

THESIS

**IMPLEMENTATION OF DATA MINING IN COAL PRODUCTION
CONTROL WITH DASHBOARD INTEGRATION
AS DECISION SUPPORT SYSTEM**

Compiled and submitted by

MUH. SANDI ARISTA IKHSAN YAHMID

D111181330



**MINING ENGINEERING STUDY PROGRAM
FACULTY OF ENGINEERING
HASANUDDIN UNIVERSITY
MAKASSAR
2023**

LEGALIZATION

IMPLEMENTATION OF DATA MINING IN COAL PRODUCTION CONTROL WITH DASHBOARD INTEGRATION AS DECISION SUPPORT SYSTEM

Submitted by

MUH. SANDI ARISTA IKHSAN YAHMID
D111181330

Has been defended in front of the Examination Committee which established for the Completion of Mining Engineering Undergraduate Program of Faculty of Engineering Universitas Hasanuddin on January 26, 2023 and declared eligible

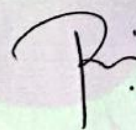
Approved by,

Supervisor,

Co-Supervisor,



Dr. Aryanti Virtanti Anas, S.T., M.T.
NIP.197010052008012026



Dr-Eng. Rini Novrianti Sutarjo Tui, S.T., M.BA., M.T.
NIP.198311142014042001

Ad Interim Head of Study Program,



Dr. Amil Ahmad Ilham, S.T., M.IT.
NIP.197310101998021001

LEMBAR PENGESAHAN SKRIPSI

IMPLEMENTATION OF DATA MINING IN COAL PRODUCTION CONTROL WITH DASHBOARD INTEGRATION AS DECISION SUPPORT SYSTEM

Disusun dan diajukan oleh

MUH. SANDI ARISTA IKHSAN YAHMID

D111181330

Telah dipertahankan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Studi Program Sarjana Program Studi Teknik Pertambangan Fakultas Teknik Universitas Hasanuddin pada tanggal 26 Januari 2023 dan dinyatakan telah memenuhi syarat kelulusan.

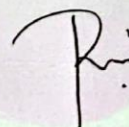
Menyetujui,

Pembimbing Utama,

Pembimbing Pendamping,

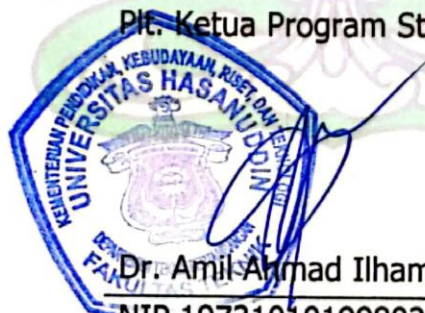


Dr. Aryanti Virtanti Anas, S.T., M.T.
NIP.197010052008012026



Dr-Eng. Rini Novrianti Sutarjo Tui, S.T., M.BA., M.T.
NIP.198311142014042001

Pt. Ketua Program Studi,



Dr. Amil Ahmad Ilham, S.T., M.IT.
NIP.197310101998021001

STATEMENT OF AUTHENTICITY

The undersigned below:

Name : Muh. Sandi Arista Ikhsan Yahmid
Student ID : D111181330
Study Program : Mining Engineering
Degree : Undergraduate (S1)

Declare that my bachelor thesis titled

Implementation of Data Mining in Coal Production Control With Dashboard
Integration as Decision Support System

is my own writing and not a claim of others. The thesis which I wrote is really my own work. If in the future, it is proven or can be proven that a part or a whole of this thesis is the work of others, I am willing to accept any penalty for such act.

Makassar, January 26, 2023



Muh. Sandi Arista Ikhsan Yahmid

ABSTRACT

Mining companies generally have a business model that utilizes contractors for the mining operations so the mine owners are responsible for supervising and controlling mining operations by contractors which are expected to run effectively and efficiently with the help of available technology. The objective of this study is to create a real-time dashboard as a decision support system based on the production achievement of mining contractors after being analysed using data mining techniques with K-Means clustering model. Data processing method used refers to the Cross Industry Standard Process for Data Mining (CRISP-DM) with Python programming language and the dashboard created using Tableau Desktop. There are 4 clusters formed based on the data used. The evaluation of the model was carried out with a silhouette score of 0.425, which means that the distribution of the data has not been clearly grouped. A real-time production control dashboard created as the last stage of CRISP-DM as well as a decision support system for mine owner companies. The information provided includes production achievement, achievement of production parameters, time use management and recommendations for interventions based on contractor clusters.

Keywords: data mining, machine learning, coal production, visualization dashboard, contractor.

ABSTRAK

Perusahaan pertambangan pada umumnya memiliki model bisnis yang bekerja sama dengan kontraktor untuk operasi penambangan sehingga pemilik tambang bertanggung jawab untuk mengawasi dan mengendalikan operasi penambangan oleh kontraktor yang diharapkan dapat berjalan secara efektif dan efisien dengan bantuan teknologi yang tersedia. Penelitian ini bertujuan untuk membuat dashboard real-time sebagai sistem pendukung pengambilan keputusan berdasarkan pencapaian produksi kontraktor tambang setelah dianalisis menggunakan teknik data mining dengan model klasterisasi K-Means. Metode pengolahan data yang digunakan mengacu pada Cross Industry Standard Process for Data Mining (CRISP-DM) dengan bahasa pemrograman Python dan dashboard dibuat menggunakan Tableau Desktop. Ada 4 cluster yang terbentuk berdasarkan data yang digunakan. Evaluasi model dilakukan dengan silhouette score sebesar 0,425 yang berarti sebaran data belum terkelompokkan dengan jelas. Dashboard pengontrolan produksi dibuat sebagai tahap terakhir CRISP-DM sekaligus sebagai sistem pendukung keputusan untuk perusahaan tambang. Informasi yang diberikan meliputi pencapaian produksi, pencapaian parameter produksi, manajemen penggunaan waktu dan rekomendasi intervensi berdasarkan klaster kontraktor.

Keywords: penambangan data, pembelajaran mesin, produksi batubara, dasbor visualisasi, kontraktor

PREFACE

Alhamdulillah, alhamdu lillahi rabbil 'alamin. First of all, the author's deepest thank To Allah SWT, the lord of the universe because of His grace that the writer can complete this thesis entitled "Implementation of Data Mining in Coal Production Control with Dashboard Integration as Decision Support System". This thesis was carried out as a final project and one of the requirements in completing the undergraduate education level at the Department of Mining Engineering, Faculty of Engineering, Hasanuddin University.

