data warehouse. In addition, it is observed in the past conducted experiments that some parameters were not recorded. A data warehouse is needed in order to complete missing information.

The experiment results of the Rock Mechanics Laboratory of the METU Department of Mining Engineering have been stored in different formats, such as spreadsheets and hard copy reports. These methods lack security and reliability for data analysis. In addition, the collective analysis of the data is not possible due to integration limitations from the variety of data storage methods. In this thesis study, related studies available in the literature are examined to determine potential solutions to these data-related issues. It is observed that an infrastructure, which is composed of a database and data warehouse, is needed for the storage, collective analysis and backfilling the missing information of the METU Mining Engineering Department's rock mechanics experiment test results.

1.2 Aim and Scope of the Study

Preliminary analysis showed that rock types in the experiment results were the mostly missing data in the experiment results. This thesis study aims to develop a rock mechanics database (RMDB) and rock mechanics data warehouse (RMDW) to store, analyze, and employ decision tree and random forest algorithms to backfill the rock type information. These infrastructures provide access to a fast, reliable, and secure environment to retrieve information. In order to achieve the objective of developing a RMDB and RMDW, experiment results data was collected from static spreadsheets and hard copy reports. The gathered data was then cleaned with data preprocessing techniques before being transferred to tables in SQL Server Management Studio (SSMS) software. The RMDW was developed in MS SQL Server Data Tools (SSDT) software via a connection to the RMDB. Online Analytical Processing (OLAP) cubes were created for multidimensional analysis in the data warehouse. The analytical potential of the developed infrastructure is presented in different applications throughout this thesis. In addition, missing rock type information of filtered historical experiments were back filled with a decision tree and random forest algorithms via Rattle software.

In this study, the primary focus is the experiment results data generated between the years 2000-2021 and was the sole data used for the data gathering and cleaning processes. The expansion of the data back to the 1960s will increase the amount of data and will provide a better representation of the various experiments conducted on different rock samples.

1.3 Thesis Outline

This thesis study includes five chapters and one appendix. Chapter 1, the problem statement, presents the objectives and scope of the study. In Chapter 2, the literature review of the study is provided. Chapter 3 outlines the methodology followed during the database, data warehouse development processes, and utilized data mining algorithms. Chapter 4 gives the results of the conducted collective analysis in RMDW and data mining case study. Finally, the conclusion and recommendations are stated in Chapter 5.

Appendix A provides the created tables in RMDB.

CHAPTER 2

LITERATURE REVIEW

2.1 Rock Material Related Databases

Although there is extensive and comprehensive literature in the field of rock mechanics, the number of research studies where the experimental results are defined as data, integrated with different data types, and used for knowledge discovery is limited. The purpose of developing a database is to ensure that the data is stored in a digital platform safely, which also enables the data to be queried easily when required (Zhu, Li, & Zhuang, 2011). To date, various databases have been used to store and process the information of the mechanical and physical properties of materials. Most of these examples focus on engineering materials, such as ceramics and metals. MatWEB, NIMS Materials Database, and MATBASE are some examples of such available databases. Databases created for the properties of natural materials, such as rock, generally focus on geological and mineralogical information (e.g., petrographic data, mineral composition, microstructure). Mindat is an example of a web-based database that provides general information for minerals (Ralph, J., 1993). The RockPro software, which is developed for engineering applications, is a rock mechanics database tool for storing and reporting recorded data in underground operations (www.esgsolutions.com). This software is used to store data from pillar stability, rock burst, and support designs for specific projects. A more comprehensive rock mechanics database that stores the mechanical parameters of rocks is the RocProp database developed by RocScience (Turichshev, 2002). This database contains more than 700 rock mechanics-related test results. Each record includes information about the rock (rock type, country, location, unit weight), reference, and test results. The test results are divided into four categories according to both the experiment type and the failure parameters used. Uniaxial/triaxial compressive strength tests, direct and indirect tensile strength tests are examples of data that can be obtained from this database. Similarly, data collected by wave speeds representing compressive strength and shear test conditions are also available (Liolios & Exadaktylos, 2011).

There are limited number of studies focusing on the development of the rock mechanics test results database in the literature. Liolios and Exadaktylos (2011) developed a hierarchical database of rock mechanics-related test results. This database aims to store the results of rock mechanics, including rock sampling sites, test procedures, data reduction, and model calibration methods. The database developed by the researchers has been prepared by using Structured Query Language (SQL) and consists of three main sections: rock, experiment, and laboratory. The rock material section, shown in Figure 2.1, consists of tables that include the properties of the tested sample. These tables are sampled rock location, microscope image, mineral content, texture, microstructure, physical properties, and visual photographs. The tables in the experiment section include the dimensions of the sample, the modeling conditions, the measurement techniques, the deformation, and the strength results obtained from the experiment. The test section is divided into five sub-sections containing the test types. These sections are Brazilian Tests (BT), Drilling Tests (DT), Shear Tests, Uniaxial Compression / Triaxial Compression (UCTC), Uniaxial Tension Tests (UT). The rock and laboratory sections also contain multiple tables, such as the experiments section. The parent table includes necessary information, while the other child tables provide additional information. While each record added to the table is kept in a column, the child tables are structured concerning the relationships defined in the parent tables. As a result, the database is designed to prevent the input of data in the sub-table if there is no available connection to the main table.

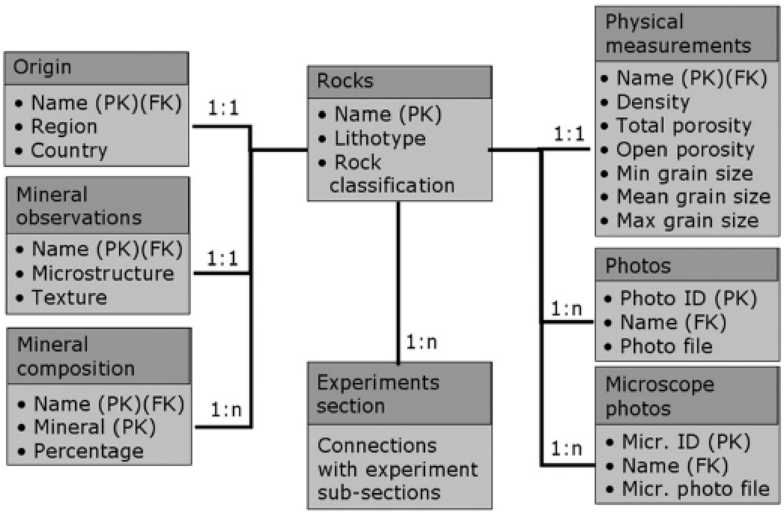


Figure 2.1 Part of the relational diagram of the database showing the rock section (Liolios & Exadaktylos, 2011)

Exadaktylos et al. (2007) used a data reduction method in the database where data is stored in two levels, Level 0 and Level A. In Level-0, the code number of the experiment, the dimensions of the sample, maximum strain, strain stress, failure pressure, time, confining stress, axial, and lateral strain values are stored in the columns. In Level-A, the test curve is divided into loading and unloading - restoring sections, and the elastic modulus value of the rock is calculated. In the next stage, the plastic behavior of the rock can be calculated according to the failure criterion and Level-A data.

The first study considered to be the pioneer of similar studies was the report prepared by Hsiung et al. (1995) for the Nuclear Regulatory Commission Contract NRC-02-93-005. This report deals with the development of a rock mechanics properties database to facilitate the analysis and stability studies of an underground gallery where nuclear waste will be stored. In this database, rock material properties, crack properties, rock thermal properties, hydrogeological properties, Rock Mass Rating (RMR), and Norwegian Geotechnical Institute Tunneling Quality Index (Q) rock mass classification methods were recorded.

Kim & Hunt (2017) developed the Earth Mechanics Institute (EMI) Rock Mechanics Database. The database aims to enable researchers to access any mechanical properties requested by a rock type on a web-based data environment securely and easily. MySQL Server infrastructure was used to create the database. The schematic given in Figure 2.2, designed by the researchers, includes the metadata of the rock, the position data, and tables of eight different test type related data. The tables of this test results are Brazilian tensile strength (BTS), cerchar abrasivity index (CAI), cohesion, density, direct shear, punch penetration, triaxial, and uniaxial compressive strength (UCS). The primary key of each test table is also treated as a foreign key to the EMI data table. Also, the position foreign key in the EMI data table is connected to the ID column in the location table.

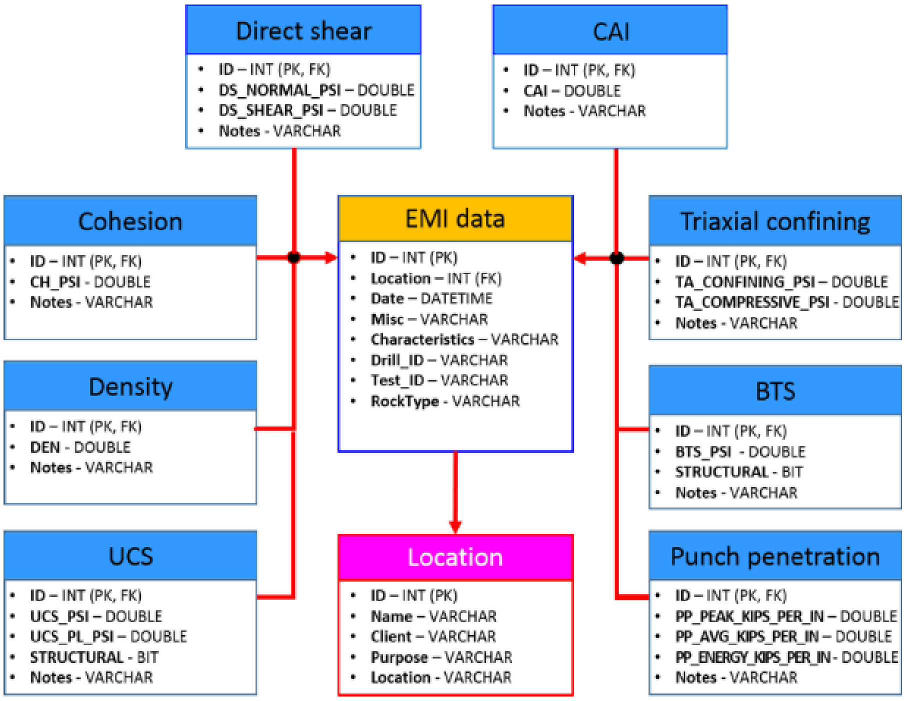


Figure 2.2 Schema design for EMI rock mechanics database (Kim & Hunt, 2017)

Another database developed for rock mechanics test results is completed by Pere et al. (2011). A geotechnical data management system has been developed to manage drilling and mapping data generated by consulting firms. This system is an infrastructure that provides verification, processing, and reporting to the user through a data collection interface. With this system, accessing and querying geotechnical data is provided in a user-friendly environment through a single data store.

Descamp et al. (2013) developed a rock mechanics database to evaluate the abrasion effect of a rock formation on the drilling process. In order to analyze the collected data, sample, petrographic analysis, petrophysical analysis, mechanical analysis, and report tables were created in the database. The scheme used by the researchers for this infrastructure is given in Figure 2.3.

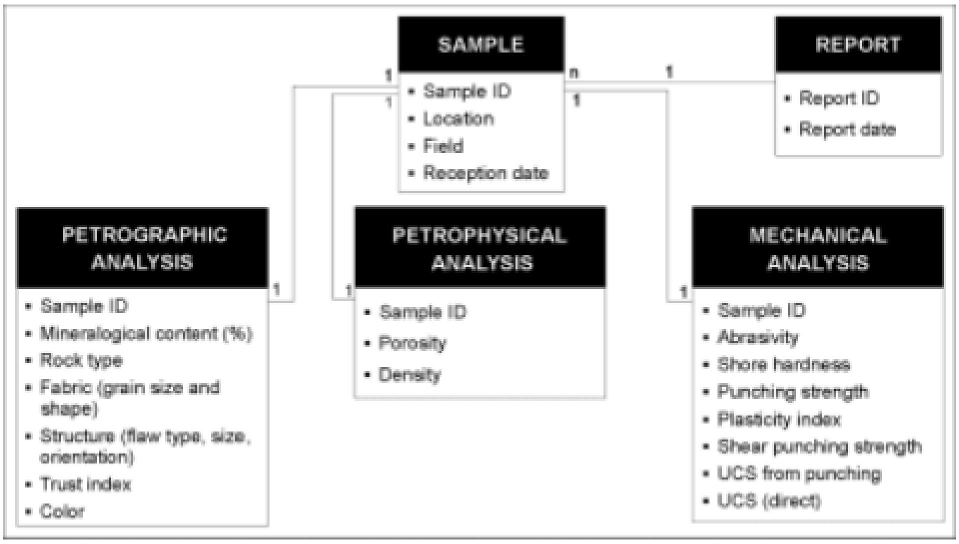


Figure 2.3 Schematic of the database prepared for the assessment of the impact of rock formation abrasion (Descamp et al. 2013)

Li, Wang, and Zhu (2012) investigated the development processes of data storage and sharing infrastructures such as Greenwood, AGS, AGSML, and DIGS to keep the output of rock mechanics test results following specific standards. The experimental methods proposed by ISRM were discussed. It has been observed that the test methods offered by ISRM did not follow the standards established so far or the structure of the databases such as RocLab Rockware serving particular demands. It was also stated that the test implementation instructions for some experiments were not clearly defined within these standards. Fifty-two proposed methods have been developed by ISRM for various experiments and applications to increase the collective utilization of the rock mechanics experiment results. The data schematic of the Uniaxial Compression Test (UCT), which is one of these methods, is given in Figure 2.4.

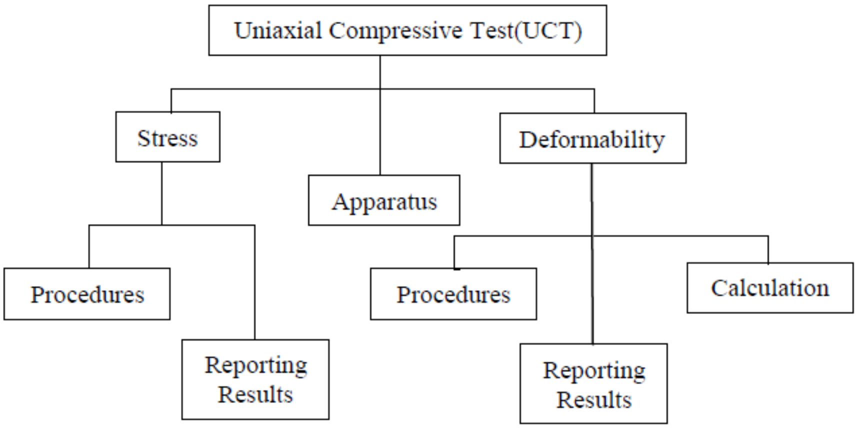


Figure 2.4 The data structure of the uniaxial compression test (Li et al., 2012)

Ng and Lau (2015) developed a database containing data such as uniaxial compression (UCS), point loading index (Is_{50}), Schmidt hammer test (R_L), porosity (η), water saturation (S), specific gravity (Gs) dryness and density (ρ_d) in their study. A total of 151 data sets were obtained from a series of experiments performed and a database was compiled for statistical analysis. The mean, standard deviation, and coefficient of variation values for each of the rock properties in this compiled database are given in Table 2.1.

Table 2.1 Uniaxial compression strength (UCS), point loading index (Is_{50}), Schmidt attractive (R_L), porosity (η), water saturation (S), specific gravity (Gs), and statistical information of dry density (ρ_d) experiments (Ng & Lau, 2015)

Data Set	Property	Min.	Max.	Average	Standard Deviation	Coefficient of Variation
151	UCS	12.00	134.80	62.38	25.27	0.41
	Is_{50}	0.86	12.54	5.99	2.22	0.37
	R_L	21.30	56.00	46.91	6.62	0.14
	η	0.45	6.13	1.22	0.74	0.60
	S	0.17	2.50	0.47	0.30	0.64
	Gs	2.46	2.66	2.61	0.04	0.01
	ρ_d	2.59	2.68	2.64	0.02	0.01

Using the database, the researchers developed a new empirical correlation to estimate the UCS of Macau granite. A comprehensive statistical analysis was performed to correlate UCS with Is50, RL, η, S, Gs, and ρd. In the first stage, the one-parameter relationships between UCS and rock index properties (Is_{50}, R_L, η, S, Gs, and ρ_d) were examined respectively and presented in Figure 2.5. In the second stage, convenient features were selected to generate rock-properties correlations by multiple regression analysis to estimate the UCS using the results from the first stage. A local empirical correlation was proposed for Macau granite with the created database, and the reliability of the correlation was measured with the independently created test database. This study reveals the contribution that rock mechanics test results can offer after a series of stages.

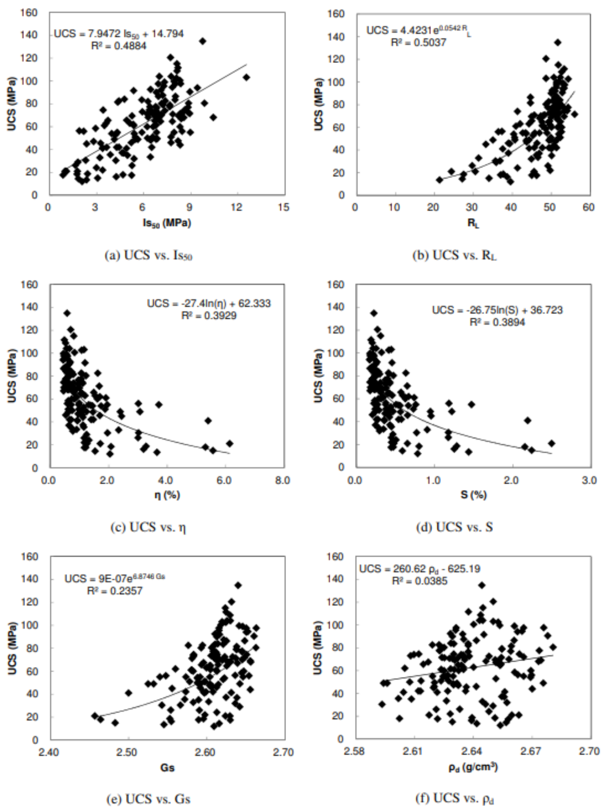


Figure 2.5 Relationships between UCS and rock index properties (Ng & Lau, 2015)

The Geological Engineering Department of Technische Universität München (TUM) has developed a web-based database solution to handle all data related to research projects in a more suitable way for advanced analysis. (Menschik, Thuro & Käsling, 2015). Researchers aimed to optimize data management by finding a quick and easy way to analyze and report collected data. Their secondary purpose is to provide easier access to data for project partners located in different locations. This database system consists of a PHP web interface and MySQL database core. Figure 2.6 shows the entity-relationship (ER) model of the main part of the rock database. This diagram provides a basis for programming the database and shows the connection between all tables, showing which combinations of cross-table queries are possible.