The author realizes that this thesis could not be separated from the help of various related parties. The author is grateful to the Ministry of Education and Culture (KEMDIKBUD) who has organized the "Magang Merdeka" internship program so that the author can study and conduct research with the intermediary of the program. The authors would like to thank to the mining company that cannot be mentioned and the great people of the company who have helped in the internship and data collection for this thesis. Thanks also to the second batch of interns who always supported and motivated during the internship location.

This thesis would not be successful without the advice and assistance of those who always guide the author during the study at the Mining Engineering Study Program, Faculty of Engineering, Hasanuddin University. Therefore, the author would like to express many thanks to Dr. Aryanti Virtanti Anas, S.T., M.T. as the Supervisor, Dr. Eng. Rini Novrianti Sutardjo Tui, S.T., M.BA., M.T. as the Co-Supervisor, Dr. Eng. Purwanto, S.T., M.T. and Dr. Eng. Ir Muhammad Ramli, M.T. as the examiner at the thesis defence. Thanks also to Rizki Amalia, S.T., M.T. as a supervisor of the Mine Planning and Valuation Laboratory as well as to all the lecturers who have educated the author from the first semester until now. The author also thanks the academic staff of the Mining Engineering

Department who were very helpful in the management towards the end of the author's study.

The author also thanks friends who have always supported and acted funny while studying for the last 4 years. Of course, the author also expresses his deepest gratitude to friends from TUNNEL 2018, PERMATA FT-UH, MENTOR 2018 FT-UH, KOMTEK 09 SMFT-UH, members of the Mine Planning and Valuation Laboratory, as well as friends from various study programs and batches at the Faculty of Engineering, Hasanuddin University and other universities who became witness to the author's journey in lectures.

It is not easy for the writer to finish all the contents of this thesis without support from Mr. Ikhsan Yahmid Yusuf and Mrs. Hasdiana as the author's parent who has given their great love and struggle to provide college opportunities for the author, as well as brother, sister, and extended family, who always provide prayers, encouragement, and good moral and material support. Hopefully this thesis can be useful for readers and all parties. Aamiin.

Makassar, January 26, 2023

Author

TABLE OF CONTENTS

	Page
ABSTRACT	v
<i>ABSTRAK</i>	vi
PREFACE	vii
TABLE OF CONTENTS	ix
LIST OF FIGURES	xi
LIST OF TABLES	xiv
LIST OF APPENDICES	xv
CHAPTER I INTRODUCTION	1
1.1 Background	1
1.2 Research Problem	2
1.3 Research Objectives	3
1.4 Research Advantages	3
1.5 Research Stages	3
CHAPTER II MINING PRODUCTION AND DATA MINING	5
2.1 Digital Transformation in Mining Industry	5
2.2 Big Data Management in Mining Industry	8
2.3 Data Mining	11
2.4 The Stages of Data Mining	13
2.5 Business Process of Mining Industry	15
2.6 Machine Learning	19
2.7 Relationship of Data Mining and Machine Learning	22
2.8 Clustering in Machine Learning with K-Means Algorithm	22
2.9 Determination of the Optimal Number of Cluster	25

	Page
2.10 Python Programming Language.....	26
2.11 Data Visualization using Tableau	28
2.12 Mining Production.....	28
CHAPTER III RESEARCH METHODS	31
3.1 Data Collection	31
3.2 Data Mining.....	33
3.3 System Design	42
3.4 System Validation.....	53
CHAPTER IV IMPLEMENTATION OF DATA MINING IN COAL PRODUCTION	55
4.1 Data Preparation	55
4.2 Modelling K-Means Clustering.....	61
4.3 Interpretation and Evaluation	63
4.4 Dashboard Design	71
4.5 Dashboard Validation.....	78
4.6 Prototype Design of Data Entry Form.....	80
CHAPTER V CONCLUSION AND SUGGESTION	82
5.1 Conclusion	82
5.2 Suggestion	83
REFERENCES.....	84

LIST OF FIGURES

Figure	Page
2.1 Digital transformation (Soofastaei, 2021).....	5
2.2 Relative digitalization by industries (Ghandi et al., 2016)	6
2.3 Frequency of digital transformation related terms in annual reports of top mining global companies (Young and Rogers, 2019).	7
2.4 Application procedure of big data (Qi, 2020).....	8
2.5 The major processes of big data (Qi, 2020)	9
2.6 The major data sources in the mining industry (Qi, 2020).....	10
2.7 The stages of data mining (Provost and Fawcett, 2013)	13
2.8 The life cycle of mine project (Revuelta, 2018)	16
2.9 Data science, data mining and machine learning (Kurniawan, 2020)	22
3.1 Import library and dataset.....	33
3.2 Data cleaning (a)	34
3.3 Data cleaning (b)	35
3.4 Data transformation (a).....	36
3.5 Data transformation (b).....	36
3.6 Syntaxes to finding optimal K.....	37
3.7 Syntaxes to training K-Means model with optimal value of K.....	38
3.8 Syntaxes to visualize clustered data	39
3.9 Syntaxes to gain more information of each cluster	39
3.10 Menu to manage analytics extension connection in Tableau.....	40
3.11 Connecting Tabpy server to Tableau	40
3.12 Deploying model into the Tableau by Jupyter Notebook	41
3.13 Using the model deployed into the Tableau	41

3.14	Design of production control dashboard as decision support system.....	42
3.15	Big number KPI of daily production	44
3.16	Horizontal bar chart of daily production	44
3.17	Big number KPI of WTD production.....	45
3.18	Horizontal bar chart of WTD production.....	45
3.19	Big number KPI of MTD production	46
3.20	Horizontal bar chart of MTD production	46
3.21	Big number KPI of YTD production.....	47
3.22	Horizontal bar chart of YTD production.....	47
3.23	Line chart of production in last 14 days	48
3.24	Line chart of weekly production	48
3.25	Line chart of monthly production.....	49
3.26	Horizontal bar chart of time usage	49
3.27	Donut chart of operational issue based on PA and UA	50
3.28	Vertical bar chart of operational issue based on productivity	50
3.29	Horizontal bar chart of PA parameter by equipment	51
3.30	Horizontal bar chart of UA parameter by equipment.....	51
3.31	Horizontal bar chart of productivity parameter by equipment.....	52
3.32	Symbolic shape of clustering result	52
3.33	Research flowchart.....	54
4.1	Missing value checking result.....	55
4.2	Outlier in dataset	56
4.3	Boxplot visualization of production parameters	56
4.4	Descriptive statistic after cleaning outliers	57
4.5	Boxplot visualization of production parameters after cleaning outliers.....	57
4.6	Multiplication of the PA, UA and Pdty parameters with quantity.....	58

4.7	Group data by contractor.....	59
4.8	Create new columns containing the average parameter of each contractor	59
4.9	Create new column containing the deviation of parameter	60
4.10	Handling missing value of the new data	60
4.11	Finding elbow with SSE value.....	61
4.12	Silhouette score of each cluster	62
4.13	Result of clustering in 3D plot.....	62
4.14	Number of samples by contractor of Cluster 0.....	64
4.15	Number of samples by contractor of Cluster 1.....	65
4.16	Number of samples by contractor of Cluster 2.....	66
4.17	Number of samples by contractor of Cluster 3.....	68
4.18	Data connection in Tableau	71
4.19	Production control dashboard	75
4.20	Production control dashboard (tool tip)	76
4.21	Production control dashboard (highlighted and filtered).....	77
4.22	Prototype design of daily production form	80
4.23	Prototype design of daily time usage form	81

LIST OF TABLES

Table	Page
2.1 Subjective interpretation of silhouette score (Kaufman and Rousseeuw, 1990)	.26
3.1 Attribute description of data parameter	32
3.2 Attribute description of data TUM.....	32
4.1 Number of samples in each cluster.....	63
4.2 Centroid of each cluster.....	63
4.3 Characteristic of Cluster 0.....	64
4.4 Characteristic of Cluster 1.....	65
4.5 Characteristic of Cluster 2.....	66
4.6 Characteristic of Cluster 3.....	67
4.7 Interview questions and answers from subject matter expert of the company...	68
4.8 Intervention recommendation for each cluster	70
4.9 Daily OB production by contractors	78
4.10 Week-to-date OB production.....	78
4.11 Month-to-date OB production	79
4.12 Year-to-date OB production	79

LIST OF APPENDICES

Appendix	Page
A Calculated Fields in Tableau.....	88
B Python Code for Data Mining	101
C Python Code for Model Deployment	108
D Sample Dataset Parameter	110
E Sample Dataset TUM (Time Usage Management).....	112

CHAPTER I

INTRODUCTION

1.1 Background

The mining industry in recent decades had to face challenges in its operations by increasing productivity to cope with natural factors such as decreasing ore grades, deeper deposits and harder rock masses as well as increasing environmental and social awareness, which has prompted the industry to continuously strive to improve its processes throughout the value chain as a whole. Innovation plays an important role with the right solutions to overcome these difficulties, define mining activities (Sanchez and Hartlieb, 2020). One of the mining commodities is coal. Compared to other energy sources such as oil, gas and renewable energy, the operation of coal energy sources is required to be much more efficient. The development of renewable energy for the next few years makes the competition between these energy sources even tighter, especially in terms of operational efficiency (BP Energy, 2020; Jezard, 2017).

The key to make mining more efficient is to select the right technology and to make the best use of available data. Mining companies nowadays should improve data access and relevance by focusing on getting real, applicable insights from data and sharing them clearly and effectively with the right levels of the organization (World Economic Forum, 2017). The application of digital technology makes the amount of data generated increase drastically because sensors from IoT devices or digital applications record every data in real time. Data management becomes a crucial aspect when all objects are connected and constantly exchange information (Chaulya and Prasad, 2016). Massive data will not be useful unless we can explore the information contained in it. This is where the need for data science to help get the benefits of data. Data science in

practice intersects with data mining and machine learning techniques to perform data analysis accompanied by in-depth knowledge of the industry they are involved in or also called domain expertise (Kurniawan, 2020).

One of the processes that produce the most data is in the core process of the coal mining business that is the mining operation stage which is divided into overburden removal and coal getting. Rupprecht (2015) said that mining companies generally have a business model that utilizes contractors for the entire mining cycle especially in the mining operations. Both parties must trust and understand each other's business process with the aim of encouraging maximum contribution. Contractors need to understand the mine owner's expectations, requirements, and quality constraints in order to deliver optimal results. Mine owners also need to understand the realities of mining, production, stripping considerations, and other operational issues. Mine owners are responsible for supervising and controlling mining operations by contractors which are expected to run effectively and efficiently with the help of available technology.

Therefore, this study was conducted by applying methods in data science, namely data mining techniques with a machine learning model approach to extract information from coal production data that held by contractors to support mine owner's decision-making processes in operational activities. The information obtained is expected to help mine owner to make decisions more effective and efficient.

1.2 Research Problem

Mine owner needs to monitor contractor's production performance, evaluate if there is a problem in production and make appropriate interventions according to the problem. So, the research problem is how the production control team of mine owner can quickly find out which contractors are not achieving their production targets. After finding out which contractors are not achieving their production targets, how can the

production control team quickly identify the cause of the contractors not achieving their production targets so that interventions can be carried out according to the problem.

1.3 Research Objectives

The objective of this study is to create a real-time dashboard as a decision support system based on the production achievement of mining contractors along with their production parameters after being analysed using data mining techniques with clustering model. The clustering model can provide recommendations for appropriate actions according to the characteristics of contractor's issues.

1.4 Research Advantages

This study is useful as a reference for the use of data processing technology to maximize the potential of technology application in mining companies. A data-driven decision making can help mining companies to make strategic decisions to solve problems more effective and efficient. Dashboard prototype as output in this research can be used and developed by the mining companies to monitor and control production performance of mining contractors.

1.5 Research Stages

There are several stages performed in this study start with topic determination to thesis seminar. Data were collected by directly observation and given by the company. Stages have carried out in this study are as follows:

a. Topic determination

This stage is the initial process to determine the direction of this study. Predetermined topic becomes guideline in the discussion and problem solving.

b. Literature study

Literature study is the stage of collecting various references according to the research topic. References can be obtained from national and international journals, as well as company data and/ or support data that is valid and accountable.

c. Problem identification

This stage aims to identify issues that will be discussed in the research. The problem has a direct relation to the research topic.

d. Data collection

Data collection was done to obtain the data needed in the discussion of problems in research. The data is obtained from interviews (primary data) and data from company (secondary data).

e. Data processing

Data Processing was done using several tools and applications e.g. Python, Jupyter Notebook, Microsoft Excel, MySQL Workbench and Tableau Public. Data processing method used is data mining refers to the Cross Industry Standard Process for Data Mining (CRISP-DM).

f. Thesis composing

Thesis composing is carried out in accordance with the applicable format in the Manual Book of The Final Assignment of The Mining Engineering Department Student of Hasanuddin University.

g. Thesis Seminar

Thesis seminar will be done to present the result and discussion of the thesis.

CHAPTER II

MINING PRODUCTION AND DATA MINING

2.1 Digital Transformation in Mining Industry

The fourth industrial revolution was happening when the world is facing with the digital decade. A massive amount of data in mining industry is collected from many equipment and machines working in the mining sites that are much more than ever before (Pacheco, 2019). These data can potentially make great opportunities for mining innovation to find new solutions for business problems through digital transformation in this industry (Soofastaei, 2019). Technology driven process consists of three main components of digital transformation which is data, connectivity, and decision-making as shown in Figure 2.1 (Soofastaei, 2021).