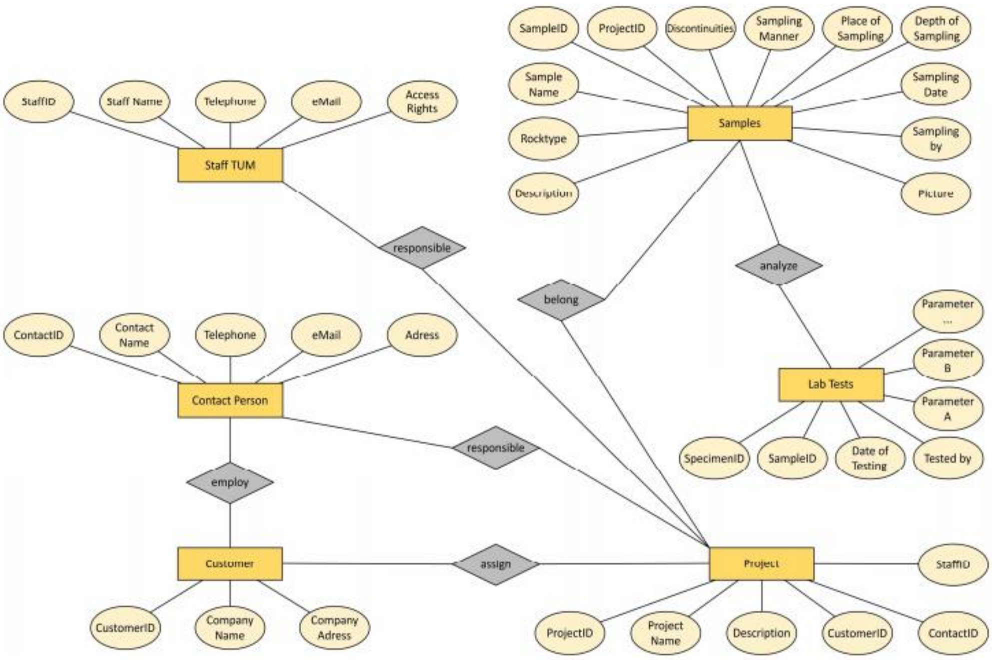


Figure 2.6 Entity-relationship model of rock database created in TUM (Menschik, Thuro & Käsling, 2015)

Although the database system has a modular design, it consists of three modules: general project data, laboratory data, and user management modules (Menschik, Thuro & Käsling, 2015). In Figure 2.7, the general and schematic layout of the database is shared. The user can perform a simple statistical analysis within the database and export the data for further analysis. This study is another example of how a well-structured database can provide significant value to mining projects with different data sources by providing the opportunity to manage and analyze data easily.

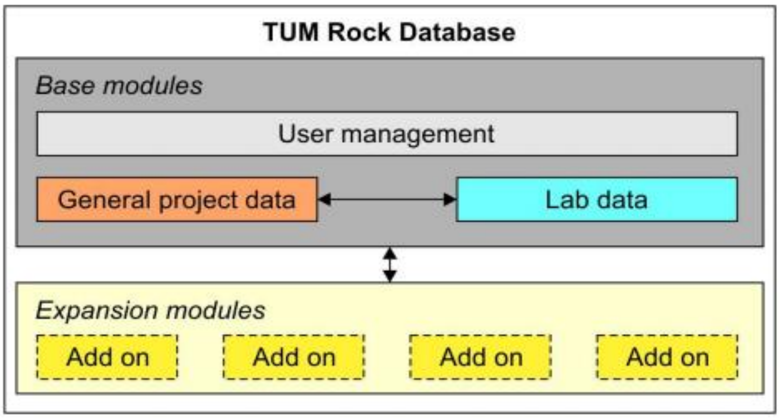


Figure 2.7 General and schematic layout of the TUM rock database (Menschik, Thuro & Käsling, 2015)

Chuanyao et al. (2015) presented another example of database and data warehouse application in mining engineering. A concept called “Genetic Mineral Processing Engineering (GMPE)” has been introduced to take advantage of what is called “genetic characteristics”, which are related to the bedding formation, ore and mineral properties, and beneficiation. The steps of the researchers’ work on this subject are a) testing and filtering “Genetic Traits”; b) Establishment of GMPE’s database and data warehouse; c) Intelligent decision-making system for the ore beneficiation process; d) Validation test run; e) A virtual ore processing facility. The technical route of the study can be seen in Figure 2.8. Genetic properties are collected with testing and stored and analyzed in a database and data warehouse environment. Processing method and technical parameters are recommended with an intelligent decision-making system. Verification and optimization stages are completed with different data analysis and modelling tools with recommended parameters and processing method.

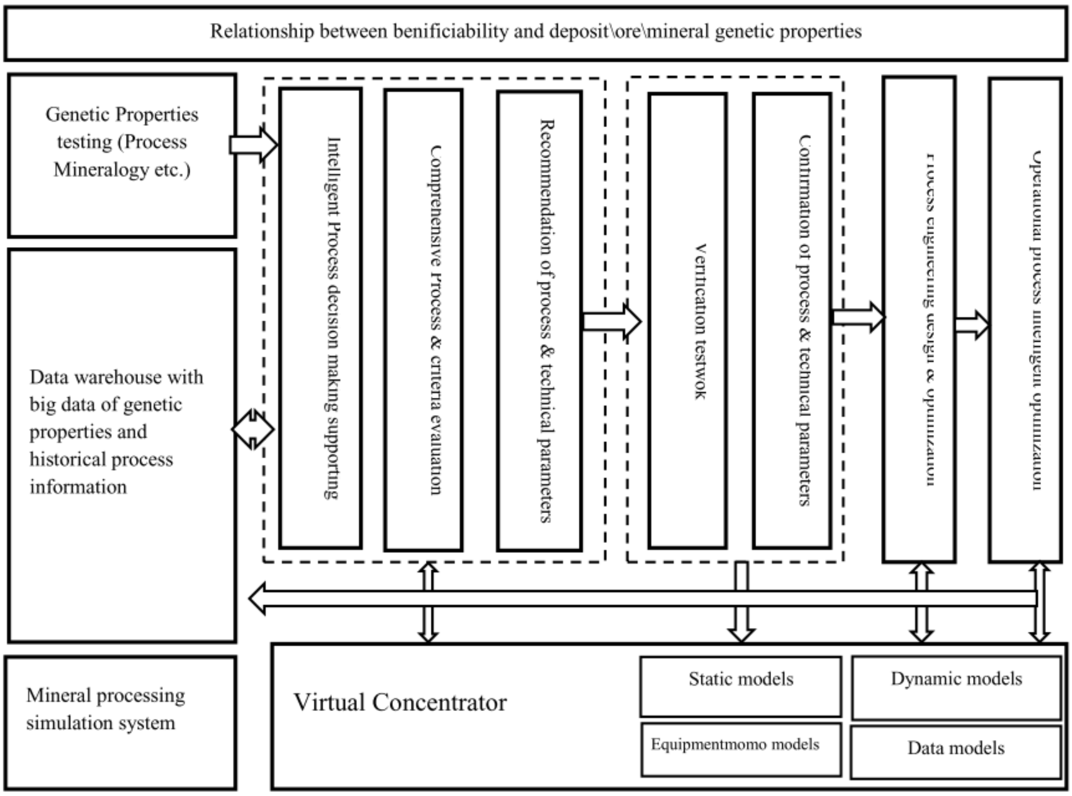


Figure 2.8 The technical route of Genetic Mineral Processing Engineering (Chuanyao et al., 2015)

Jones (2015) presented a compilation of the physical properties of the rocks in the Bushveld Complex in his study. The database consists of more than 900 measurements and 190 heat capacity values, each for thermal conductivity and density. The database consists of three main tables. Table I contains specific rock types. Table II contains the average thermal conductivities and densities of different rock types in different stratigraphic units of the Bushveld Complex. Table III includes heat capacity measurement data.

Gering et al. (2017) conducted a case study to demonstrate the contribution of an Enterprise Data Warehouse (EDW) used to manage drilling and blasting information on decision-making and process improvement and discuss how technical difficulties encountered during implementation were resolved. This case study examines the process of implementing a drill-blast workflow at multiple sites within Freeport McMoRan Inc (FMI) using both field data and information obtained using EDW. Particle analysis from drilling-blasting, explosive data used, and data generated by drilling systems can be visualized and reported using MineSight software (Figure 2.9). In this way, instantaneous feedback is provided on how far the explosion achieved its fragmentation targets. In addition to visual inspection of blasting success by data warehouse users, it is possible to benefit from further analysis by looking at the relationship between fragmentation, rock type, and other parameters.

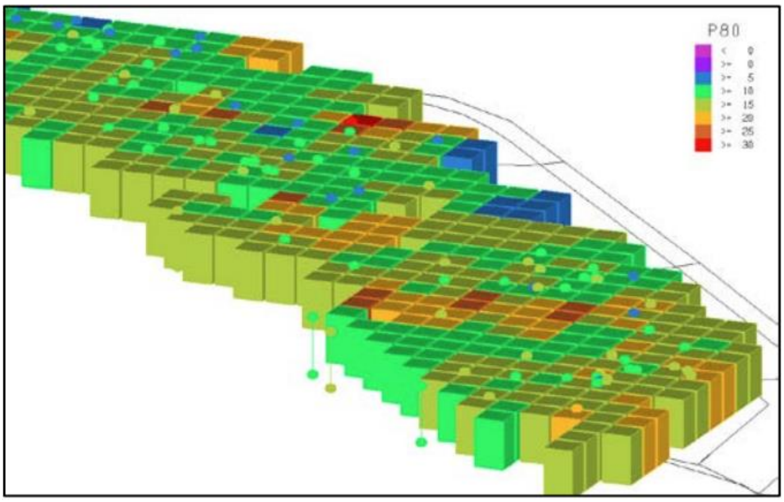


Figure 2.9 Fragmentation analysis image created in Minesight software (Gering et al., 2017)

The literature review summarized in this chapter show that storing rock properties-related data in a database makes data management relatively easy and efficient. It is possible to analyze data obtained from different databases or different tables in a database through a data warehouse. Various researchers used these database and data warehouse infrastructures to convert raw data to decision-making mechanisms through different intelligent systems, graphs, and tables. Such tools can add value by producing knowledge from raw data and provide the required infrastructure for data analysis and data mining.

2.2 Data Analysis and Data Mining of Rock Material Related Data

Data analysis is the process of cleaning, transforming, and modeling data to find useful information, reach conclusions, and support decision-making (Brandt & Brandt, 1998). The data analysis process includes several distinguishable stages, many of which are iterative and may require additional work in earlier stages with the results obtained afterward.

A good data analysis methodology is essential to predict the project's progress and accurately calculate the impact of the outputs to be obtained at the end of the project. Various methodologies have been created for data analysis to date. OSEMN (obtain, scrub, explore, model, interpret), SEMMA (sample, explore, modify, model, asses), KDD (knowledge discovery in databases), CIRSP-DM (cross-industry standard process for data mining) and TDSP (team data science process) systems can be given as examples of these methodologies (Azevedo & Santos, 2008; Saltz et al., 2018; Shafique & Qaiser, 2014). The knowledge discovery (KDD) scheme in a database, which is one of these systems, is given in Figure 2.10 as an example. The diagram shows the necessary stages for a data science project. At these stages, the process of collecting the data in the database, transferring it to the data warehouse after certain stages, and then transforming it into information with various data analysis methods is summarized.

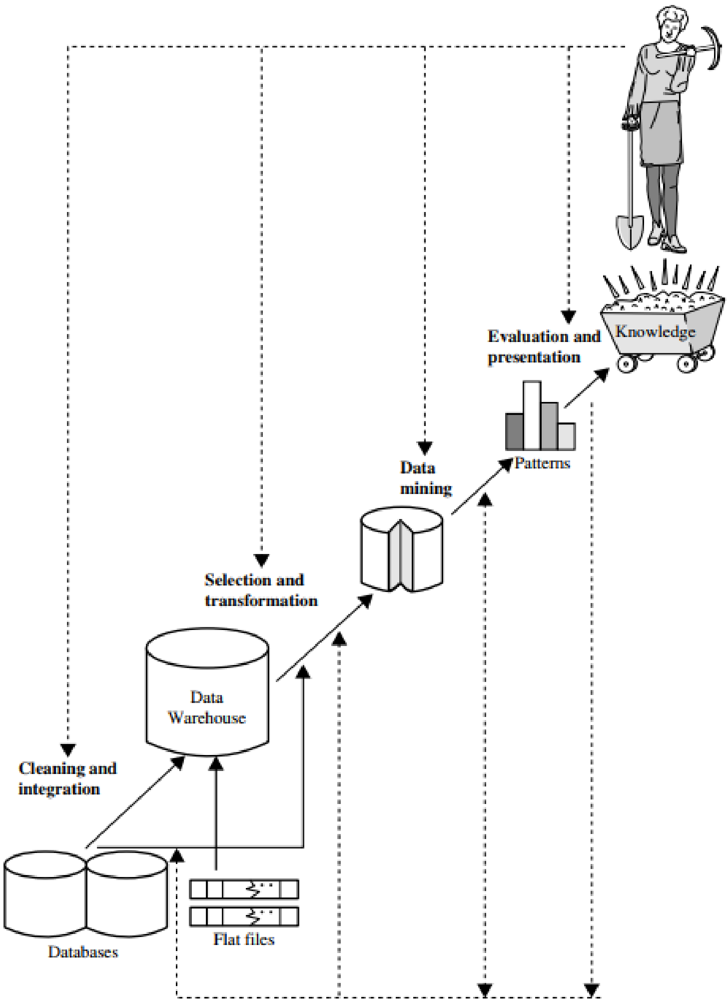


Figure 2.10 Data mining as a step in the process of knowledge discovery (Han, 2006)

There are many different approaches to data analysis. Data analysis may include different techniques in different fields and allows data collection from a wide variety of sources obtained through different sources such as traffic cameras, satellites, recorders (Zohuri & Moghaddam, 2017). Data collected for data analysis in the mining industry is commonly obtained through experimental research and equipment-based technologies. The collected data must be organized and preprocessed before being used for data analysis. A data warehouse is one of the systematical implementations where these processes are carried out, and data is sourced from different systems and integrated for business intelligence purposes.

Technologies used in mining engineering are continuously improving, as seen in every industrial production activity. Digitization for the mining industry includes the use of computer-assisted or digital devices, methods, systems, and digitized data to increase work productivity and efficiency while reducing costs (Barnewold & Lottermoser, 2020). One of the most important and widely applicable current digital developments in the mining industry, which benefits from versatile processes and high-capacity equipment, is the analysis of the collected data and its use in decision-making processes.

Another tool used in decision making process is OLAP, a decision support technology, which is among the data visualization studies (Codd et al., 1993; Maniatis et al., 2005). OLAP research is closely related to data warehouses, which are considered the information sources where OLAP is performed. The general data flow path involves collecting data from various sources to data warehouse systems and then using this data in the multidimensional analysis process using OLAP applications (Tsois et al., 2001). The multidimensional analysis primarily involves calculating aggregated information using large volumes of detailed data. Data is analyzed based on detailed or derived characteristics (dimensions) using an almost static business model (hierarchies). The application of the OLAP cubes in the mining industry is still extremely limited.

OLAP databases can store data structures in the form of multidimensional tables, also known as data cubes, which form an essential part of information systems. Manipulating and presenting such information through interactive multidimensional tables and visual graphics support analysts during review (Caron & Daniels, 2013).

Data warehouses and metadata technologies are required to integrate various types of data originating from different subsystems (Chen et al., 2007). Data warehouses are electronic information systems that host data from another application, external subsystem, or source. Data warehouses contribute to database queries and reporting tools because of their ability to analyze data from various databases. Data warehouses are structured to perform analytics, subject-oriented, and clustering operations instantly (Chen et al., 2007). Metadata plays a crucial role in the design, implementation, and maintenance of the data warehouse. It is also used for data organization, querying information, and interpreting results. Location and information about data warehouse system units are recorded as metadata.

The uniform data warehouse platform hosts many data mining methods, as shown in Figure 2.11, and supports OLAP for multidimensional data and decision-making for high-level users (Chen et al., 2007). It includes feature extraction, classification, fault diagnosis, prediction, association rule extraction, statistical analysis, and data mining functions with many algorithms.

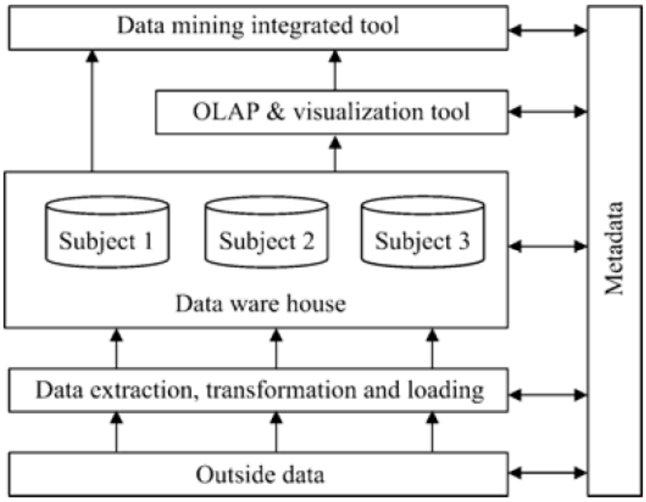


Figure 2.11 Uniform data warehouse platform (Chen et al., 2007)

Zheng et al. (2008) designed an intelligent system to calculate the rational production capacity of an underground metal mine. This system is a meta-synthesis of an artificial neural network (ANN). The entire system includes various subsystems, as shown in Figure 2.12.

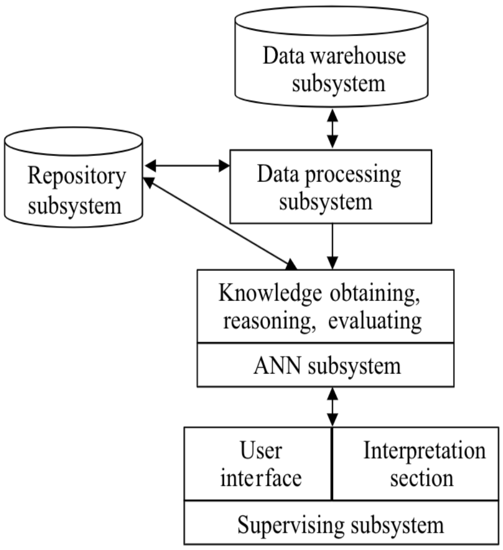


Figure 2.12 Zheng et al. (2008) prepared intelligent system

In this mine optimization design, it is proposed to use a data warehouse subsystem to store raw data, data generated during optimization, and system function indexes. The data processing subsystem is used to convert large amounts of data to ANNs and training samples, while the warehouse subsystem, which is another system, is used to accumulate relevant information. The ANN subsystem serves to validate the scale of the mine’s production. The auditing subsystem is used as a user interface and a section for interpretation (Zheng et al., 2008).

There are limited number of studies using data warehouse, relational databases, and OLAP tools in the mining sector. Chen et al. (2007) proposed a three-layer model for digital communication within a mine through two primary platforms, a uniform transmission network and a data warehouse. The three-layer digital mining model is shown in Figure 2.13.

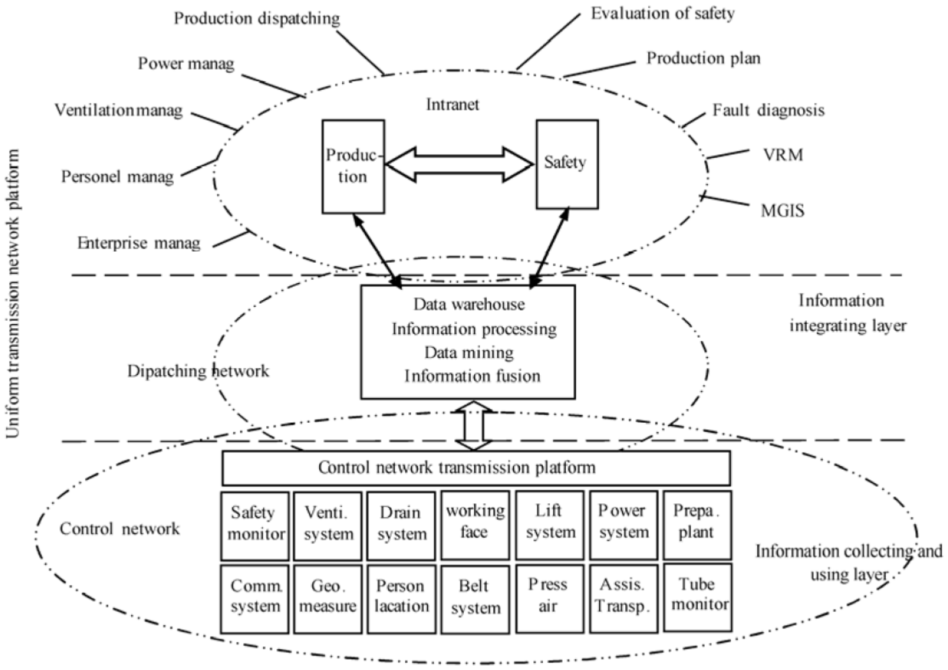


Figure 2.13 Three-layer digital mining model (Chen et al., 2007)

As given in Figure 2.13, the lowest layer is the information gathering and ordering layer. The middle layer is the information integration layer. Each subsystem operates using its own data structure. Data warehouse technology and metadata techniques are key technologies in the information integration layer. The top layer is the management and decision-making layer.