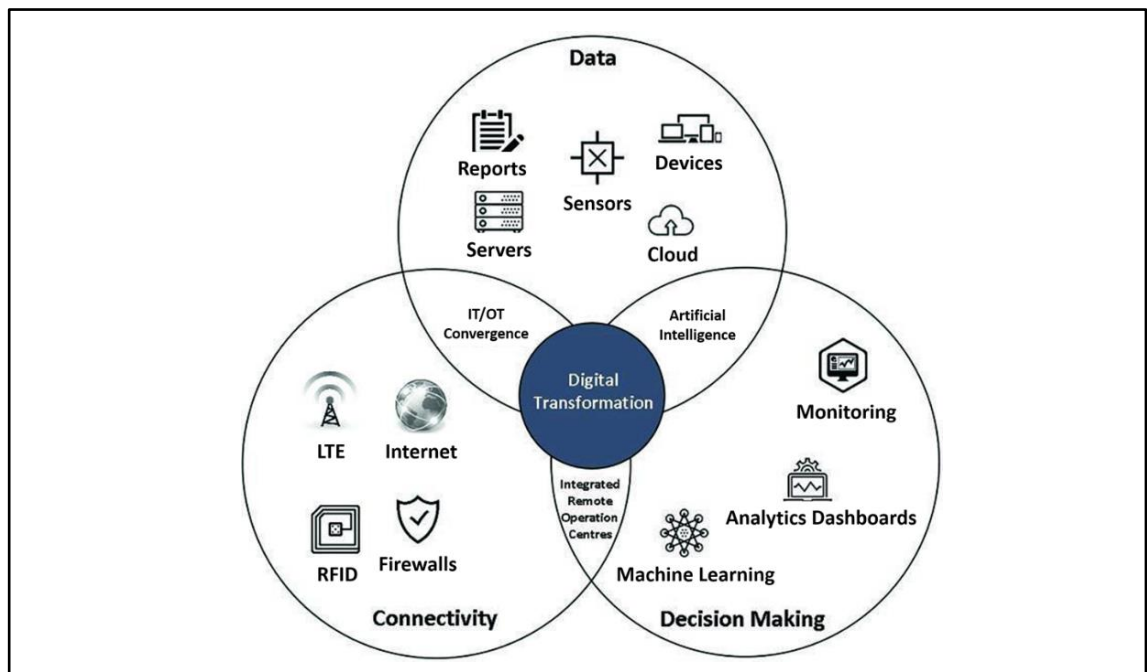


Figure 2.1 Digital transformation (Soofastaei, 2021)

The mining companies in the past could choose to be later or early adopters of the new technologies. However, this is no longer the reality nowadays, considering the

journey on digital transformation becomes an essential plan for all companies working in the mining industry. Overall, there are four highlighted summaries according to Soofastaei (2021) for digital transformation in the mining industry as follows:

- a. Mining companies should start digital transformation program as an essential revolution in this industry.
- b. There are three foundational components of the digital mining transformation process. These components are information, intranet and internet connectivity, and decision support.
- c. Digital transformation delivers a conversation on how it will be an essential part of the achievements of mining businesses into the future era.
- d. Digital transformation recognizes the strategic fields in which organizations of higher learning can supply the required resources to support the mining industry.

The mining industry needs to critically use digital transformation to increase the safety, productivity, and efficiency. However, mining industry is behind most other industries based on Ghandi et al. (2016) as shown in Figure 2.2. Mining stands out as one of the least digitized industry categories.

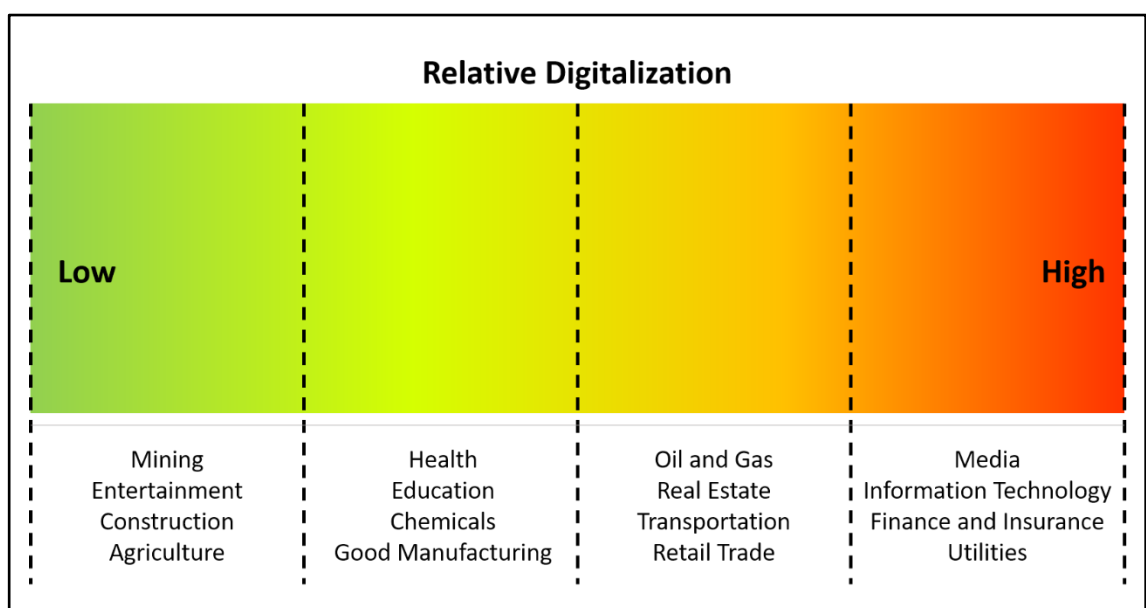


Figure 2.2 Relative digitalization by industries (Ghandi et al., 2016)

A completed review of the yearly published documents from the industries top ten mining companies shows that six out of ten stated that digital transformation is a part of its policy, three out of ten corporations list qualitative consequences from digitalization programs, and only one out of ten might provide quantitative value for the benefits of digital transformation (Sganzerla et al., 2016). However, Soofastaei (2021) said that this story has been changed, and currently, there are many other completed successful digital transformation plans in big mining companies globally.

Figure 1.3 demonstrates the number of companies which mentioned specific terms and the total number of mentions of each cited terms related to the digital transformation in some published annual reports by top mining global companies conducted by Young and Rogers (2019). The terms technology and data exist for the most part in the reviewed documents. Innovation were cited very regularly and ordinarily close related to the term technology, which indicates that technology and innovation are the foremost popular concern through top mining businesses.

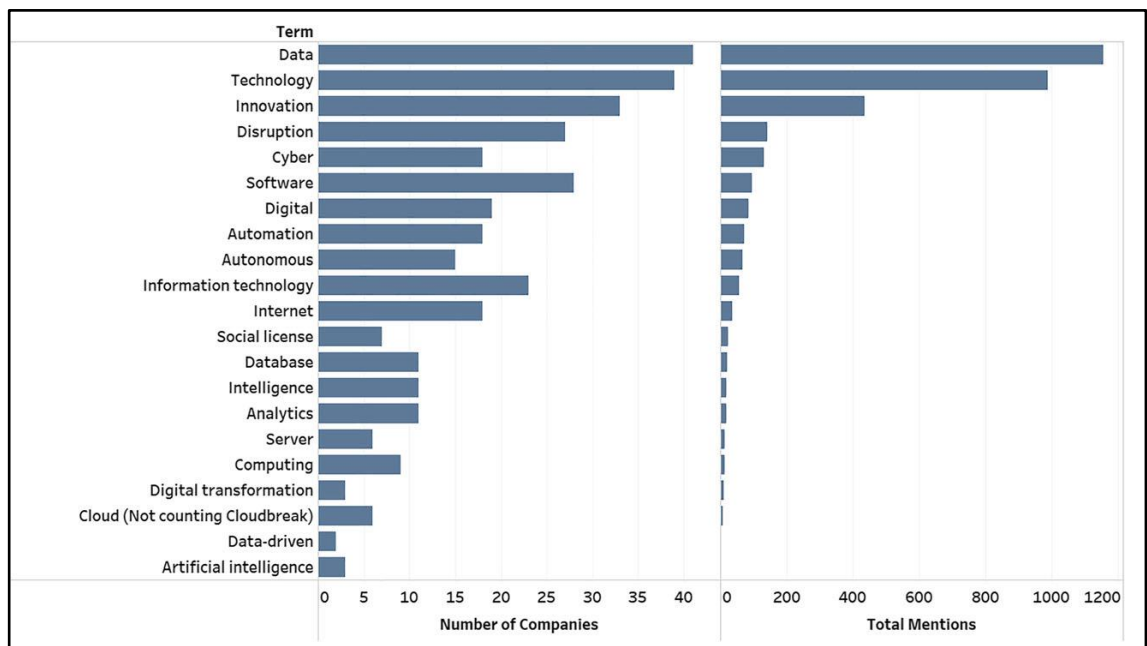


Figure 2.3 Frequency of digital transformation related terms in annual reports of top mining global companies (Young and Rogers, 2019).

2.2 Big Data Management in Mining Industry

Driving deeper into the characteristics of big data, different authors have proposed different criteria for the term. Ward and Barker (2013) identified three important characteristics of big data to differentiate it from other data we commonly encounter, namely size, complexity, and technology. Another way to identify big data is by using the attributes of the 5Vs (Bilal et al., 2016; Kaplinski et al., 2016), which is the abbreviation for volume (large amount of data in petabytes or terabytes), variety (different types of structured, semi-structured, or unstructured data), velocity (continuous streams and analysis of the data in near real-time or virtually real-time), veracity (integrity of data), and value (valuable information behind the data).

The application of big data often consists of three steps: source data collection, data structuring, and knowledge discovery. Figure 2.4 illustrates the application procedure of big data. Data structuring is not necessary if structural data are directly collected from mining operations. Various platforms have been developed to facilitate the application of big data, such as horizontal and vertical scaling platforms (Qi, 2020).

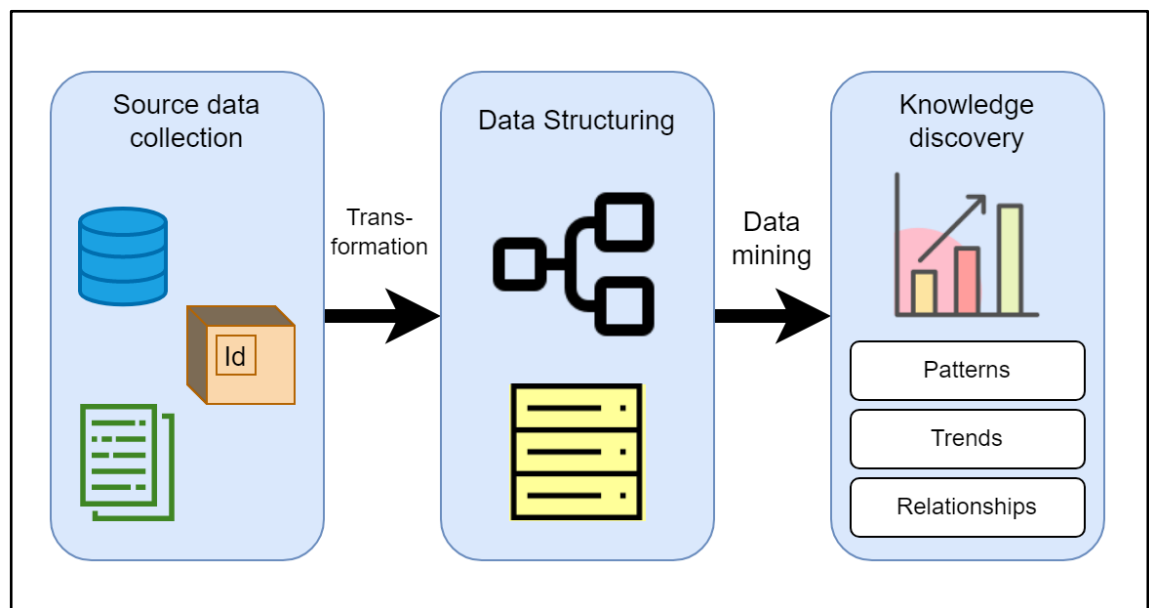


Figure 2.4 Application procedure of big data (Qi, 2020)

Big data processes can be classified into two main classes, namely, BDM and BDA. BDM primarily deals with data collection, pre-processing, storage, and sharing, which are essential for the implementation of BDA (Provost and Fawcett, 2013). The objective of BDA is mainly interpretation and knowledge discovery. Figure 2.5 illustrates the major processes of big data. BDM and BDA focus on different aspects of big data. Moreover, BDM is the prerequisite of BDA, and the successful application of big data in the mining industry relies heavily on BDM. Thus, this study discusses the basics of BDM in the mining industry (Qi, 2020).