An implementation of data warehousing and data mining with mining-related data is presented by Erkayaoğlu and Dessureault (2019). Operational data from an open-pit copper mine is gathered from several sources such as drill navigation system, fleet management system, and processing plant. A data warehouse is built using these data sources and integrated with each other. After, data mining algorithms are used to predict the fragmentation distribution after blasting operation via online analytical processing (OLAP) cubes. As a result, parameters affecting each stage in the mine-to-mill operation are determined to establish a data-driven framework to help the decision-making process.

Kahraman and Dessureault (2018) created a mine planning database to aid the decision-making process in a control room. The overview of data collection for the database is given in Figure 2.14. The researchers built a data warehouse using historical data kept in OLAP cubes and data produced by a fleet management system. Mine planning spreadsheets, including information about plans, targets, and equipment availability, were combined on a mine planning database. It was observed that right decision-making tools reduced the pressure of the fleet management system controllers and increased the adherence to the mine plan.

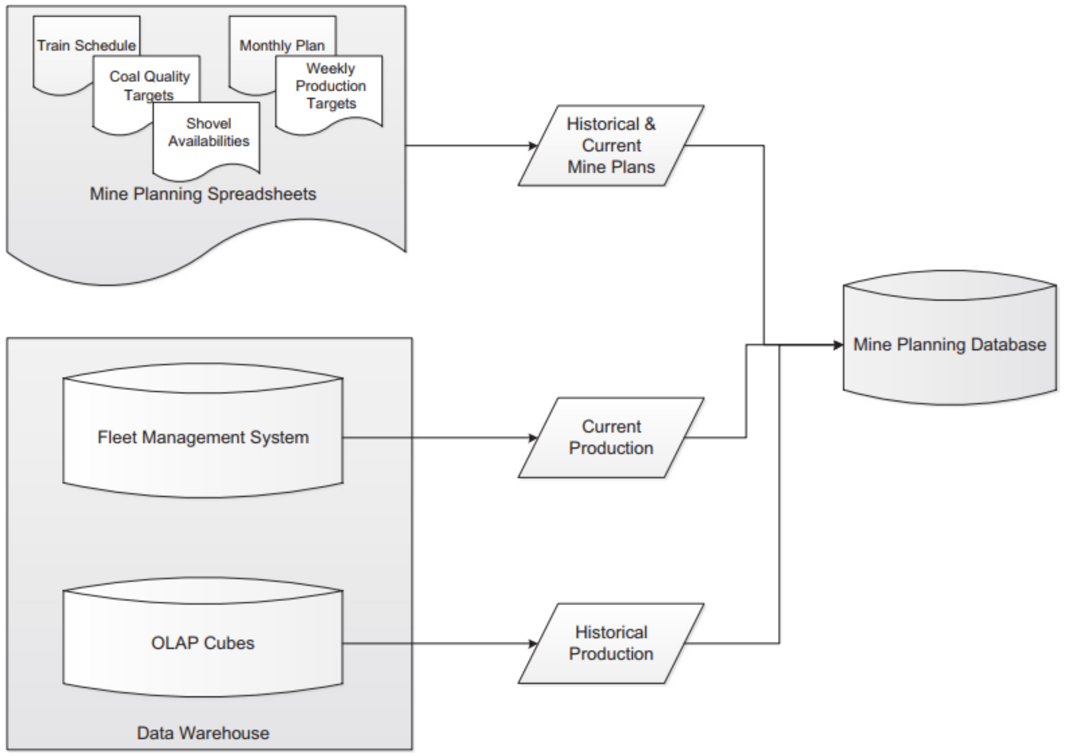


Figure 2.14 Overview of the data collection for the database (Kahraman & Dessureault, 2018)

Another application of data mining on mining-related data is published by Miranda, Sousa, Roggenthen, and Sousa (2012). Geotechnical information gathered from underground works in Portugal is used for knowledge discovery purposes in a database. New regression models were developed using multiple regression and ANN to calculate strength and deformability parameters and rock mass rating index. A similar perspective is also followed in tunnel boring machine (TBM) related data studies.

Several databases were deployed to create models to predict the performance of TBM in different rock conditions and to develop rock mass characteristics in the literature. Yağız (2006) created two different models, including classification and regression tree analysis and multivariate regression analysis to predict the performance of a TBM. Yağız (2008) created a database using rock mechanics data produced by two different sources. These sources are TBM and the rock mechanics laboratory test results, such as UCS, indirect tensile strength, brittleness/toughness. The results obtained in the laboratory and in the field were analyzed statistically, and the effect of rock mass properties on TBM performance was measured. A new empirical formula was created to measure TBM performance utilizing the developed database. Gong and Zhao (2009) conducted a study to develop a rock mass model to observe the effect of rock mass properties on a TBM penetration rate. Another study on TBM performance is conducted by Hassanpour et al. (2011). The data obtained from tunnel projects opened using TBM in hard rock units of different strengths were combined in a database to create a new TBM performance estimation model. Tóth, Zhao, and Einstein (2013) measured the TBM performance in mixed ground conditions. Delisio, Zhao, and Einstein (2013) published an article to analyze and predict TBM performance in blocky rock conditions.

Data generated by the TBM is an example to stream data. Since streaming data is continuously generated, a database or a data warehouse is needed to manage, manipulate, and analyze the data due to the large volume. Thus, a database or a data warehouse were utilized TBM-related studies presented in this literature review.

Comprehensive studies integrated with different data types are limited in order to obtain concrete and reliable information from rock mechanics test results. This requires the development of a data warehouse. While the data warehouse provides an infrastructure that allows the analysis of data transferred from different systems, databases make it possible to access the data by ensuring that the data can be easily and securely stored. Databases create the necessary infrastructure for updating historical data already recorded and efficiently storing large amounts of data. At the same time, it might ensure data security by making data accessible only to authorized users. In addition to these, it ensures that the data is stored correctly and consistently, with the predefined data integrity constraints in the database. It also ensures that the access and search of the data in the database are straightforward and understandable by querying.

CHAPTER 3

DATABASE AND DATA WAREHOUSE DEVELOPMENT

3.1 Database Development

A database is a collection of data that are linked together. The term “data” refers to known facts that can be recorded and have implicit meaning. The presented database definition is fairly broad. For example, one may consider the words that make up this page of text to be connected data, and therefore a database. However, the term database is frequently used in a more limited sense. The following are the implicit properties of a database (Elmasri & Navathe, 2017):

- A database is a representation of some component of the real world, sometimes known as the mini-world or the discourse universe (UoD). The database keeps track of changes to the mini-world.
- A database is a rationally arranged collection of data that has some meaning. A database cannot be referred to be a random collection of data.
- For a specific purpose, a database is created, developed, and populated with data. It has a target audience and some pre-determined applications that these users are interested in.

There are several types of databases, such as relational, noSQL, and columnar. Although noSQL databases have recently become popular due to their flexibility while storing unstructured data, the relational database type was used in this study after considering the rock mechanics experiments test results structure and conducting literature research.

The methodology followed in the relational database development can be presented as the following (Silberschatz, Korth & Sudarshan, 1997):

- Data gathering, cleaning and characterization
- Conceptual design
- Logical design
- Physical design

3.1.1 Data Gathering, Cleaning, and Characterization

A commonality of all the presented data analysis systems in Chapter 2 is that the studies began with collecting and preprocessing data. Similarly, this study starts with data collection and cleaning. The results of the experiments carried out in the METU Mining Engineering Rock Mechanics Laboratory from the year 2000 to the present were compiled and any errors and typos were corrected by using various data cleaning methods.

To begin, a static spreadsheet file was used to compile and consolidate the data. In the defined static raw data file, different worksheets were created for each experiment, and columns were defined for the expected outputs related to the experiment. With the increase in the number of experiments performed, it has been observed that the results of the experiments have been reported in different ways throughout this time period. In these cases, in order to prevent data loss, additional columns were added to the worksheets to include those experiment results. In addition, the existing rock types were divided into rock classes as metamorphic, igneous, and sedimentary at this stage. This classification enabled the analysis to be performed according to the generalized rock types in the RMDW.

Once the data entry was completed, the next step was the data cleaning and extraction process. To start, test results using different measurement units were gathered under a standard measurement unit. For example, while the elastic modulus value is shared with the GPa unit by some researchers, it is shared with the MPa unit by others. Such measurement unit conversions were revised to make them common in order to ensure the accuracy of future mathematical operations. Likewise, the thousands and decimal separators needed to be standardized, as the separators used in the reports differ depending on the language. For this purpose, all of the entered data were revised and standardized as a comma for the thousands separator and point for the decimal number. During this process, it was observed that some test results were left blank for various reasons. These reasons were digitally defined so that they can be separated from the database with certain filters. The numerical values used for the missing data are listed in Table 3.1.

Table 3.1 Numerical equivalents of missing results in experiments

Description not available.	-111
The test result could not be determined because the sample has failed from the crack, or the sample was not suitable for testing.	-222
The result of the experiment could not be determined healthy.	-333
The failure has occurred from the filling.	-444
Irregular failure has occurred.	-555

During the collection of the data, certain anomalies in the experiment results that occurred due to the researchers' post-experiment report and similar reasons were examined. This step is necessary to increase the reliability of the data entered into the database. Thus, a reliable data set could be presented to the researchers during the use of the database for advanced analysis. For this stage, the experimental results were analyzed with histograms and box plots. Mean, maximum, and minimum values were calculated for each parameter in the experiment results. Errors and typos such as negative failure load values or outliers beyond the limit of the testing equipment were corrected. In Figure 3.1, the histogram of the UCS test results before populating the related table in the database is shared as an example. It is observed that most of the UCS values are between 0-100 MPa. A set of codes that filtered the experimental results in the Python programming language using Python libraries such as pandas, matplotlib, and NumPy were used to prepare the histograms.

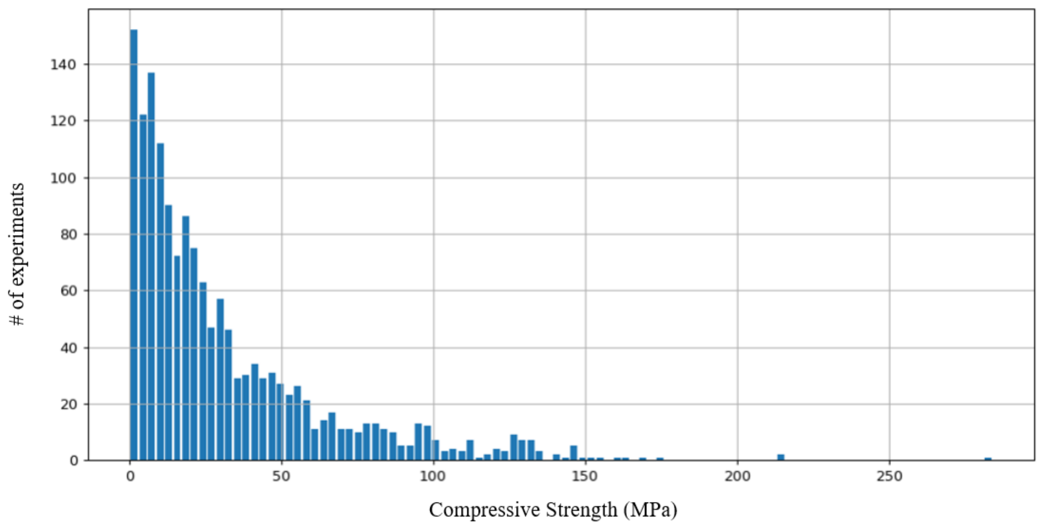


Figure 3.1 Distribution of UCS test results at the end of the data gathering, cleaning and characterization stage

3.1.2 Conceptual Design of the Database

The conceptual design phase was completed before defining the tables to be used in the database and the relations between the tables. Data model for rock mechanics laboratory results sharing (Li, Wang & Zhu, 2012) for the ISRM were taken into consideration in order to ensure that the model created at this stage complies with these suggested methods. During the data-gathering stage, some data was discovered to be non-conforming to the suggested data model. Therefore, the tables were redefined to include new columns covering this data. These add-ons are essential for the database to provide the necessary flexibility to researchers in the future. The ER diagram was created for the RMDB considering the literature. The simplified version of the ER diagram is shared in Figure 3.2. The connection of the experiments tables (uniaxial/triaxial compressive strength, static deformability, density and porosity, indirect tensile strength, direct shear, slake durability and point load) to information tables (experiment information, ID table, and rock types) can be seen in the ER diagram.

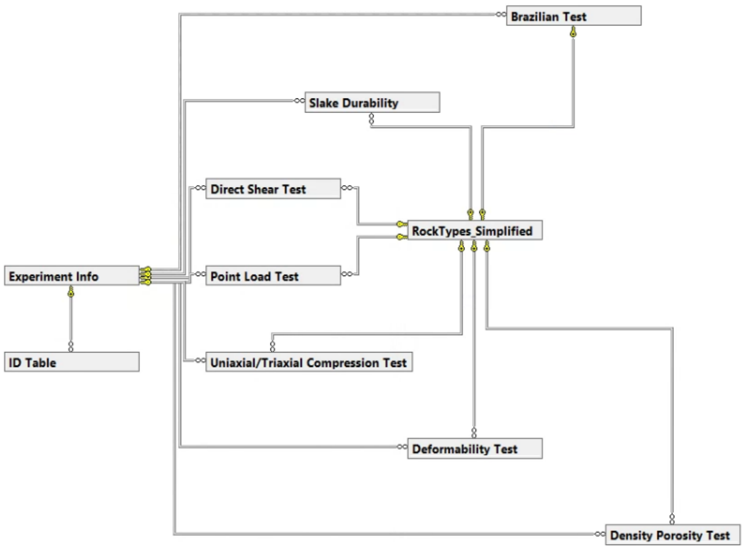


Figure 3.2 Simplified ER diagram of RMDB

In every experiment table in the database, properties about the intact rock sample are kept, such as id of the sample, drill hole name, length and diameter of the sample, depth of the sample, and rock type. Depending on the experiment, other properties are also added to the tables. In addition, each table in the database is shared with their data type in Appendix A.

3.1.3 Logical Design of the Database

Tables were created in the SSMS software in this stage. Keys are defined in order to establish connections between tables. For each sample, a key consisting of the experiment year, type, and number of experiments was created. The static deformability, Brazilian, and experiment information tables are given in Figure 3.3.

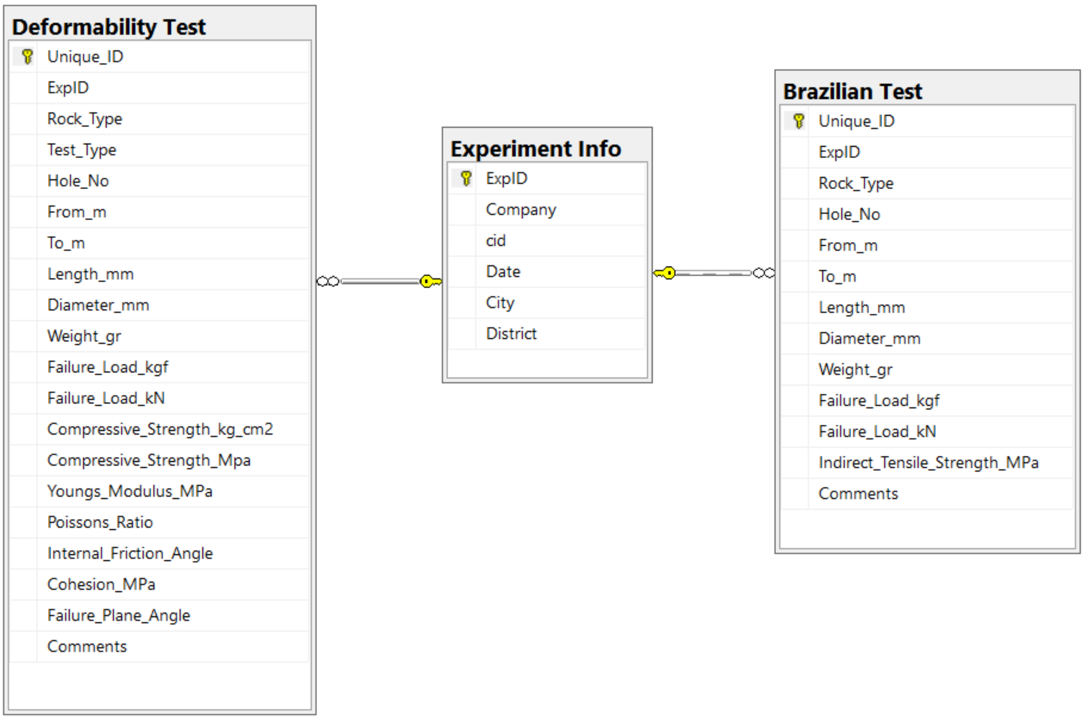
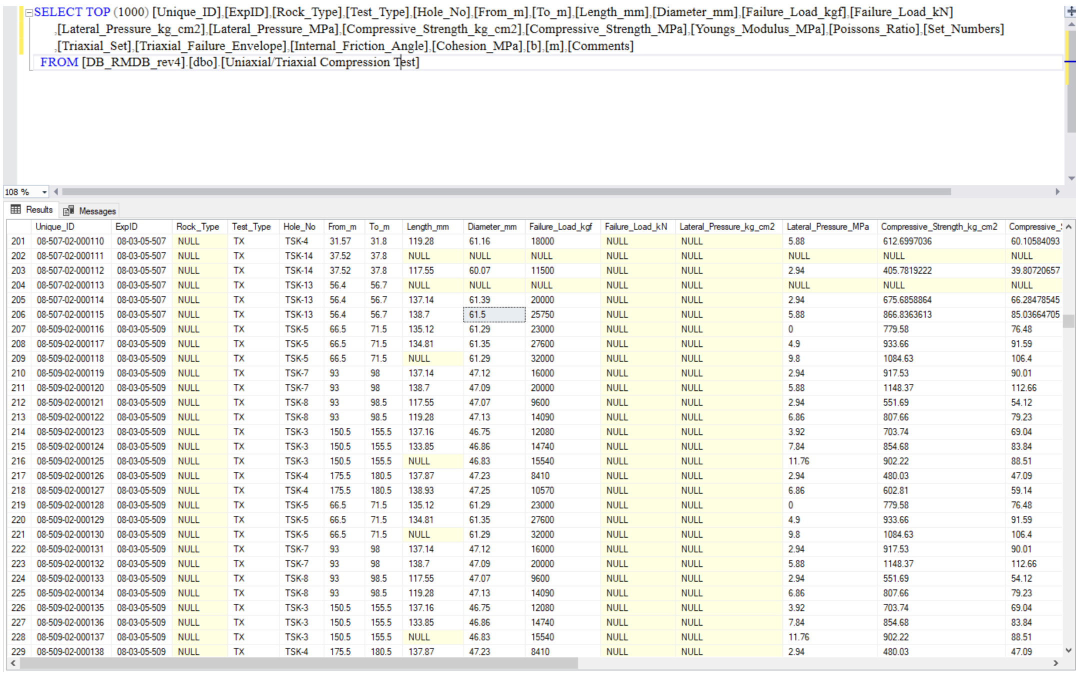


Figure 3.3 Display of deformability test, experiment information, and Brazilian Test tables of RMDB

While deformability and Brazilian test tables include the properties of the intact rock sample, and results of the experiments, the experiment information table includes metadata about the experiment and experiment ID.