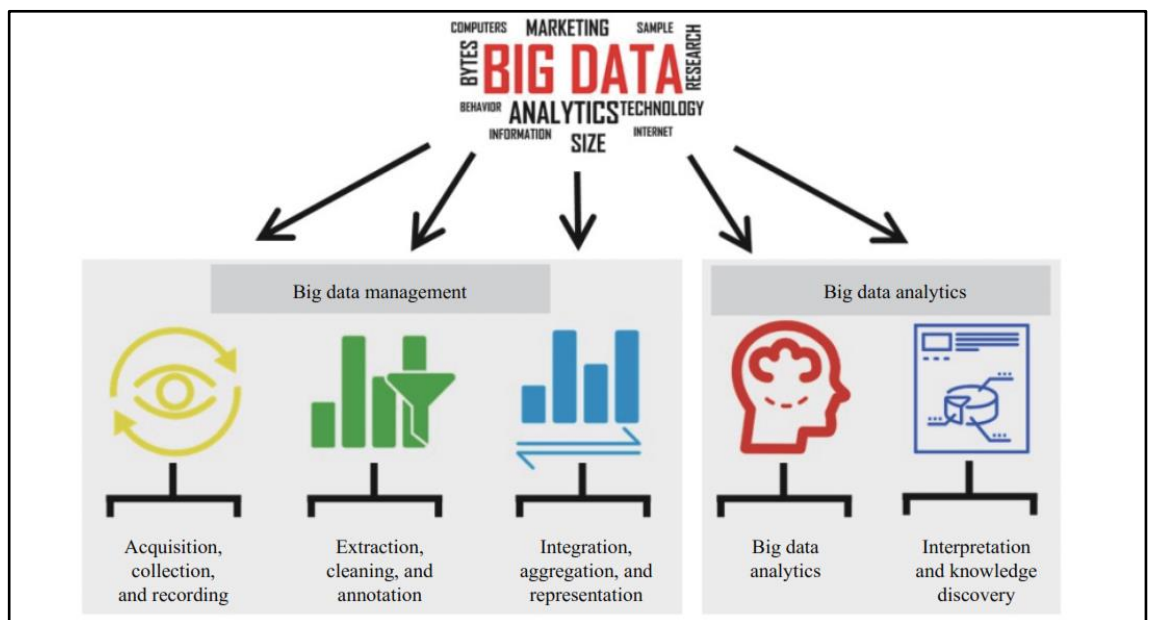


Figure 2.5 The major processes of big data (Qi, 2020)

Data sources in the mining industry can be classified into direct and indirect data sources. Direct data sources, such as Global Positioning System (GPS), conventional geodetic measurement, and commodity price monitoring, represent the data taken from instruments that are specified for data collection. Indirect data sources are data collected as a by-product of mining operations, such as drill and blasting, plant control, and rail track system. Fig. 4 illustrates the major data sources in the mining industry (Qi, 2020).

With the development of ICT, the data sources in the mining industry will expand. The use of sensors and smart chips to measure system performance in the mining industry has increased sharply, providing indicators for potential failures. Drone systems can be employed to explore new mining areas, monitor landscapes, and measure mine tailing ponds (Lee and Choi, 2016). Promoting the concept of big data is thus significant for the sustainable development of the mining industry.

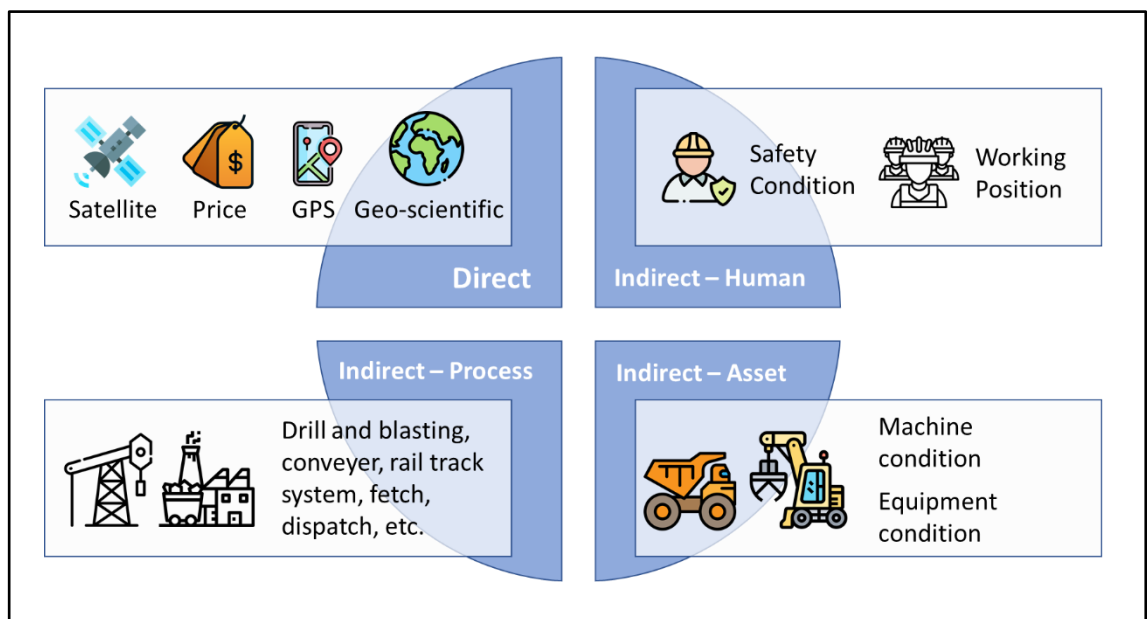


Figure 2.6 The major data sources in the mining industry (Qi, 2020)

BDM, together with BDA, can contribute to every stage of mining operation. The ultimate objective of big data to the mining industry is to reduce mining interruptions, increase efficiency, reduce costs, and increase profit margins. The following are the specific benefits of embracing big data in the mining industry (Qi, 2020):

1. Improve efficiency. Once it has been properly compiled, analyzed, and evaluated, big data can speed up mining operations, such as ore excavation, extraction, and separation, as a result of the power of automation, agile decision-making, and real-time optimization.

2. Energy management. Big data can help save energy in every mining operation, making the mining process more energy saving and sustainable.
3. Improve logistics. With automated and optimized transportation, big data can help improve transportation efficiency, reduce operational costs, and identify areas for improvement.
4. Intelligent business. Through real-time monitoring and analysis of mineral prices, the decision-making process can be improved by big data. Business operation efficiency can improve, and big data can help identify the cost distribution in all mining operations. Moreover, collaborations between various departments and even various companies can be solidified and become smarter with advanced sharing and communication platforms.
5. Safety, security, and reliability. Big data plays a vital role in the safety, security, and reliability of mining operations. The advancement of ICT makes real-time monitoring of practitioners (location, heart rate, temperature, etc.), equipment (posture, operating condition, etc.), and the environment (CO₂, gas concentration, temperature, dust level, etc.) possible. Precursors of mining accidents can be identified based on the above real-time data.
6. Improved operational management. Big data can help analyze machine reliability and optimize the need for spares. Moreover, smarter procurement of future services can be achieved, which can make prices more negotiable and reduce the overall procurement costs. Predictive maintenance can be achieved with the utilization of big data to reduce interruptions in mining operations

2.3 Data Mining

Data mining is the process of extracting previously unknown information and patterns from large amounts of data. Data mining uses artificial intelligence, statistics,

mathematics and machine learning in the process of extracting information for decision making (Bhatia, 2019). According to Azevedo and Santos (2008), data mining is often called knowledge discovery in database (KDD), which is a process that includes collecting, using historical data to find patterns or relationships in large datasets. The output of this data mining can be used to improve decision making in the future. In general, the main functions of data mining are descriptive functions and predictive functions. The descriptive function is used to understand more about the observed data so that it can find out its characteristics, while the predictive function is used to find a certain pattern from the data that can be used to predict other unknown variables. According to Provost and Fawcett (2013) some other functions of data mining are:

1. Classification. The classification function is used to predict whether an individual belongs to a population category.
2. Regression. The regression function is used to estimate the numerical value of one or more variables, usually based on the value of the other variables.
3. Similarity matching. This function is used to identify similarity of known data.
4. Clustering. This function is used to divide each data into several previously unknown groups based on the similarity between the data.
5. Association rule discovery. This function is used to find relationships between data entities based on data records that occur.
6. Profiling. This function is used to characterize the characteristics of an individual, group or population of data.
7. Link prediction. This function is used to predict the relationship between data and estimate the strength of the data relationship.
8. Data reduction. This function is used to optimize very large data by replacing it with smaller data that has a lot of important information from large data sets. Smaller data sets can be better at providing information if processed properly.

2.4 The Stages of Data Mining

There are many methods for analysing data. One of the most popular is the Cross Industry Standard Process for Data Mining (CRISP-DM) which is used by more than 40% of data analysts. About 27% of analysts use their own method and the rest use various other methods (KDnuggets, 2014). CRISP-DM is an industry-independent process model for data mining. It consists of six iterative phases, namely business understanding, data understanding, data preparation, modelling, evaluation and deployment as shown in Figure 2.7 (Provost and Fawcett, 2013; Schroer et al., 2021).

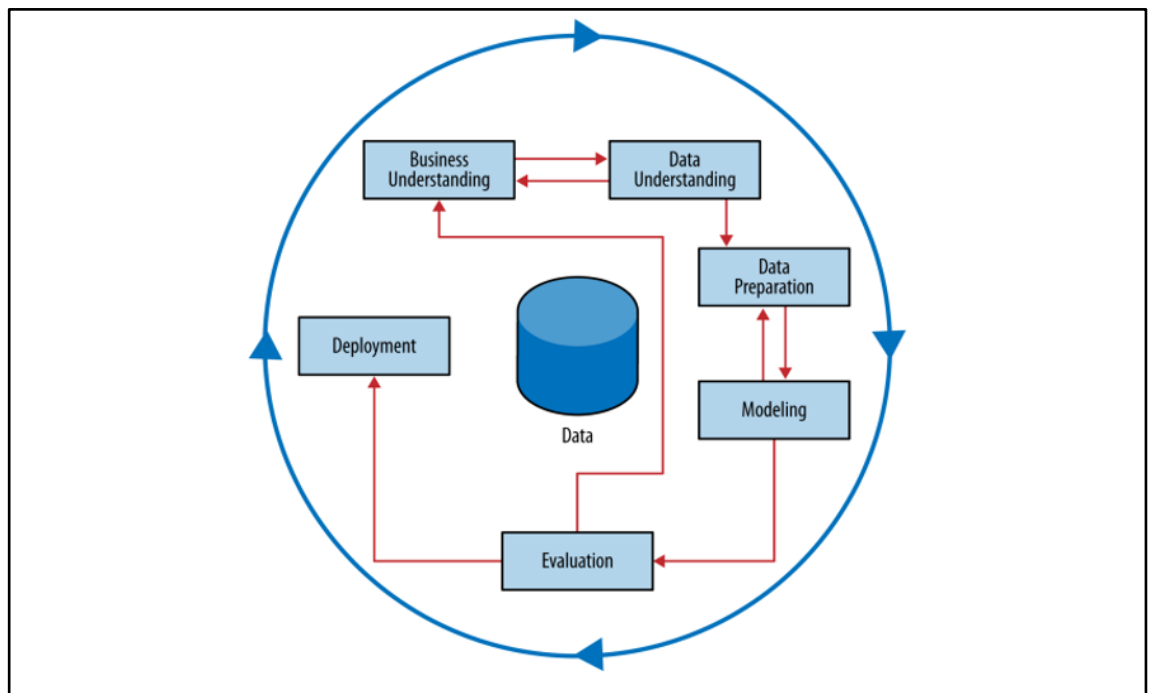


Figure 2.7 The stages of data mining (Provost and Fawcett, 2013)

The stages in data mining are actually carried out iteratively according to the problem to be solved. More detailed explanation for each stage in CRISP-DM are as follows:

1. Business understanding. This stage aims to understand the workflow of a business process. The better understanding and purpose of the process, the clearer the choices that can be made and the more effective the results obtained.

2. Data understanding. This stage aims to understand the raw data used. An understanding of the strengths and limitations of data is important because it is rare to find data that matches the problem to be solved.
3. Data preparation. This stage is also called data pre-processing. The general steps of data preparation consist of several sub-processes, including:
 - a. Data cleaning. The data cleaning process is the process of removing noise and irrelevant data. In general, the data obtained, both from a company's database and experimental results, have imperfect entries such as missing data, invalid data or just a typo. In addition, there are also data attributes that are not relevant to the data mining hypothesis they have. It is also better to discard irrelevant data. Data cleaning will also affect the performance of data mining techniques because the data handled will be reduced in number and complexity.
 - b. Data integration. Data integration is the merging of data from various databases into one new database. Not infrequently the data needed for data mining does not only come from one database but also comes from several databases or text files.
 - c. Data transformation. Data is converted or combined into a format suitable for processing in data mining. Some data mining methods require special data formats before they can be applied. For example, some standard methods such as association analysis and clustering can only accept categorical data input. Therefore, data in the form of continuous numeric numbers needs to be divided into several intervals. This process is often called data transformation.
4. Modelling. The modelling stage is the main stage of data mining techniques carried out to analyse the cleaned data. The modelling technique, method or algorithm used is very dependent on the goal or problem to be solved. Machine learning algorithms take a lot of roles in this stage.