3.1.4 Physical Design of the Database

The data, which was digitized and cleaned in the data collection stage, were transferred to the tables created in the database, considering the data types. The indexing process has been completed in order to improve the query performance and the query times of the possible data sources to be added in the future. The result of a query performed on the uniaxial/triaxial compressive strength table is given in Figure 3.4.



```
SELECT TOP (1000) [Unique_ID],[ExpID],[Rock_Type],[Test_Type],[Hole_No],[From_m],[To_m],[Length_mm],[Diameter_mm],[Failure_Load_kgf],[Failure_Load_kN]
      ,[Lateral_Pressure_kg_cm2],[Lateral_Pressure_MPa],[Compressive_Strength_kg_cm2],[Compressive_Strength_MPa],[Youngs_Modulus_MPa],[Poissons_Ratio],[Set_Numbers]
      ,[Triaxial_Set],[Triaxial_Failure_Envelope],[Internal_Friction_Angle],[Cohesion_MPa],[b],[m],[Comments]
  FROM [DB_RMDB_rev4].[dbo].[Uniaxial/Triaxial Compression Test]
```

108 % Results Messages

	Unique_ID	ExpID	Rock_Type	Test_Type	Hole_No	From_m	To_m	Length_mm	Diameter_mm	Failure_Load_kgf	Failure_Load_kN	Lateral_Pressure_kg_cm2	Lateral_Pressure_MPa	Compressive_Strength_kg_cm2	Compressive_
201	08-507-02-000110	08-03-05-507	NULL	TX	TSK-4	31.57	31.8	119.28	61.16	18000	NULL	NULL	5.88	612.6997036	60.10584093
202	08-507-02-000111	08-03-05-507	NULL	TX	TSK-14	37.52	37.8	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
203	08-507-02-000112	08-03-05-507	NULL	TX	TSK-14	37.52	37.8	117.55	60.07	11500	NULL	NULL	2.94	405.7819222	39.80720657
204	08-507-02-000113	08-03-05-507	NULL	TX	TSK-13	56.4	56.7	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
205	08-507-02-000114	08-03-05-507	NULL	TX	TSK-13	56.4	56.7	137.14	61.39	20000	NULL	NULL	2.94	675.6858864	66.28478545
206	08-507-02-000115	08-03-05-507	NULL	TX	TSK-13	56.4	56.7	138.7	61.5	25750	NULL	NULL	5.88	866.8363613	85.03664705
207	08-509-02-000116	08-03-05-509	NULL	TX	TSK-5	66.5	71.5	135.12	61.29	23000	NULL	NULL	0	779.58	76.48
208	08-509-02-000117	08-03-05-509	NULL	TX	TSK-5	66.5	71.5	134.81	61.35	27600	NULL	NULL	4.9	933.66	91.59
209	08-509-02-000118	08-03-05-509	NULL	TX	TSK-5	66.5	71.5	NULL	61.29	32000	NULL	NULL	9.8	1084.63	106.4
210	08-509-02-000119	08-03-05-509	NULL	TX	TSK-7	93	98	137.14	47.12	16000	NULL	NULL	2.94	917.53	90.01
211	08-509-02-000120	08-03-05-509	NULL	TX	TSK-7	93	98	138.7	47.09	20000	NULL	NULL	5.88	1148.37	112.66
212	08-509-02-000121	08-03-05-509	NULL	TX	TSK-8	93	98.5	117.55	47.07	9600	NULL	NULL	2.94	551.69	54.12
213	08-509-02-000122	08-03-05-509	NULL	TX	TSK-8	93	98.5	119.28	47.13	14090	NULL	NULL	6.86	807.66	79.23
214	08-509-02-000123	08-03-05-509	NULL	TX	TSK-3	150.5	155.5	137.16	46.75	12080	NULL	NULL	3.92	703.74	69.04
215	08-509-02-000124	08-03-05-509	NULL	TX	TSK-3	150.5	155.5	133.85	46.86	14740	NULL	NULL	7.84	854.68	83.84
216	08-509-02-000125	08-03-05-509	NULL	TX	TSK-3	150.5	155.5	NULL	46.83	15540	NULL	NULL	11.76	902.22	88.51
217	08-509-02-000126	08-03-05-509	NULL	TX	TSK-4	175.5	180.5	137.87	47.23	8410	NULL	NULL	2.94	480.03	47.09
218	08-509-02-000127	08-03-05-509	NULL	TX	TSK-4	175.5	180.5	138.93	47.25	10570	NULL	NULL	6.86	602.81	59.14
219	08-509-02-000128	08-03-05-509	NULL	TX	TSK-5	66.5	71.5	135.12	61.29	23000	NULL	NULL	0	779.58	76.48
220	08-509-02-000129	08-03-05-509	NULL	TX	TSK-5	66.5	71.5	134.81	61.35	27600	NULL	NULL	4.9	933.66	91.59
221	08-509-02-000130	08-03-05-509	NULL	TX	TSK-5	66.5	71.5	NULL	61.29	32000	NULL	NULL	9.8	1084.63	106.4
222	08-509-02-000131	08-03-05-509	NULL	TX	TSK-7	93	98	137.14	47.12	16000	NULL	NULL	2.94	917.53	90.01
223	08-509-02-000132	08-03-05-509	NULL	TX	TSK-7	93	98	138.7	47.09	20000	NULL	NULL	5.88	1148.37	112.66
224	08-509-02-000133	08-03-05-509	NULL	TX	TSK-8	93	98.5	117.55	47.07	9600	NULL	NULL	2.94	551.69	54.12
225	08-509-02-000134	08-03-05-509	NULL	TX	TSK-8	93	98.5	119.28	47.13	14090	NULL	NULL	6.86	807.66	79.23
226	08-509-02-000135	08-03-05-509	NULL	TX	TSK-3	150.5	155.5	137.16	46.75	12080	NULL	NULL	3.92	703.74	69.04
227	08-509-02-000136	08-03-05-509	NULL	TX	TSK-3	150.5	155.5	133.85	46.86	14740	NULL	NULL	7.84	854.68	83.84
228	08-509-02-000137	08-03-05-509	NULL	TX	TSK-3	150.5	155.5	NULL	46.83	15540	NULL	NULL	11.76	902.22	88.51
229	08-509-02-000138	08-03-05-509	NULL	TX	TSK-4	175.5	180.5	137.87	47.23	8410	NULL	NULL	2.94	480.03	47.09

Figure 3.4 Example of a query showing the first 1000 data in the uniaxial/triaxial compressive strength table

The result of the query given in Figure 3.4 shows how data is stored in the uniaxial/triaxial table. It can be seen that rock type is not shared in the given experiment. This was also common in other different experiments. Rock type of the intact rock samples was missing nearly 85 % of the recorded experiments.

3.2 Data Warehouse Development

A data warehouse is a system that periodically gathers and secures data from many sources and stores it in a dimensional or normalized data storage. It usually contains historical data and can be requested for data analysis. Data warehouses, unlike databases, are not updated with each transaction. In general, updates are made in batches at regular intervals, with the low-density times being preferred. (Rainardi, 2008)

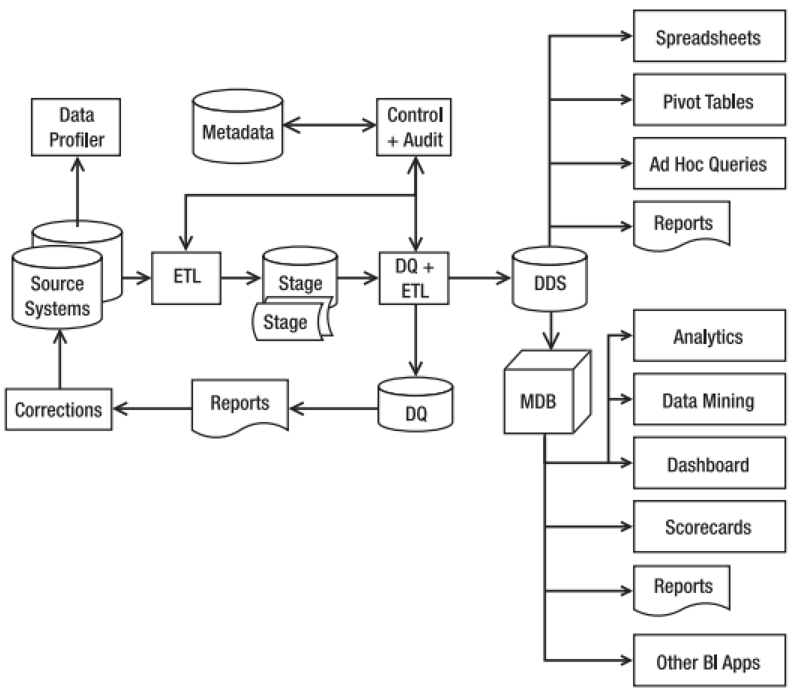


Figure 3.5 A diagram of a data warehouse system (Rainardi, 2008)

A diagram of a data warehouse system is given in Figure 3.5. When we inspect the diagram from left to right, data is loaded from source systems such as databases with extract, transform and load (ETL) system. ETL refers to a system that can connect to source systems, read data, transform it, and load it into a target system such as dimensional data source (DDS). The reason why DDS is used in a data warehouse system is that DDS data can be modified according to needed analysis.

A data warehouse system, RMDW, was developed to analyze the data stored in RMDB. SQL Server Data Tools is used to create the DDS. SQL Server Data Tools software (SSDT) is a development tool that allows designers to create different databases, analysis services data models, integration services packages, and reporting services reports. In order to load to data and to the visual basic environment, a connection is made to RMDB. After tables are loaded to the data warehouse environment, Online Analytical Processing (OLAP) cubes are deployed. View from SSDT is given in Figure 3.6.

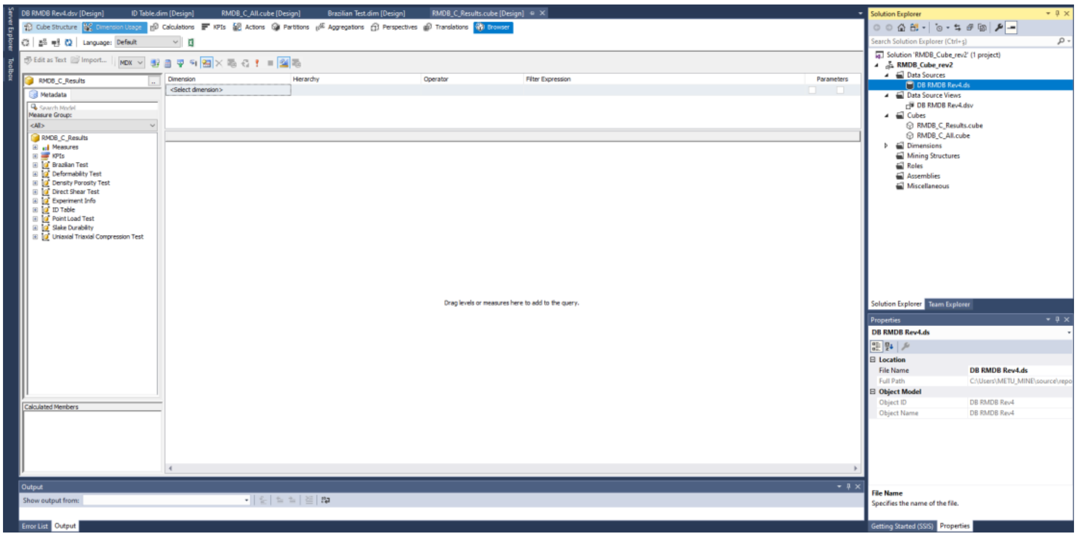


Figure 3.6 A sample view from SSDT

An OLAP cube is a multidimensional data array optimized for rapid data analysis. A three-dimensional cube example consisting of customer, product, and time dimensions is given in Figure 3.7. The term “cube” refers to a multidimensional dataset called a hypercube if the number of dimensions exceeds three. This dataset may come from different and unconnected sources. For two-dimensional data, it is possible to use the traditional method of organizing the data in rows and columns using a spreadsheet, but for multidimensional data, using a spreadsheet may not be the most convenient option.

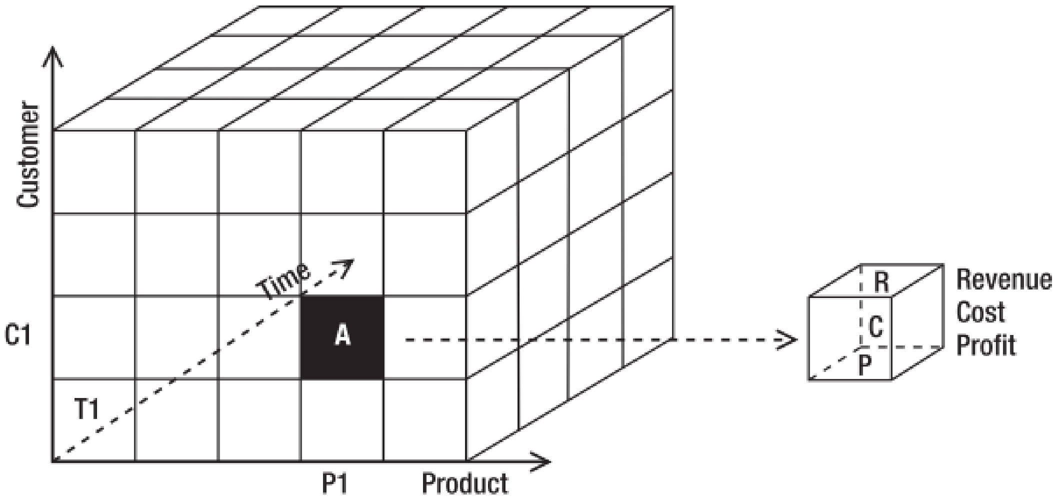


Figure 3.7 Three-dimensional cube example consisting of customer, product and time dimensions (Rainardi, 2008)

OLAP assists users during hardware resource-intensive operations such as grouping, aggregating, and combining data by precomputing and pre-aggregating data to make analysis faster. OLAP databases are divided into one or more cubes designed to facilitate report creation and viewing.

After OLAP cubes are deployed, Microsoft Power BI software is used for the data analysis and visualization. Microsoft Power BI is a business analytics service offered by Microsoft. It provides interactive visualizations and business intelligence capabilities for users to create their own reports and dashboards. It also supports different programming languages such as R and Python, which makes usage of the software flexible. Created dashboards, figures, and tables are presented in Chapter 5.

3.3 Data Mining Case Study

Created data warehouse system is not only used for data visualization but also knowledge discovery purposes to back fill data via data mining. In order to backfill the missing information, a set of experiment results retrieved from the database is used to train a decision tree and random forest model for rock type classification.

3.3.1 Data Analysis

Data of 300 experiments carried out within the scope of a geotechnical project were used in this case study. Since nearly 90 % of the records in the RMDB had no rock type information, it is not possible to successfully backfill the rock type information using all records in the RMDB. That is why a sample project data consisting of only 300 experiments is chosen to measure the performance of the classification.

The types of experiments used in this project are; UCS, triaxial compressive strength, indirect tensile (Brazilian), and density experiments. Experiments were carried out in different rock formations, including ore, phyllite, granite, and hornfels. Conducted experiments numbers according to experiment and rock type used in the case study are given in Table 3.2.

Table 3.2 Conducted experiments numbers according to experiment and rock type used in the case study

Experiment Type	**Ore**	**Phyllite**	**Granite**	**Hornfels**
Static Deformability	16	13	11	36
Triaxial Compressive Strength	12	7	7	24
Indirect Tensile (Brazilian)	7	6	13	22
Density	28	20	18	60
Total (Rock Type)	**63**	**46**	**49**	**142**
Total	**300**			

Different experiments were joined in 76 different records according to depth of the drillhole to classify rock types. These 76 records were divided into train, test and validation sets as 80 %, 10 %, 10 %. Density experiments conducted on different rock types used in the case study had an average density of 3.1 gr/cm^3. The density values were changing between 2.6 gr/cm^3 and 5.0 gr/cm^3. The box plot given in Figure 3.8 shows the distribution of the density according to rock type. It can be seen from the figure that while the rock types with the lowest density are granite and phyllite, the rock type with the highest density is ore.

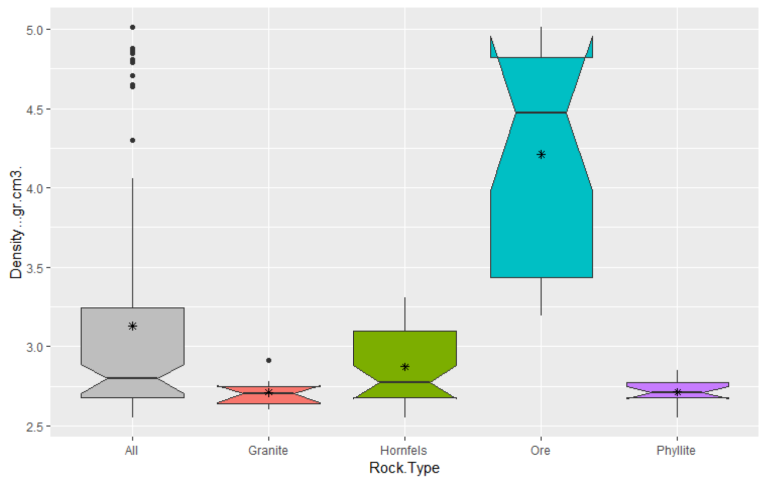


Figure 3.8 Box plot showing the density distribution of the density experiments according to rock type

One of the 76 experiments to be used in the case study was considered invalid due to cracks in the rock. Analyzes were carried out using 75 test results. It is understood from the box plot given in Figure 3.9 that the average UCS of the tests carried out is 100 MPa. While the highest UCS values were observed in the granite rock, the highest strength variability is also seen in this rock type. It can be concluded that the highest average UCS value is in the ore type of rock, followed by granite, hornfels, and phyllite, respectively.

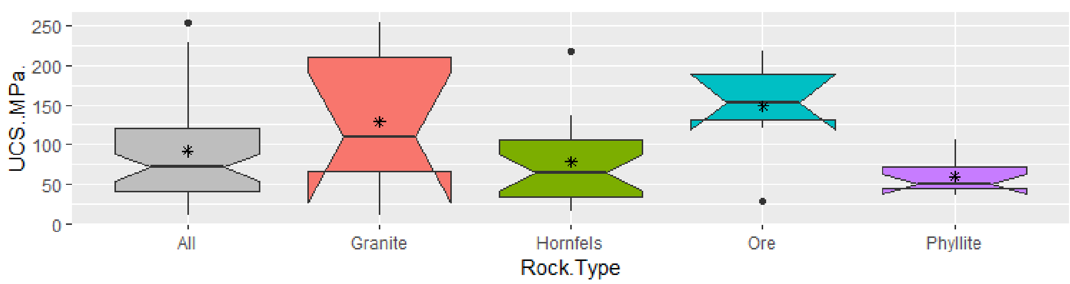


Figure 3.9 Box plot showing the UCS (MPa) distribution of the static deformability experiments according to rock type

When the modulus of elasticity (E) of rocks was examined, it is seen that ore had the highest modulus of elasticity value (Figure 3.10). The variability was found to be very low in this rock type. The average elasticity of all rock types was found as approximately 26 GPa, ranging between 5 GPa to 67 GPa.

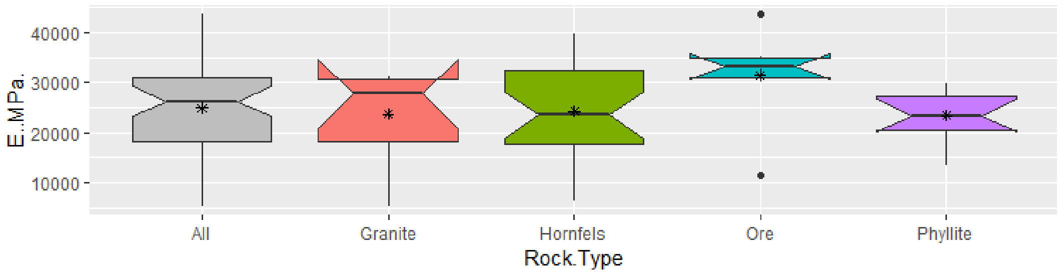


Figure 3.10 Box plot showing the modulus of elasticity (MPa) distribution of the static deformability experiments according to rock type

Another result of the static deformability experiment, the Poisson's ratio, was found to be approximately 0.10 for all rocks. Figure 3.11 shows the distribution of the Poisson's ratio.

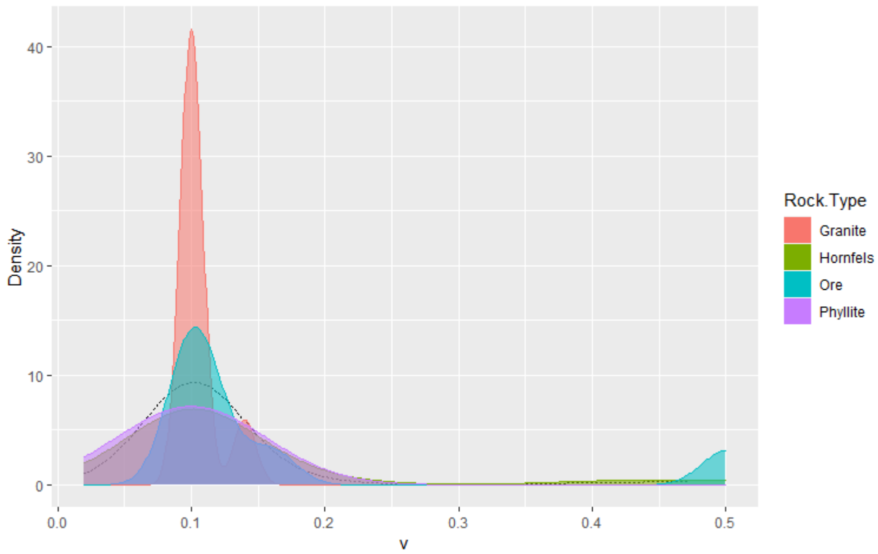


Figure 3.11 Histogram showing the Poisson's ratio distribution of the static deformability experiments according to rock type

Indirect tensile strength experiments, another experiment included in the case study experiments, it is seen that the average of all rock types is approximately 14.5 MPa. Tensile strength values ranged from 5 MPa to 31 MPa. The highest variability is observed in hornfels type. Furthermore, average values of granite, hornfels, and phyllite were similar and were found to be approximately 15 MPa. On the other hand, ore rock type had the smallest average tensile strength value, 11.4 MPa.

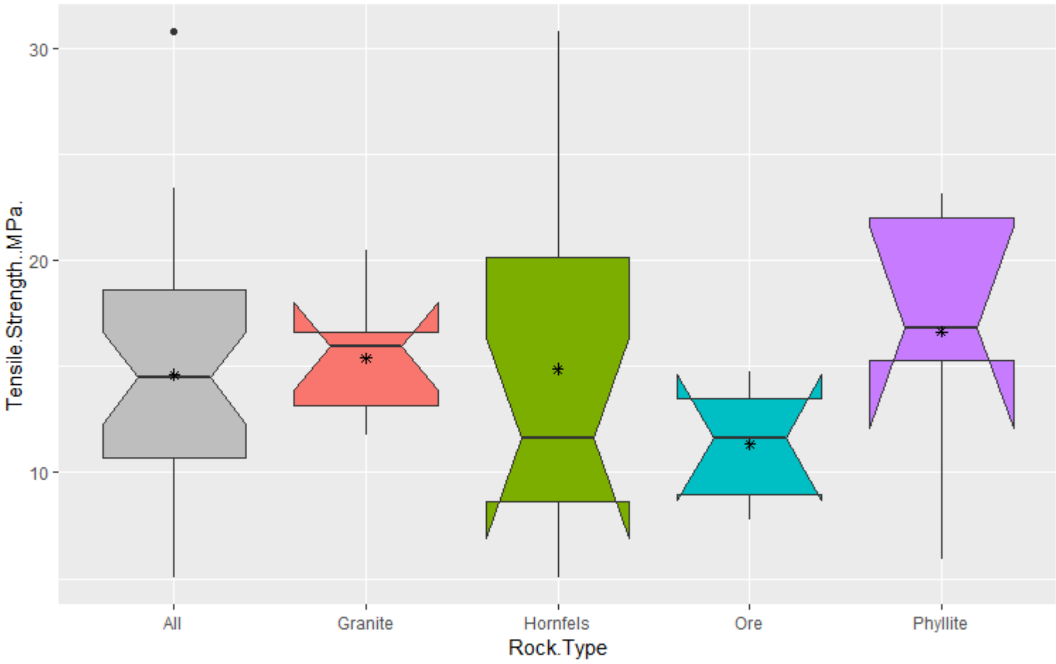


Figure 3.12 Box plot showing the tensile strength (MPa) distribution of the Brazilian experiments according to rock type

The triaxial compressive experiments carried out with 2 MPa, 4 MPa, and 6 MPa confining pressures. When the σ_1 values of all rocks were examined, the values ranged from 49 MPa to 365 MPa, while the average value was recorded as 162 MPa (Figure 3.13). The highest mean value was determined in granite rock with 252 MPa. This rock type was followed by ore, hornfels, and phyllite, respectively.

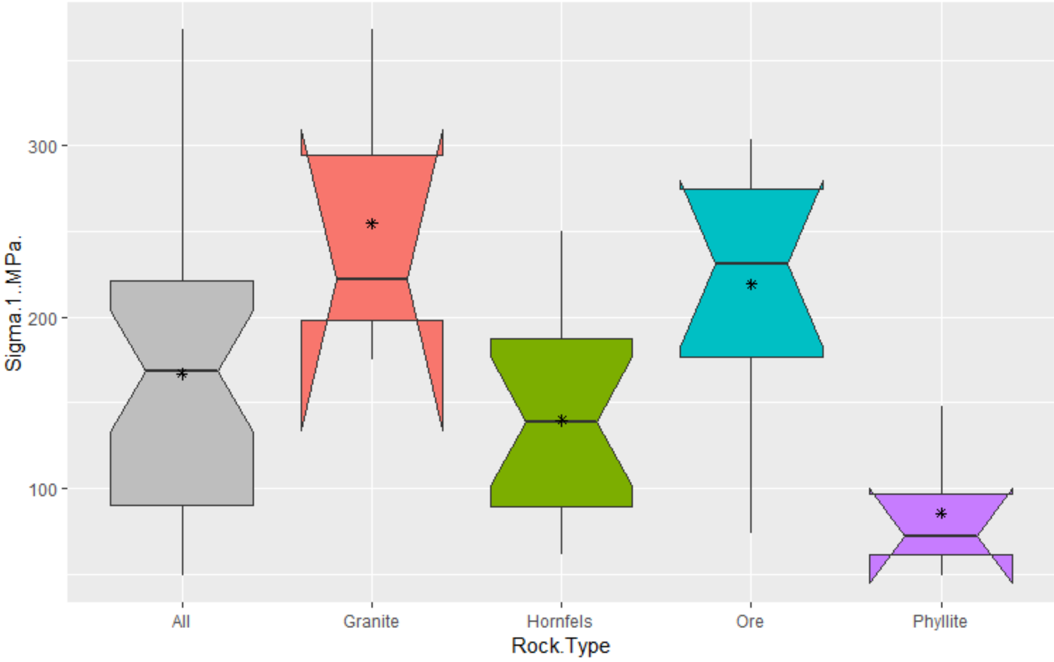


Figure 3.13 Box plot showing the σ_1 (MPa) distribution of the triaxial compressive experiments according to rock type

With the completed several sets of the triaxial compressive test, cohesion (c) MPa and internal friction angle (ϕ) were determined for each rock type with the constructed Mohr's Envelopes. Table 3.3 shows the cohesion and friction angle values of rocks. Cohesion values in ore and phyllite rocks were determined as approximately 14 MPa. This value was found as high compared to granite and hornfels rock. In addition, when internal friction angles are inspected, one can see that values vary between 46 degrees and 69 degrees.

Table 3.3 Cohesion and internal friction angle values of rocks used in the case study

Rock Type	**c (MPa)**	**ϕ (°)**
Ore	14.65	65
Phyllite	14.03	46
Granite	9	68.8
Hornfels	11.28	60.8

3.3.2 Classification Algorithms

After a detailed data analysis, the test results of the static deformability, triaxial compressive strength, indirect tensile (Brazilian), and density experiments were joined according to the lithological properties and the depths from which the intact rock samples were obtained. Combined test results were extracted from the data warehouse and loaded to Rattle software for the data-mining operation.

Rattle is a R-based graphical user interface (GUI) for data mining (Figure 3.14). It provides statistical and visual summaries of data, converts data so that it can be easily modeled, creates both unsupervised and supervised machine learning models from the data, visually displays model performance, and scores new datasets. One of the most valuable features is that all of the interactions with the graphical user interface are saved as a R script that can be run in R without having to use the Rattle interface.

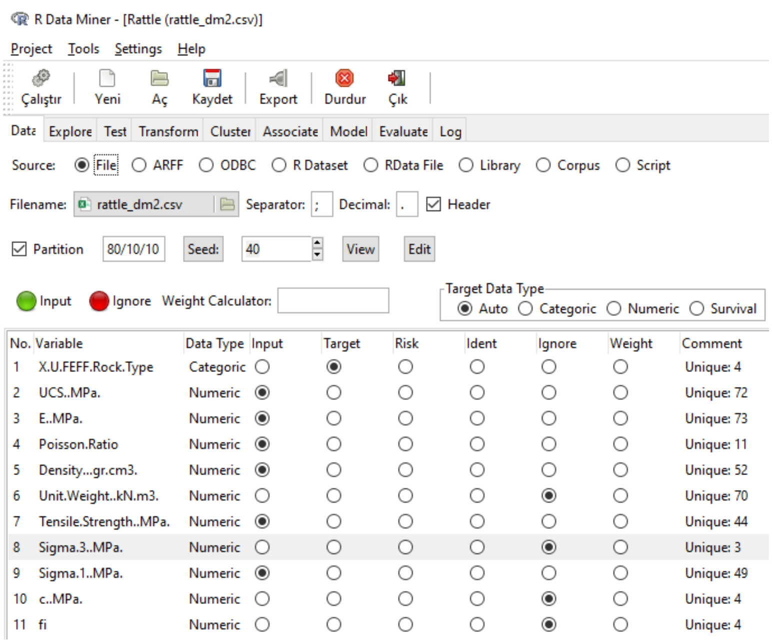


Figure 3.14 GUI and data load window showing parameters used in the case study

Figure 3.14 is a screenshot taken from the data import window in Rattle software showing input and target parameters used in the model. The variable name, data type and use of the variable such as, input, target, risk can be adjusted from this window. The software indicates the number of unique values in the comments column to provide general information about the properties of the variable. Data can be loaded into the software from different sources, such as comma separated file (.csv), text (.txt) and MS Excel (.xlsx, .xls) files. Open Database Connectivity (ODBC) is another way to transfer data to software via a database connection. At the same time, there are sample data sets available that could be used to run data mining algorithms for training purposes. Partitions percentage and seed number can also be adjusted from this window. From the partition pane, loaded data can be randomly divided into certain percentages as train, test, and validation data sets. In addition, seed number allows users to divide the data into the same train, test, and validation data sets.

Data used in this study were randomly split as training, validation, and testing sets, 80 %, 10 %, and 10 %, respectively. The reliability of the data mining model is tested and confirmed with the validation and test data sets.

The main purpose of the case study is to successfully classify the rock types from conducted tests via data mining algorithms. After the rules of classifications are determined, a classifier can be used for backfilling the missing rock type information. As a classifier, a decision tree and random forest classifier were used. These classifiers are the most commonly used classification data mining algorithms in many studies (Ba’abbad et al., 2021). Especially, decision trees are easy to use classifiers that can be interpreted for important predictors or values for each split in the tree.

A decision tree is a recursive split of the instance space that is used to classify data. The decision tree is made up of nodes that create a rooted tree, which is a directed tree with no incoming edges and a node named "root." Each of the other nodes has one incoming edge. An internal or test node is a node with outgoing edges. All additional nodes are referred to as leaves (also known as terminal or decision nodes). Each internal node in a decision tree divides the instance space into two or more subspaces based on a discrete function of the input attribute values (Maimon & Rokach, 2005).

A decision tree given as an example shown in Figure 3.15 is used to determine whether or not a potential client would reply to a direct mailing. Internal nodes are shown as circles, whereas leaves are shown as triangles. The analyst may use this classifier to anticipate a potential customer's reaction and identify the behavioral characteristics of the whole potential customer group when it comes to direct mailing. Each node is labeled with the property it is testing, and its branches are labeled with the values that correspond to that attribute.

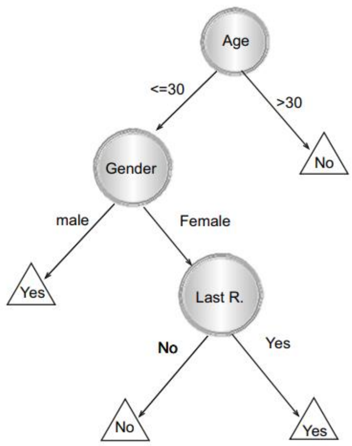


Figure 3.15 A Decision tree example showing direct mailing potential (Maimon & Rokach, 2005)

Decision-tree classifiers are a common classification approach. The reason is that created model is easy to interpret. Figure 3.16 shows the decision tree parameters that can be changed in the software. The minsplit parameter defines the minimum number of observations required at a node in the tree before it may be considered for splitting. The smallest amount of observations that can be made in every decision tree leaf node is known as the min bucket parameter. The depth of the tree is defined as the maximum depth of any node in the final tree. Depth 0 is assigned to the root node. The complexity parameter is used to determine the ideal tree size and regulate the size of the decision tree. Classifier performance can be increased by fine tuning these parameters.

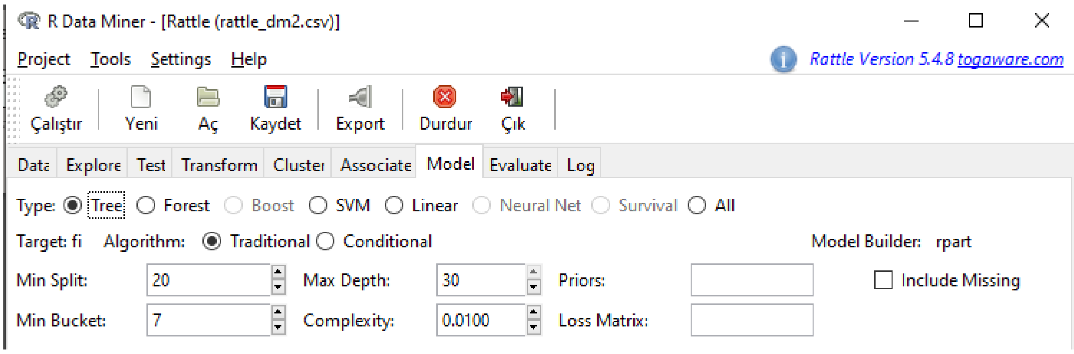


Figure 3.16 Decision tree model interface

Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other problems that works by training a large number of decision trees (Kantardzic, 2011). For classification tasks, the random forest's output is the class chosen by the majority of trees. Ensemble approaches combine many learning algorithms to achieve higher predictive performance than any of the individual learning algorithms could.

Random forest algorithm shows different correlation and strength with changing introduced variable and sample size. Breiman (2001) states that as the number of trees in a forest grows larger, the generalization error converges to a limit. The strength of individual trees in the forest and their connections determine the generalization error of a forest of tree classifiers. The effect of different sample size and number of variables on algorithm strength and correlation have been studied in the related literature. Regardless of the sample size, error was proportional to number of variables, the less the variable the more the error. Furthermore, it is seen that with the increase of the sample size, the algorithm strength increased for a longer time before reaching to the limit. Figure 3.17 shows the effect of number of inputs on larger sample data set. It is observed that correlation and strength increased with increasing number of variables. In addition to this, when the number of inputs was low, the variation of out of bag (OB) and test set error were higher. Variation decreased with the introduction of more inputs to algorithm. Later, errors showed slight linear increase with the increasing number of inputs.

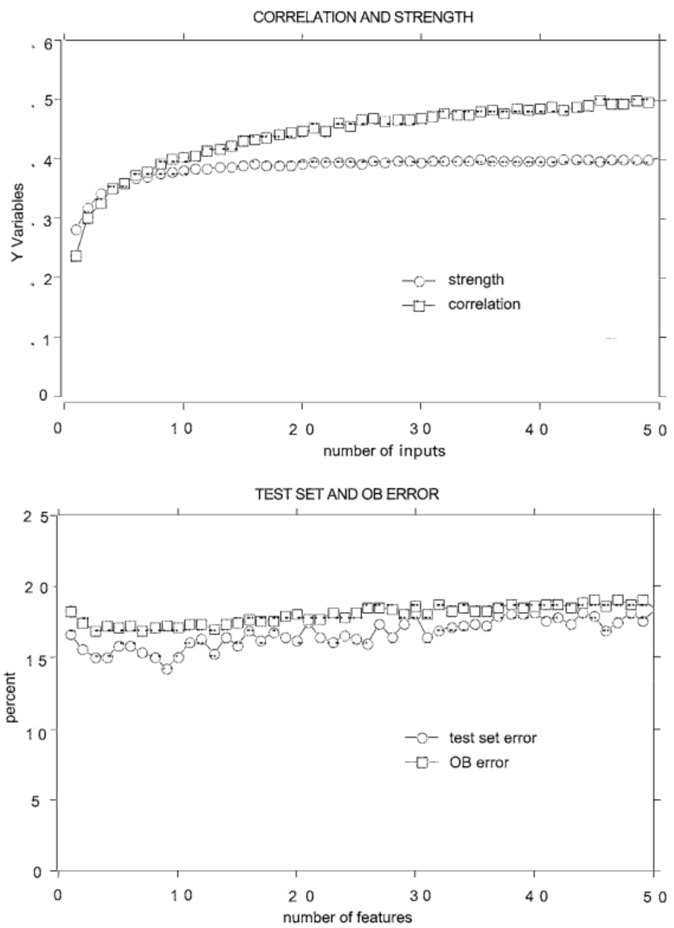


Figure 3.17 Effect of number of inputs on large data set in random forest algorithm

Random forest models can be trained in the Rattle software. The random forest interface of Rattle software enables to change parameters such as number of trees, variables, and sample size to achieve better predictive performance.

The Boosting meta-algorithm is a fast, simple, and straightforward method for creating data mining models. Adaptive boosting has been referred as the "best of the shelf classifier in the world' by Hastie et al (2001). This method creates several models from a dataset by employing a different learning technique that doesn't necessarily the best. This is actually preferred for this type algorithm and often referred as the weak learner. Decision trees are the most common choice to represent the boosting algorithms. In a very simplified manner, boosting is a technique that connects weights with dataset observations and increases (boosts) the weights for those observations that are difficult to predict properly. In order to fine tune the data a series of models can be constructed by simply modifying the weights assigned. An additive model is then obtained as the finished product. One must be careful as boosting algorithms can fail when the data is limited, and models are overly complex (Williams, 2011).

Neural Networks tries to mimic the spirit of neurobiology in artificial networks to create networks and devices in a similar manner to solve computational problems easily (Hopfield, 1988). The initial interest to this method dates to 1940s. Artificial neurons are the basic processing elements for this architecture, which is composed of three different types of layers; input, hidden and output layers. Based on the input signal and characteristic of the neurons the effect is represented by connection weights. By adjusting these weights based on the algorithm the learning ability is gained (Abraham, 2005).

Support Vector Machines (SVM) functions the best on nonlinear, sparse and high dimensional problems. Depending on the tuning options selected, this can be challenging and time intensive. This is the main disadvantage of SVM. On the contrary, dealing with the support vectors rather than the whole dataset makes it less taxing and size of the training set is not a concern. Also, these models are less effected by the outliers, makes it advantageous (Ma, 2014). In order to obtain the vectors in this method, observations are clustered depending on the target variable to create a straight line. It is very seldom that a straight line can separate the observations because data is not distributed in such a way. In that case, the input can be remapped in alternate ways to generate new variables. This increases the chance of creating a sort of gap between observations of different classes. By repeating this process, decisions could be made on which hyperplane represents the observations and therefore the conclusion.

The discussed classification algorithms could also be used in this study, such as boosting algorithms, ANN and SVM. Compared to decision tree and random forest algorithms, interpretation of the result is relatively more difficult to interpret. Because of this reason, only decision tree and random forest algorithms were utilized, and relative performances were measured in this study.

Developed models and their predictive performances are presented in Chapter 4.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Data Analysis in the RMDW

The findings obtained after completing the development of the database, data warehouse and OLAP cubes are presented in this chapter. The map in Figure 4.1 is a heat map showing the locations where the samples were taken from, where the experiments examined in this project were carried out.

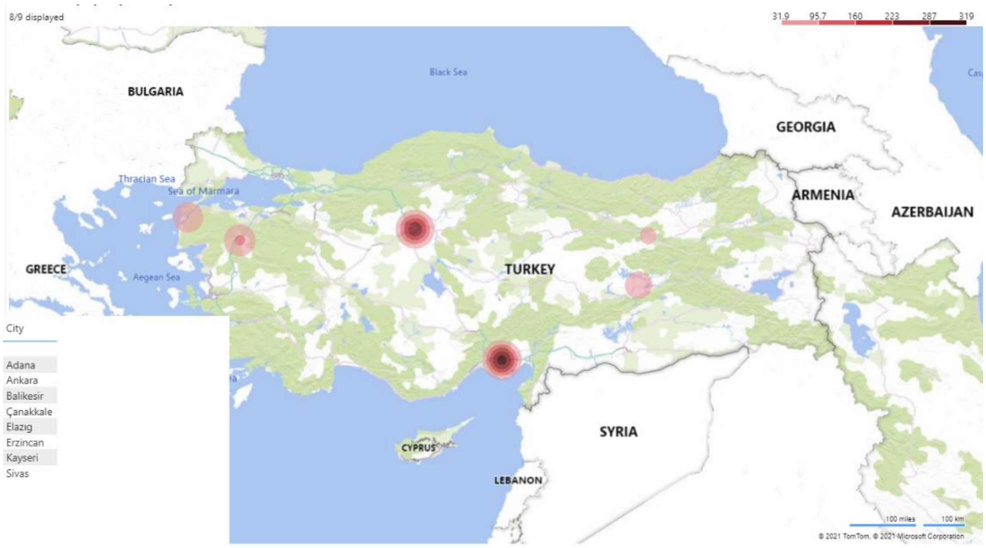


Figure 4.1 Distribution of number of experiments conducted by location

This map gives the researcher visual clues as to the locations of the samples and how they varied according to their density. The change in the radius of each circle represents the density of the number of experiments. Geotechnical projects completed for certain operations can be seen in Figure 4.1 as having comparably higher number of experiments recorded in the data warehouse.

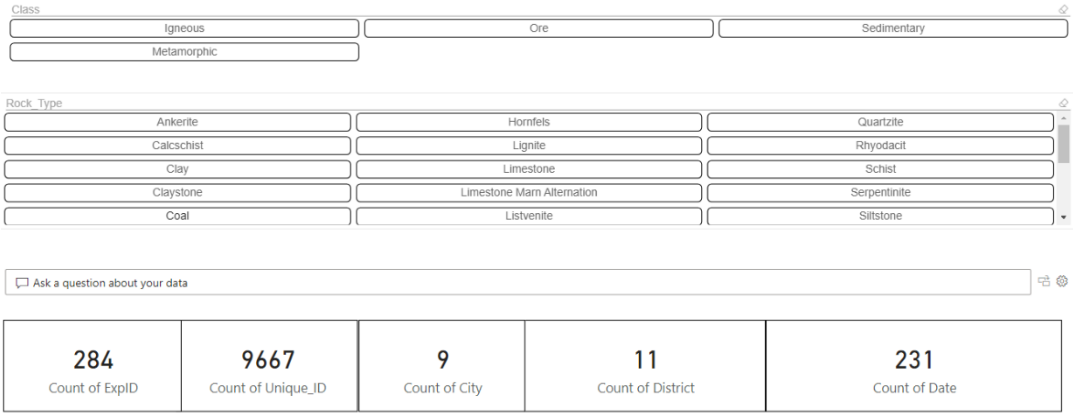


Figure 4.2 User panel with general information in the data warehouse

The panel shared in Figure 4.2 makes it possible to access general information such as the total number of experiments, region, and city in the data warehouse. Users can filter by rock class and type through menus at the top of this panel. Another tool available here is the question box, which makes it possible to query within the data warehouse without the need for any programming language knowledge; it is possible to get answers in this box by asking questions about the experiments verbally. The graph in Figure 4.3 shows the percentage (%) of all experiments performed.

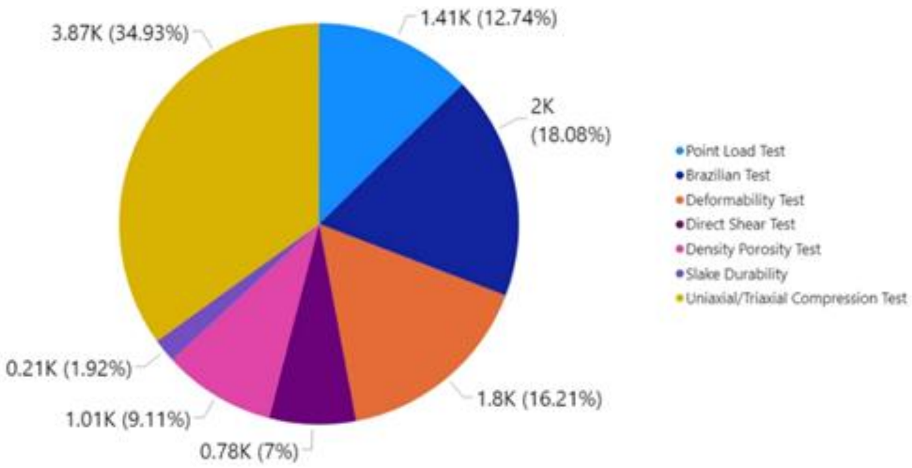


Figure 4.3 Graph showing the distribution of experiments

The most commonly performed experiments carried out between the years 2000 and 2020 are uniaxial/triaxial compression (34.93 %). Other tests include indirect tensile strength (18.08 %), static deformability (16.21 %), point loading (12.74 %), density and porosity (9.11 %), direct shear test (7.00 %), and slake durability (1.92 %). In the panel presented in Figure 4.4, data can be interactively classified according to igneous, metamorphic, and sedimentary rock types.

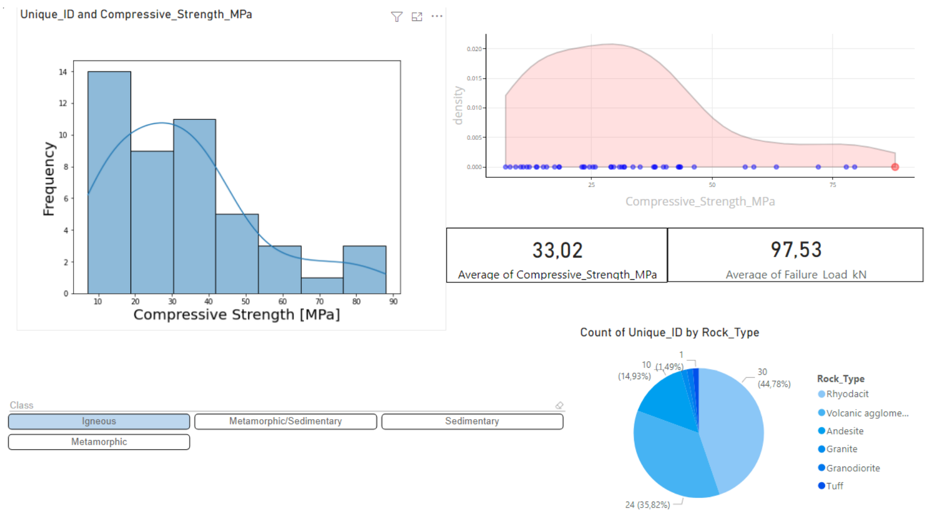


Figure 4.4 User panel of UCS test results

Similarly, after selecting the rock type, it is possible to filter by both these features on the graph. When the igneous rocks were examined, the average compressive strength value was approximately 33 MPa, and the failure load was approximately 100 kN.

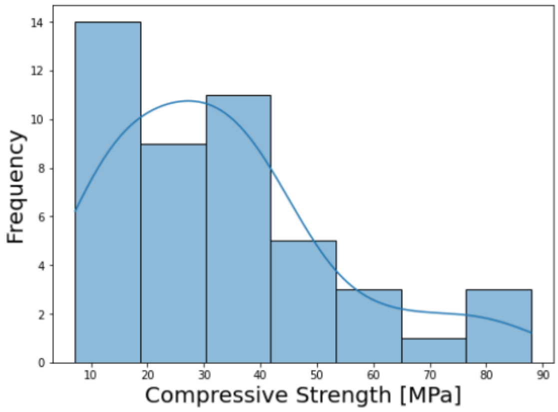


Figure 4.5 UCS test results (MPa) frequency graph of igneous rocks

It is seen that most of the test results range between 10-40 MPa when the UCS (MPa) histogram of the igneous rocks is examined (Figure 4.5). It is possible to conclude that the data do not follow a normal distribution. When the UCS test results were compared, the highest values were determined as the igneous rocks, followed by metamorphic and then sedimentary. Figure 4.6 shows the kernel density graph of UCS test. Extreme values can be determined from this figure.

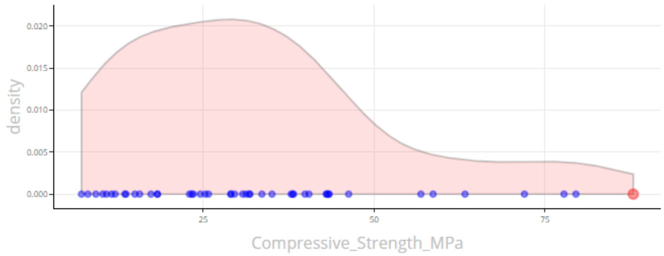


Figure 4.6 UCS test results (MPa) density graph of igneous rocks showing the extreme values

ID 1 and ID 2 values are reported as a part of the results from stability tests against dispersion in water. In this experiment, dry rock grains varying between 40 g and 60 g are rotated 200 times in a water drum within 10 minutes, after which the sample is dried and the percentage of weight loss is recorded as ID 1. The test is then repeated, using the same rock sample. The second weight loss percentage is defined as ID 2. Figure 4.7 represents the user panel developed for the results of this experiment.

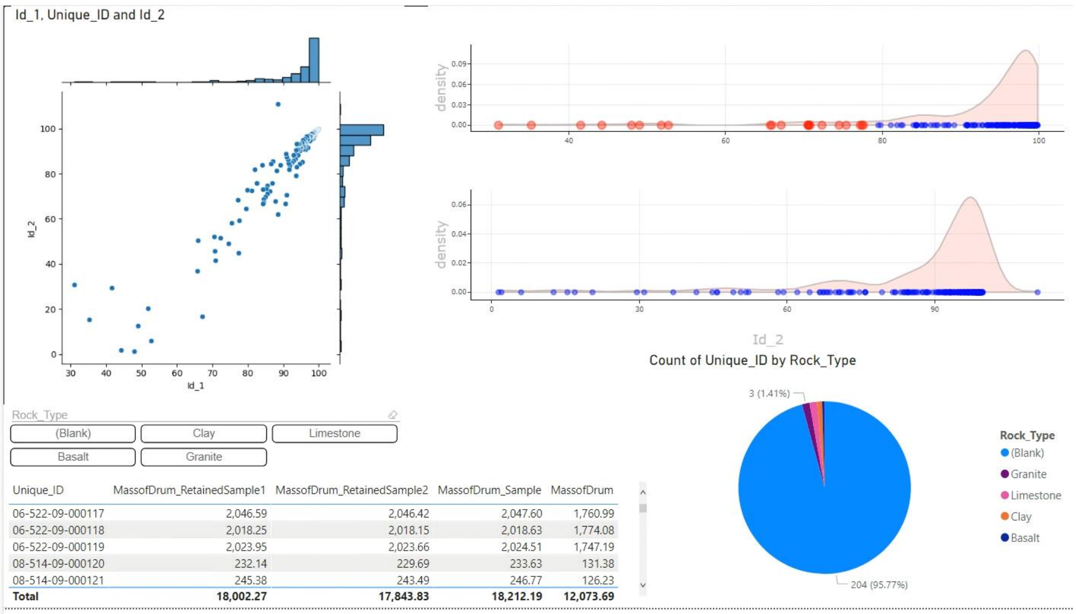


Figure 4.7 User panel for slake durability test results

In the panel shown in Figure 4.7, the upper left graphic was created by the seaborn library in the Python programming language and shows how the ID 1 and ID 2 values change between 0-100. Also, the distribution parameters is presented to the user via the histogram on the graph. The two images in the upper right were created with the kernel density functions of the ID 1 and ID 2 parameters to determine the extreme values through the Tukey algorithm. In addition, it is possible to obtain interactive visuals by switching between various rock types and different experiments through the table containing the rock types and test results below.

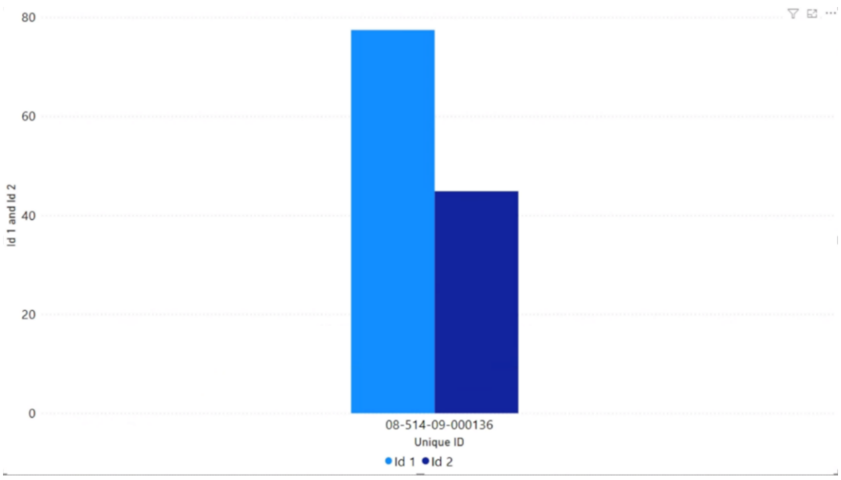


Figure 4.8 Results of a single slake durability test

The stability index (ID 1) obtained from the first cycle is always higher than the stability index (ID 2) obtained from the second cycle, as expected. This was observed for all experiments performed and validates the slake durability test results. Tensile strength and failure load are reported as a result of indirect tensile testing.

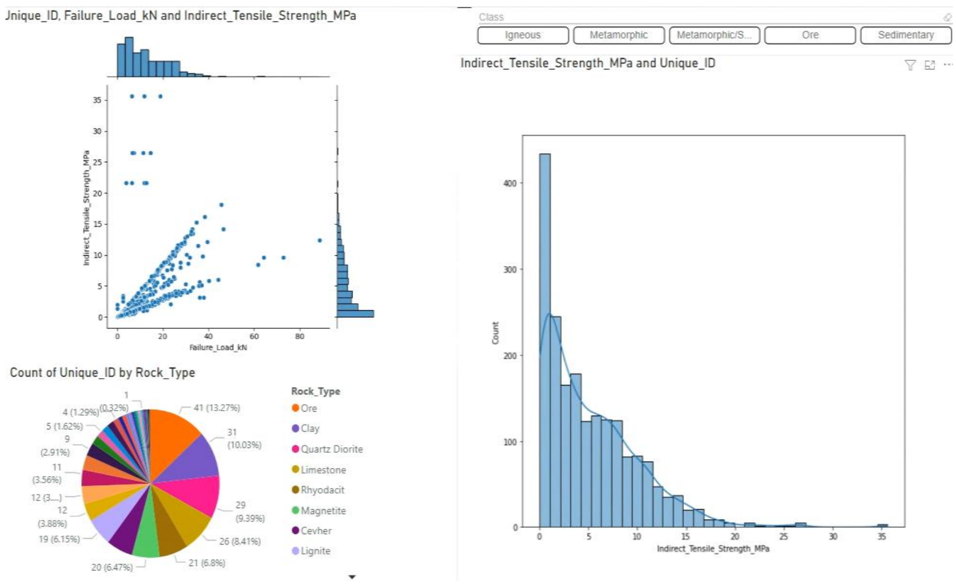


Figure 4.9 User interface of indirect tensile strength test

As shown in Figure 4.9, although the rock type is unknown for 85 % of the experiments carried out, the remaining 15 % contains the results of 35 different rock types. While the indirect tensile strength histogram shows a logarithmic distribution, a linear behavior was observed between failure load and failure strength, as expected. This also represents the data reliability of the test results and could be used as a data quality indicator. The bar graph seen in Figure 4.10 shows the average indirect tensile strength (MPa) of each rock type.

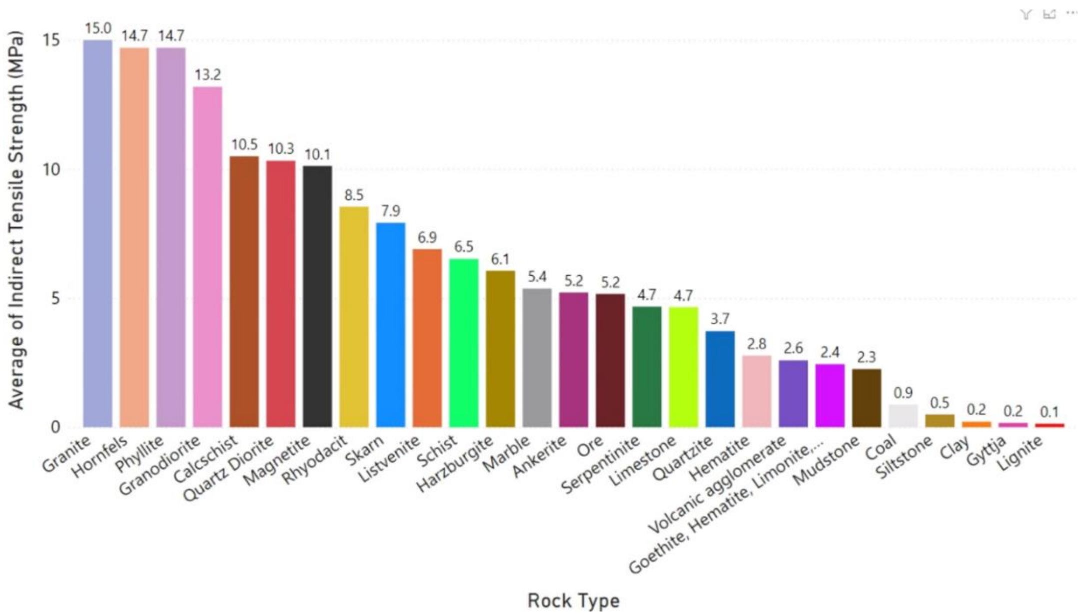


Figure 4.10 Distribution of indirect tensile strengths (MPa) by rock types

According to the indirect tensile strength tests results, granite is the rock type with the highest strength value, followed by hornfels, phyllite, and granodiorite rocks. The rock type with the lowest indirect tensile strength was determined to be lignite based on the available rock type information.

As shown in Figure 4.11, direct shear test results were filtered as sedimentary rocks via an interactive user interface rock class chart.

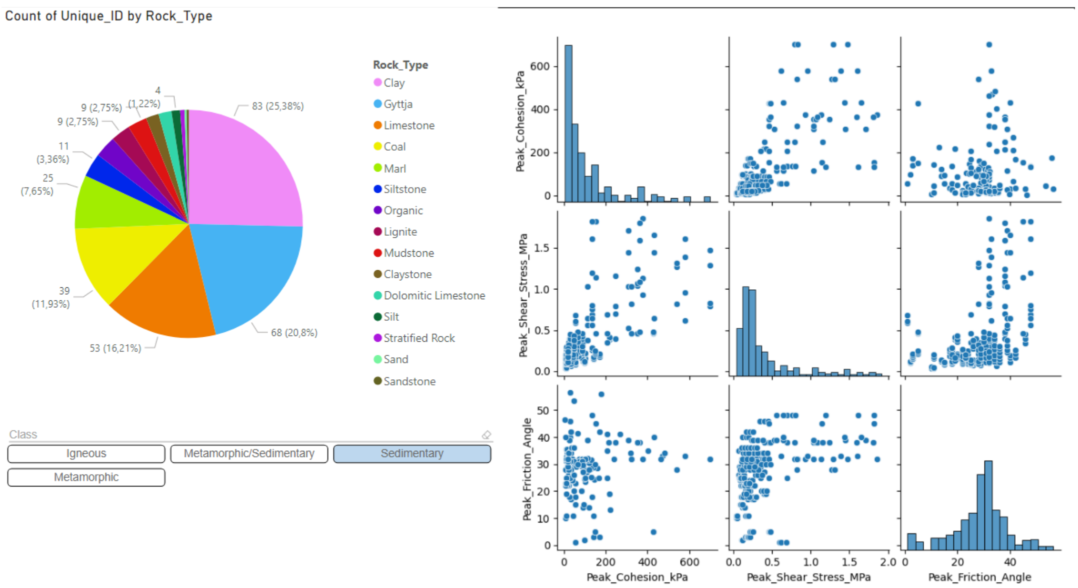


Figure 4.11 User interface of direct shear experiment

As a result, it can be observed that the rock types are dispersed with a similar percentage to each other. This was also observed when no filter was applied. The histogram of each test result can be observed in the graphs located on the diagonal in the graph on the right. The other axes of the same image present information about the relations of two different data with each other. Per the visuals created, no obvious correlation was found between the cohesion and friction angles in the direct shear testing. To analyze the relationship between the point loading index (MPa) and UCS (MPa), a scatter plot was created with all the data in the data warehouse (Figure 4.12).

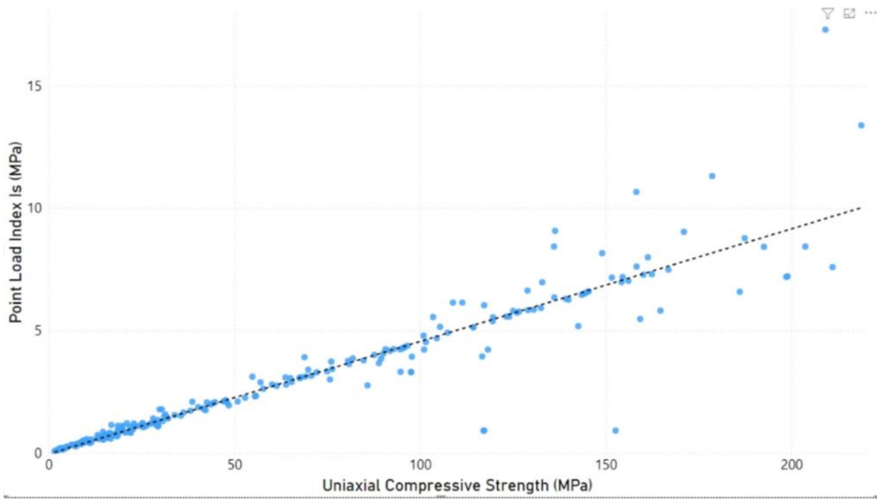


Figure 4.12 Point load index (MPa) – UCS test result (MPa) graph

The expected linear relationship between the point load index and UCS values was clearly observed and conforms to the relationships defined in related literature. Figure 4.13 shows the distribution of rock types for the slake durability tests.

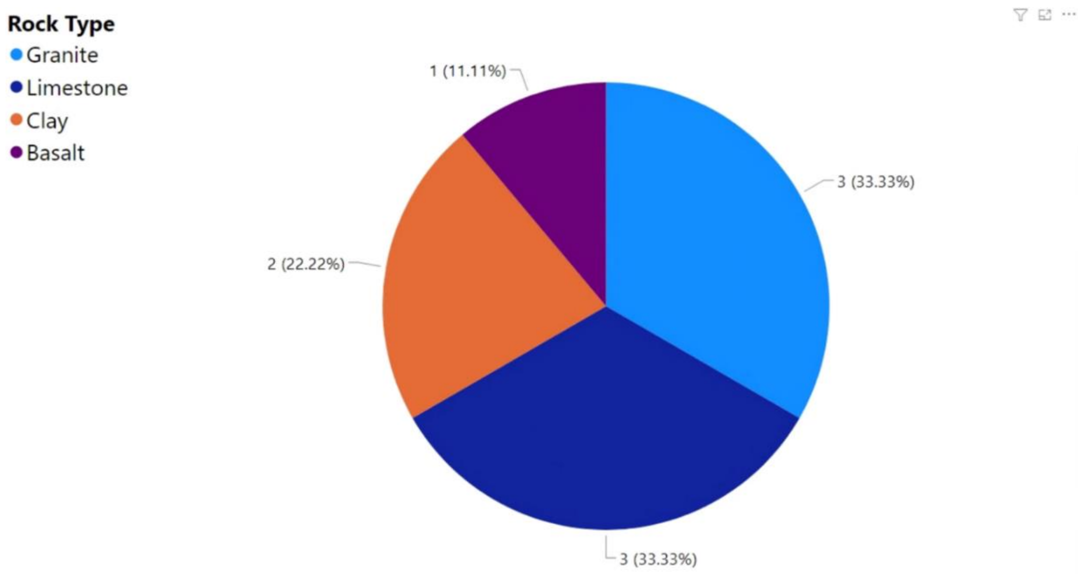


Figure 4.13 Rock type distribution graph of the slake durability test

The most frequently tested rock types in this experiment are granite (33.33 %) and limestone (33.3 3%), followed by clay (22.22 %) and basalt (11.11 %).

The most frequently tested rock type in direct shear tests is clay (19.53 %), followed by gyttja (16.00 %), limestone (12.47 %), coal (9.18 %), rhyodacite (8.47 %), marl (5.88 %), schist (5.18%), diabase (3.29 %), siltstone (2.59 %), lignite (2.12 %), mudstone (2.12%), serpentinite (2.12 %), claystone (1.41 %), dolomitic limestone (1.41%), volcanic agglomerate (0.94 %), granite (0.24 %), agglomerate (0.24 %) and others (3.76%).

The distribution of the rock types in the static deformability test were as follows: coal (25.95 %), clay (16.79 %), quartz diorite (12.98 %), limestone (8.40 %), harzburgite (6.87 %), magnetite (6.11 %), calcschist (3.82 %), granite (3.05 %), lignite (2.29 %), gyttja (1.53 %), marble (1.53 %) and others (10.68 %).

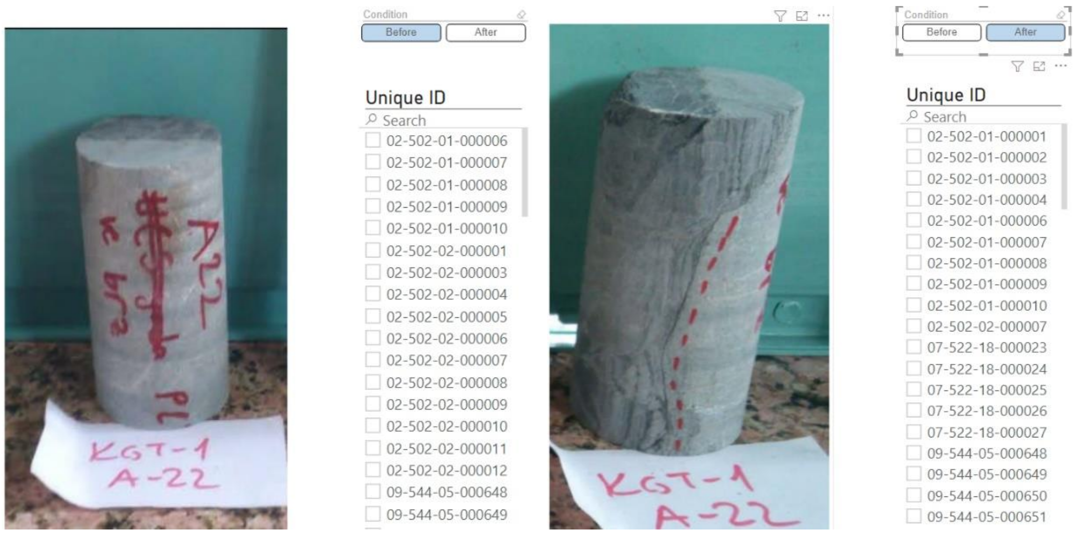


Figure 4.14 Photographs of a rock sample before and after the experiment

In the data warehouse, the test results and sample properties - as well as the pre-test and post-experiment photos of the tested rock samples - can be accessed via the unique key. In Figure 4.14, before and after photographs of a rock sample from an experiment are shown. These pictures can be helpful while extracting visual information from the rock samples, such as crack conditions, lithology, or rupture angle. Expert opinions could be added to the data warehouse as a manual source of data in case the before and after test photos are examined in a systematic manner. Another analysis could based on the size effect for samples to interrelate strength parameters. Figure 4.15 shows the relationship between the rock sample diameter to the compressive strength values.

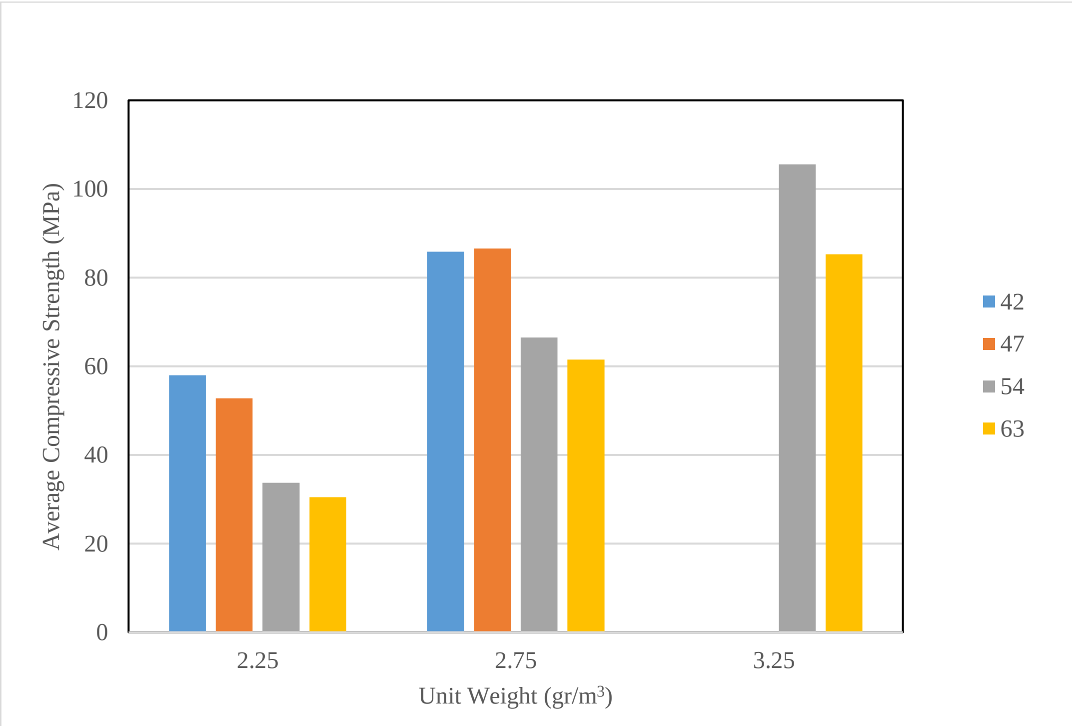


Figure 4.15 Relationship of rock sample diameter (mm) to UCS (MPa) with changing unit weight (gr/m^3) in deformability experiment

A total of 1796 data recordings were reviewed to understand this phenomenon. Out of these, only 386 results included both unit weight and diameter information together with the compressive strength. The sample diameter and density figures are grouped into specific categories by their unit weights to represent rock types.

According to these results, as the sample diameter size increases, the average compressive strength value decreases especially for the diameters 54 mm and 63 mm. This result is observed for all groups of unit weight, in other words for available information about rock types. The size effect concept could be investigated in more detail in case the before and after photos taken for each specimen are analyzed. This way, additional information about the rock types could be identified and the size effect could be compared for the same type of rocks represented by similar test results.

4.2 Results of the Data Mining Case Study

The case study's primary objective was to successfully backfill the rock type information by classifying rock types using data mining techniques based on the results of various experiments. A decision tree and a random forest classifier were employed as classifiers. The variables that were introduced to the software as inputs were UCS (MPa), E (MPa), Poisson's ratio, density (gr/cm^3), unit weight (kN/m^3), tensile strength (MPa) and σ_1(MPa). C (MPa) and φ (°) variables are not included in this model; since these values are calculated specifically for the rock type, it would create a bias toward a successful classifier.

Among the input variables, density and σ_1 were the parameters selected by the algorithm to produce a classification tree. The developed decision tree model is given in Figure 4.16.

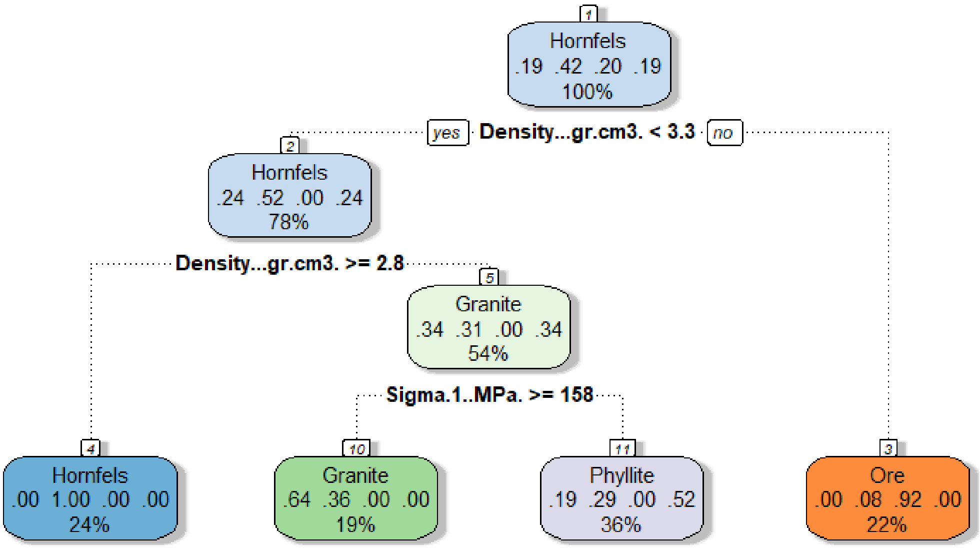


Figure 4.16 Decision tree model developed in the case study

The decision tree algorithm used density values to classify ore and other types of rock. Phyllite, hornfels, and granite rocks were classified using density and σ_1 parameters. Table 4.1 provides the information for the tree constructed in the model. It can be concluded that model errors decrease as the number of splits in the model increase; the largest tree will always yield the lowest relative error rate and the lowest relative error rate will always be found in the largest tree. However, selecting the tree with the lowest relative error is not the best option because this tree will have a bias toward the training set as the number of splits in the tree increases.

Table 4.1 Complexity table of the decision tree model

Tree #	**Complexity parameter**	**# of splits**	**Relative error**	**Cross validation error**	**Standard error**
1	0.32353	0	1	1	0.11164
2	0.11765	1	0.67647	0.67647	0.11018
3	0.01	3	0.44118	0.64706	0.10925

The success of the classification can be measured by the confusion matrix, which shows the correct classification and misclassification of the target variable. Table 4.2 provides information about the success of the decision tree model. This confusion matrix is created using the validation set. It is seen from this table that the algorithm had a 100 % error rate for granite and a 50 % error rate for phyllite. The overall error and averaged class error in this validation were 57.1 % and 37.5 %, respectively.

Table 4.2 Confusion matrix of the decision tree model

	Granite	Hornfels	Ore	Phyllite	Error (%)
Granite	0	1	0	2	100
Hornfels	0	1	0	0	0
Ore	0	0	1	0	0
Phyllite	0	1	0	1	50

Another algorithm used in the case study was the random forest algorithm. The random forest algorithm uses a multitude of decision algorithms to improve classification performance. As there are many decision trees in this algorithm, visualization of the tree is not possible.

The confusion matrix of the random forest model is given in Table 4.3. The shared matrix shows that almost all rock types were correctly classified except phyllite. The overall error and averaged class error in this validation were 14.2 % and 12.5 %, respectively.

Table 4.3 Confusion matrix of the random forest model

	Granite	Hornfels	Ore	Phyllite	Error (%)
Granite	3	0	0	0	0
Hornfels	0	1	0	0	0
Ore	0	0	1	0	0
Phyllite	0	1	0	1	50

Variable importance can also be measured in this algorithm. In Figure 4.17, the variable importance of the random forest algorithm is given. For each rock type, Poisson's ratio was the most important feature. Except for the Poisson ratio, the order of importance of the variables varied according to the rock type.

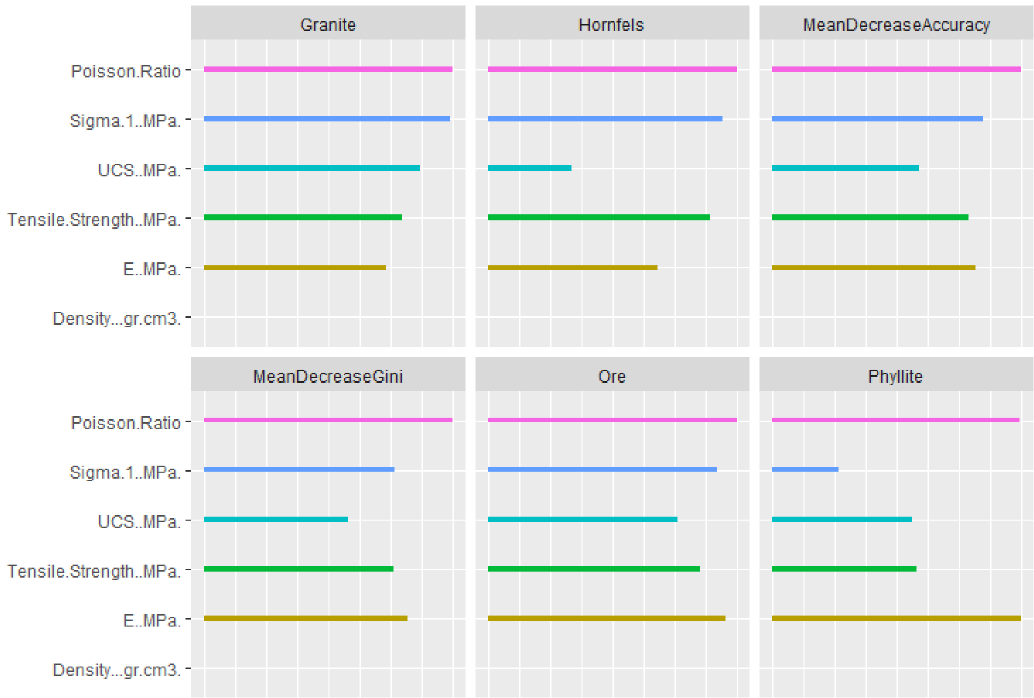


Figure 4.17 Variable importance of the random forest algorithm

When the two algorithms are compared, the random forest model outperforms the decision tree model, as it has a lower error rate. The reason for this lower error rate is that the random forest algorithm can create more complicated trees. By increasing the amount and variety of data, better results can be obtained from both classifiers.

CHAPTER 5

CONCLUSION

The experimental results of the Rock Mechanics Laboratory of the METU Department of Mining Engineering have been stored in a scattered manner in electronic tables without a standard in print or digital media. This storage method is a weak infrastructure for both data security and reliability. At the same time, it does not provide an infrastructure suitable for collective analysis by restricting data access for researchers. For this reason, the results of experiments carried out in the rock mechanics laboratory during the years 2000 to 2021, within the scope of this project, were transferred to test tables within a database after a series of data cleaning and processing methods. With the determined rules in the tables, different test results of the same test type were stored with the same standards. Tables are linked to one another with a unique test number key, allowing querying between all tables. In this way, it is possible to perform a collective query within all tables. For example, it is possible to view the test results of different types of experiments carried out under a single project code with a single query.

To provide data access for all researchers and enable collective analysis, the data in the database was transferred to the data warehouse through the SQL Server Data Tools package of the Visual Studio software. Data transfers take place via a connection between the database and the warehouse, and new information entries made to the database can also be updated in the data warehouse at certain times. OLAP cubes form the necessary infrastructures for the data analysis in the data warehouse and pre-determined calculations can be performed in these cubes, if necessary. In this way, the queries made during analysis are calculated and kept in the cubes' memory. This method not only provides an instant query opportunity for the researcher analyzing the data, but also ensures that the query results for any researcher connected to the cube are identical. At the same time, requiring permission to change the data in the data warehouse and OLAP cubes increases the reliability of the data by providing security.

It is possible to connect to OLAP cubes to the libraries of R, Python, C#, C++ programming languages, as well as with software such as Microsoft Excel and Microsoft PowerBI, which are used through the graphical user interface. In this project, Microsoft PowerBI software was used for data analysis because it allows the use of programming languages such as R and python, in addition to the ready-made tools it offers.

The results of 9,667 experiments with 284 different project codes were used in the analyzes made in the data warehouse through OLAP cubes. The experimental results in this data set vary in the amount of information shared about specific experiment details. The main reasons for this are company confidentiality requirements pertaining to sample information, missing information and the samples having pre-existing weaknesses such as natural fractures and cracks.

In the analyzes carried out, 3,870 of 9,667 experiments performed between 2000-2021 were uniaxial/triaxial compression tests. The reasons for these tests being so predominant could include the amount of information that can be determined through these tests, the ease of sample preparation for cylindrical core samples, and comparably shorter testing times as compared to other tests, such as direct shear. Another finding is that the highest values in the average UCS results belong to rhyodacite, andesite, and limestone rocks, while the lowest values include rocks such as lignite, claystone, and clay. The results show that it is possible to match the available rock types by their UCS results in a consistent way.

When the results of the slake durability tests are examined, they verify that the stability index 1 is always greater than the stability index 2 in the parameters ID 1 and ID 2, which are the parameters of the test results. This provides information to researchers about the reliability of the experimental results.

When the indirect tensile test results are analyzed according to rock types, similar to the uni/triaxial compression test results, the tensile strengths of granite, andesite, and limestone rocks are observed to be much lower than rocks such as claystone, clay and lignite.

Researchers can access the test results and sample information carried out between the years 2000-2021 through the data warehouse with this study. In addition to the test results, the data also includes information such as width, length, diameter, density, weight, volume, location, photos of rock samples before and after testing, rock class, and rock type. Researchers also have the opportunity to make instant queries through the OLAP cubes that store this information.

A case study was conducted in order to show the potential knowledge discovery capability of the data warehouse system. For this case study, a project data set containing 300 rock mechanics experiments were queried from the data warehouse. The reason for choosing this data set was that different experiments were carried out on different rock types in the project. After analyzing this data set in detail, decision tree and random forest algorithms were trained to classify the type of rock tested through the Rattle software. Validation results show that the random forest model outperformed the decision tree model with respective 57.1 % and 14.2 % error rates. This case study shows that valuable information can be extracted from the data warehouse system when reliable sources of data are integrated.

The missed opportunity in this study was that the data warehouse could not be integrated with the data sources from an operating mine. Integration with a fleet management system, drill monitoring system or fragmentation analysis would enable researchers to relate the test results to dig rate, drill rate, blasting efficiency, and others. This integration would provide the potential utilization of data mining techniques for knowledge discovery to improve the decision-making process in the field.

It is planned to keep the infrastructures created as a result of this study up to date by adding future experiments to the database. As future research, a graphical user interface will be developed as it is of key importance for better representation of the results and could enhance the utilization for users. In addition to these systems, an automatic reporting system can be set up via a template. Researchers can report the experiments carried out through the project code with the desired detail. Furthermore, a simultaneous outlier detection system during experiments in the rock mechanics laboratory could be created within this existing data warehouse system. With the prepared data warehouse and OLAP cubes, they can make the information ready for data mining after a series of manipulation and filtering processes. Lastly, before and after the experiment photographs of the rock samples could be analyzed with available image processing techniques. A potential relationship between the visual features and experiment results could be investigated in detail.

REFERENCES

Abraham, A. (2005). Artificial neural networks. *Handbook of measuring system design*.

Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Ba'abbad, I., Althubiti, T., Alharbi, A., Alfarsi, K. & Rasheed, S. (2021). A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications. *Journal of Data Analysis and Information Processing*, 9, 162-174.

Barnewold, L., & Lottermoser, B. G. (2020). Identification of digital technologies and digitalisation trends in the mining industry. *International Journal of Mining Science and Technology*, *30*(6), 747-757.

Brandt, S., & Brandt, S. (1998). *Data analysis*. Springer-Verlag.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Caron, E., & Daniels, H. (2013). Explanatory business analytics in olap. *International Journal of Business Intelligence Research (IJBIR)*, *4*(3), 67-82.

Chen, J. Y., Zhang, S., & Zuo, W. R. (2007). Digital coal mine integrated automation system based on controlnet. *Journal of China University of Mining and Technology*, *17*(2), 210-214.

Chuanyao, S., Junwu, Z., & Muxin, J. (2018). Research on genetic mineral processing engineering. *Non-Ferrous Metals: Mineral Processing Section*, *1*, 1-7.

Codd, E. F. (1993). Providing OLAP (on-line analytical processing) to user-analysts: An IT mandate. *http://www. arborsoft. com/papers/coddTOC. html*.

Delisio, A., Zhao, J., & Einstein, H. H. (2013). Analysis and prediction of TBM performance in blocky rock conditions at the Lötschberg Base Tunnel. *Tunnelling and Underground Space Technology*, *33*, 131-142.

Descamps, F., Tshibangu, J. P., da Silva, N., & Regnard, S. (2013, October). A database to manage rock mechanical properties and assess formation abrasiveness for drilling. In *ISRM International Symposium-EUROCK 2013*. OnePetro.

Elmasri, R., & Navathe, S. (2017). *Fundamentals of database systems* (Vol. 7). Pearson.

Exadaktylos, G., Liolios, P., & Barakos, G. (2007, July). Some new developments on the representation and standardization of rock mechanics data: from the laboratory to the full scale project. In *Rock mechanics data: representation and standardisation, 11th ISRM congress*.

Gardner, L. J., & Bosman, J. D. (2015, May). Determining the Mechanical Properties of Igneous Rocks in the Western Limb of South Africa's Bushveld Complex. In *13th ISRM International Congress of Rock Mechanics*. OnePetro.

Gering, S., Calderon-Arteaga, C., Gutierrez, L., & White, T. (2017). Drill and Blast Implementation Case Study at Multiple Sites Within Freeport McMoRan Inc. *Mining Engineering*, 69(5), 16-23.

Gong, Q., & Zhao, J. (2009). Development of a rock mass characteristics model for TBM penetration rate prediction. *International journal of Rock mechanics and mining sciences*, *46*(1), 8-18.

Han, J., Kamber, M., & Mining, D. (2006). Concepts and techniques. *Morgan Kaufmann*, *340*, 94104-3205.

Hassanpour, J., Rostami, J., & Zhao, J. (2011). A new hard rock TBM performance prediction model for project planning. *Tunnelling and Underground Space Technology*, *26*(5), 595-603.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and prediction*, Springer Series in Statistics, Springer-Verlag, New York.

Hopfield, J. J. (1988). Artificial neural networks. *IEEE Circuits and Devices Magazine*, *4*(5), 3-10.

Hsiung, S. M. S., Ghosh, A., Stirewalt, G. L., Henderson, B., Fenrich, C. A., & Chen, R. (1995). Development of Rock Mechanics Properties Database-Progress Report.

Jones, M. Q. W. (2015). Thermophysical properties of rocks from the Bushveld Complex. *Journal of the Southern African Institute of Mining and Metallurgy*, *115*(2), 153-160.

Kahraman, M. M., & Dessureault, S. (2018). Increasing adherence to mine plan through data integration and process change. *International Journal of Mining, Reclamation and Environment*, *32*(5), 341-354.

Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

Karpuz, C., & Hindistan, M. A. (2008). Kaya mekaniği: Ilkeleri, uygulamaları,TMMOB Maden Mühendisleri Odası Yayınları

Kim, E., & Hunt, R. (2017). A public website of rock mechanics database from Earth Mechanics Institute (EMI) at Colorado School of Mines (CSM). *Rock Mechanics and Rock Engineering*, *50*(12), 3245-3252.

Li, X., Wang, G., & Zhu, H. (2012). A data model for exchanging and sharing ISRM rock mechanics test data. *Electron J Geotech Eng (EJGE)*, *17*, 377-401.

Liolios, P., & Exadaktylos, G. (2011). A relational rock mechanics database scheme with a hierarchical structure. *Computers & geosciences*, *37*(8), 1192-1204.

Ma, Y., & Guo, G. (Eds.). (2014). *Support vector machines applications* (Vol. 649). New York, NY, USA:: Springer.

Maimon, O., & Rokach, L. (Eds.). (2005). Data mining and knowledge discovery handbook.

Maniatis, A., Vassiliadis, P., Skiadopoulos, S., Vassiliou, Y., Mavrogonatos, G., & Michalarias, I. (2005). A presentation model & non-traditional visualization for OLAP. *International Journal of Data Warehousing and Mining (IJDWM)*, *1*(1), 1-36.

Menschik, F., Thuro, K., & Käsling, H. (2015, May). Presentation of A Web-Based Database Design for Documentation and Analysis of Rock Properties. In *13th ISRM International Congress of Rock Mechanics*. OnePetro.

Ng, I. T., & Lau, C. H. (2015, May). Determination and Estimation of Uniaxial Compressive Strength for Macau Granite. In *13th ISRM International Congress of Rock Mechanics*. OnePetro.

Pere, V., Seery, J., de Graaf, P., & Mould, R. (2011, August). Development of an acQuire Based Geotechnical Data Management System. In *Proc. 8th Int. Min. Geol. Conf. Queenstown* (pp. 22-24).

Ralph, J. (1993). Mindat.org. Retrieved from https://www.mindat.org

RocScience (2020). RocProp Primary Database. https://www.rocscience.com/help/rocdata/rocprop/RocProp_Primary_Database.htm

Saltz, J., Hotz, N., Wild, D., & Stirling, K. (2018, January). Exploring project management methodologies used within data science teams. In *24th Americas Conference on Information Systems 2018: Digital Disruption, AMCIS 2018*. Association for Information Systems.

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217-222.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (1997). *Database system concepts* (Vol. 4). New York: McGraw-Hill.

Tóth, Á., Gong, Q., & Zhao, J. (2013). Case studies of TBM tunneling performance in rock–soil interface mixed ground. *Tunnelling and Underground Space Technology, 38*, 140-150.

Tsois, A., Karayannidis, N., & Sellis, T. K. (2001, June). MAC: Conceptual data modeling for OLAP. In DMDW (Vol. 39, p. 5).

Williams, G. (2011). Boosting. In *Data Mining with Rattle and R* (pp. 269-291). Springer, New York, NY.

Yagiz, S. (2006). TBM performance prediction based on rock properties. *Proceedings of Multiphysics Coupling and Long Term Behavior in Rock Mechanics, EUROCK, 6*, 663-670.

Yagiz, S. (2008). Utilizing rock mass properties for predicting TBM performance in hard rock condition. *Tunnelling and Underground Space Technology*, *23*(3), 326-339.

Zheng, M. G., & Cai, S. J. (2008). An intelligent system for calculating the scale of rational, enlarged production of an underground non-ferrous metal mine. *Journal of China University of Mining and Technology*, *18*(2), 214-219.

Zohuri, B., & Moghaddam, M. (2017). *Business Resilience System (BRS): Driven Through Boolean, Fuzzy Logics and Cloud Computation*. Springer International Publishing.

APPENDICES

A. Tables in RMDB

Brazilian Test

Column Name	Data Type	Allow Nulls
Unique_ID	nvarchar(50)	☐
ExpID	nvarchar(50)	☑
Rock_Type	nvarchar(50)	☑
Hole_No	nvarchar(50)	☑
From_m	float	☑
To_m	float	☑
Length_mm	float	☑
Diameter_mm	float	☑
Weight_gr	float	☑
Failure_Load_kgf	float	☑
Failure_Load_kN	float	☑
Indirect_Tensile_Strength_M...	float	☑
Comments	nvarchar(200)	☑
		☐

Figure A. 1 Brazilian test table in the RMDB

Deformability Test

	Column Name	Data Type	Allow Nulls
🔑	Unique_ID	nvarchar(50)	☐
	ExpID	nvarchar(50)	☑
	Rock_Type	nvarchar(50)	☑
	Test_Type	nvarchar(50)	☑
	Hole_No	nvarchar(50)	☑
	From_m	float	☑
	To_m	float	☑
	Length_mm	float	☑
	Diameter_mm	float	☑
	Weight_gr	float	☑
	Failure_Load_kgf	float	☑
	Failure_Load_kN	float	☑
	Compressive_Strength_kg_c...	float	☑
	Compressive_Strength_Mpa	float	☑
	Youngs_Modulus_MPa	float	☑
	Poissons_Ratio	float	☑
	Internal_Friction_Angle	float	☑
	Cohesion_MPa	float	☑
	Failure_Plane_Angle	float	☑
	Comments	nvarchar(15...	☑
			☐

Figure A. 2 Static deformability test table in the RMDB

Density Porosity Test

Column Name	Data Type	Allow Nulls
Unique_ID	nvarchar(50)	☐
ExpID	nvarchar(50)	☑
Rock_Type	nvarchar(50)	☑
Rock_Type_Comments	nvarchar(10...	☑
Test_Type	nvarchar(50)	☑
HoleNo	nvarchar(50)	☑
SampleID	nvarchar(50)	☑
SampleNo	nvarchar(50)	☑
Depth_m	nvarchar(50)	☑
From_m	float	☑
To_m	float	☑
Length_mm	float	☑
Diameter_mm	float	☑
Volume_cm3	float	☑
Natural_Weight_gr	float	☑
Dry_Weight_gr	float	☑
Wet_Weight_gr	float	☑
Submerged_Weight_gr	float	☑
Unit_Weight_gr_cm3	float	☑
Dry_Unit_Weight_gr_cm3	float	☑
Wet_Unit_Weight_gr_cm3	float	☑
Natural_Unit_Weight_gr_cm3	float	☑
Specific_Gravity_gr_cm3	float	☑
Porosity	float	☑
Unit_Volume_Weight_kN_m3	float	☑
Dry_Unit_Weight_kN_m3	float	☑
Saturated_Unit_Weight_kN_...	float	☑
Water_Absorption_by_Weight	float	☑
Water_Absorption_by_Volu...	float	☑
NaturalMoisture	float	☑
Comments	nvarchar(15...	☑
		☐

Figure A. 3 Density porosity test table in the RMDB

Direct Shear Test

	Column Name	Data Type	Allow Nulls
🔑	Unique_ID	nvarchar(50)	☐
	ExpID	nvarchar(50)	☑
	Rock_Type	nvarchar(50)	☑
	Drillhole_ID	nvarchar(50)	☑
	Set_No	nvarchar(50)	☑
	From_m	float	☑
	To_m	float	☑
	Weight_gr	float	☑
	Moisture_Content	float	☑
	Normal_Stress_MPa	float	☑
	Peak_Shear_Stress_MPa	float	☑
	Residual_Shear_Stress_M...	float	☑
	Shear_Surface	nvarchar(50)	☑
	Humidity_Status	nvarchar(50)	☑
	Shear_Envelope	nvarchar(50)	☑
	Peak_Cohesion_kPa	float	☑
	Peak_Friction_Angle	float	☑
	Residual_Cohesion_kPa	float	☑
	Residual_Friction_Angle	float	☑
	Comments	nvarchar(15...	☑
			☐

Figure A. 4 Direct shear test table in the RMDB

Point Load Test

	Column Name	Data Type	Allow Nulls
🔑	Unique_ID	nvarchar(50)	☐
	ExpID	nvarchar(50)	☑
	Rock_Type	nvarchar(50)	☑
	Drillhole_ID	nvarchar(50)	☑
	From_m	float	☑
	To_m	float	☑
	Length	float	☑
	Weight_gr	float	☑
	Loading_Direction	nvarchar(50)	☑
	Diameter_W_mm	float	☑
	D	float	☑
	Distance_btw_Platens_D_mm	float	☑
	Failure_Load_P_kN	float	☑
	Equivalent_Diameter_Square_De2_m...	float	☑
	Equivalent_Diameter_mm	float	☑
	Failure_Load_kgf	float	☑
	Point_Load_Index_Is_MPa	float	☑
	Correction_Factor_F	float	☑
	Corrected_Point_Load_Index_Is_50	float	☑
	Coefficient_k	float	☑
	Uniaxial_Compressive_Str_sc_MPa	float	☑
	Comments	nvarchar(15...	☑
			☐

Figure A. 5 Point load test table in the RMDB

Slake Durability

Column Name	Data Type	Allow Nulls
Unique_ID	nvarchar(50)	☐
ExpID	nvarchar(50)	☑
Rock_Type	nvarchar(50)	☑
HoleNo	nvarchar(50)	☑
From_m	float	☑
To_m	float	☑
Length	float	☑
Weight_gr	float	☑
Id_1	float	☑
Id_2	float	☑
Id_3	float	☑
Id_4	float	☑
Id_5	float	☑
Slake_Durability_Class	nvarchar(50)	☑
SlakingFluidTemp	float	☑
SlakingFluidNature	nvarchar(50)	☑
MassofDrum_Sample	float	☑
MassofDrum_RetainedSamp...	float	☑
MassofDrum_RetainedSamp...	float	☑
MassofDrum	float	☑
MassofDrum2	float	☑
Comments	nvarchar(15...	☑
		☐

Figure A. 6 Slake durability test table in the RMDB

Uniaxial/Triaxial Compression Test

	Column Name	Data Type	Allow Nulls
🔑	Unique_ID	nvarchar(50)	☐
	ExpID	nvarchar(50)	☑
	Rock_Type	nvarchar(50)	☑
	Test_Type	nvarchar(50)	☑
	Hole_No	nvarchar(50)	☑
	From_m	float	☑
	To_m	float	☑
	Length_mm	float	☑
	Diameter_mm	float	☑
	Weight_gr	float	☑
	Failure_Load_kgf	float	☑
	Failure_Load_kN	float	☑
	Lateral_Pressure_kg_cm2	float	☑
	Lateral_Pressure_MPa	float	☑
	Compressive_Strength_kg_c...	float	☑
	Compressive_Strength_MPa	float	☑
	Youngs_Modulus_MPa	float	☑
	Poissons_Ratio	float	☑
	Set_Numbers	float	☑
	Triaxial_Set	float	☑
	Triaxial_Failure_Envelope	nvarchar(50)	☑
	Internal_Friction_Angle	float	☑
	Cohesion_MPa	float	☑
	b	float	☑
	m	float	☑
	Comments	nvarchar(15...	☑
			☐

Figure A. 7 Uniaxial/Triaxial compression test table in the RMDB

ID Table

	Column Name	Data Type	Allow Nulls
🔑	Unique_ID	nvarchar(50)	☐
	ExpID	nvarchar(50)	☐
			☐

Figure A. 8 ID table in the RMDB

RockTypes_Simplified

	Column Name	Data Type	Allow Nulls
🔑	[Rock Type]	nvarchar(...	☐
	[Rock Type Detailed I...	nvarchar(...	☐
	Class	nvarchar(...	☐
			☐

Figure A. 9 Rock types table in the RMDB

Experiment Info

	Column Name	Data Type	Allow Nulls
🔑	ExpID	nvarchar(50)	☐
	Company	nvarchar(50)	☑
	cid	nchar(10)	☑
	Date	date	☑
	City	nvarchar(50)	☑
	District	nvarchar(50)	☑
			☐

Figure A. 10 Experiment information table in the RMDB