5. Evaluation. The evaluation stage aims to assess the results of data mining. In this stage the results of data mining techniques in the form of typical patterns and predictive models are evaluated to assess whether the existing hypothesis is indeed achieved. If it turns out that the results obtained do not match the hypothesis, there are several alternatives that can be taken, such as making feedback to improve the data mining process, trying other data mining methods that are more suitable, or accepting this result as an unexpected result that may be useful.
6. Deployment and visualization. Knowledge or information that has been obtained will be organized and presented in a special form so that it can be used by users. The deployment stage can be in the form of making simple reports, visualization with dashboard or implementing iterative data mining processes within the company.

2.5 Business Process of Mining Industry

Business understanding stage in CRISP-DM is a crucial stage and more important because decisions taken at other stages should be based on business understanding. Therefore, it is important to understand the business processes of the company or industry before applying data mining. This helps so that the system created is right on target and in accordance with the focus of solving existing problems (Sharma and Osei-Bryson, 2009).

Successful mine development and production requires a series of distinct steps, which include the discovery, identification and assessment of a mineral prospect, the construction and operation of the mine, the processing of the raw materials extracted from the mine, the closure and reclamation of the mine after production has ended. Each step in the sequence is distinct, and most mining projects progress systematically from one to the next. It is critical to remember that the time span between the discovery of a

mineral deposit and mine production is typically very long as shown in Figure 2.8 (Revuelta, 2018).

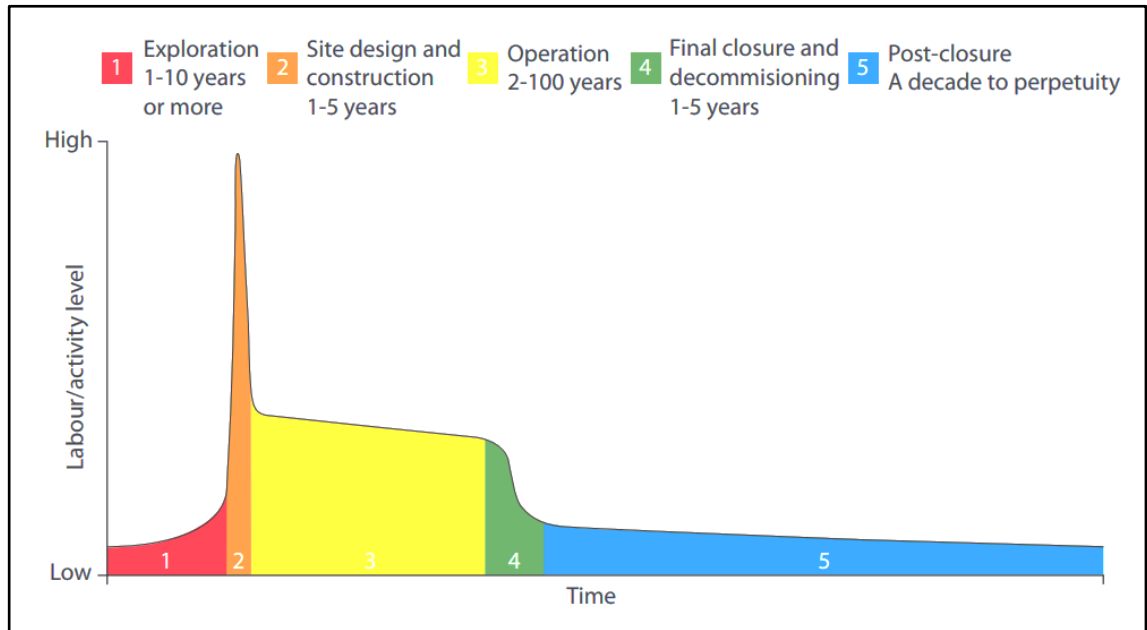


Figure 2.8 The life cycle of mine project (Revuelta, 2018)

These various phases come together to make up the mining cycle. The mining cycle generally divided into exploration, evaluation, exploitation, mineral processing and reclamation. Small mining projects may progress from exploration to mine production in a few years, with the mine closing 10 years after it begins operations. Large and complex mining projects, on the other hand, may require 20 years of exploration and several decades of mining (Revuelta, 2018).

1. Exploration

The discovery of mineral resources, also known as exploration or prospecting, is the first step in the mine cycle and involves a comprehensive set of multidisciplinary activities. Mineral resources are scarce and are buried beneath the Earth's surface. Economic mineral resources are even more difficult to obtain. Because mineral deposits are rare, finding one is difficult, and the chances of any exploration program succeeding are relatively low (Stevens, 2011).

2. Evaluation

The mineral resource evaluation process typically includes a technical, economic, and socioeconomic stage. The estimation of tonnage (quantity) and mineral or metal content (quality) from analytical data calculated in sample assays, either globally or for parts of the deposit, is the result of technical evaluation. The estimation is obtained using either traditional or geostatistical methods. The first methods are traditional but simple methods. Panel/section, polygons, inverse distance weighted, triangulation, and contour methods are some that were developed before the twentieth century (Annels, 1991).

Following technical evaluation, selecting the most appropriate economic evaluation method is critical. Many other project variables, such as production cost, capital cost, royalties, taxes, and so on, are also required to carry out this economic evaluation. It should be noted that the mining industry differs from other industries in several ways. For example, the mining industry requires many years of production before achieving a positive cash flow and requires a longer project life. Furthermore, the overall process is very capital intensive (Revuelta, 2018).

3. Exploitation

Mining will be the next step if the economic evaluation of a mining project yields positive results, indicating a high probability that the exploitation of the mineral deposit will yield benefits. The process of excavating and recovering ore and associated waste rock from the Earth's crust is known as mining. The criteria for mine method selection are based on rock competency, distance to surface, mineral characteristics, and economics. Mining methods are classified as surface or underground based on their proximity to the surface. Approximately 85% of global tonnage is produced in open-pit mines, including placer operations, with the remaining 15% produced in underground mines (Revuelta, 2018).

Surface mining is commonly used when the coal is less than 200 feet underground. Surface mining involves the removal of topsoil and rock layers known as overburden in order to expose coal seams. Large opencast mines can span many square miles and use massive equipment such as draglines to remove overburden, power shovels, large trucks to transport overburden and coal, bucket wheel excavators, and conveyors. The soil and rock overburden are first broken up with explosives and then removed with draglines or shovel and truck. Once the coal seam has been exposed, it is drilled, fractured, and mined in strips. The coal is then loaded onto large trucks or conveyors for transport to either the coal preparation plant or directly to its final destination (Bargawa, 2018; US Energy Information Administration, 2022)

Explosives are frequently required in the extraction of mineral resources. As a result, blasting is frequently used as part of the mining process. Blasting is the process of fracturing material by using an explosive loaded in special holes. Today, a variety of explosives are used, including ANFO (ammonium nitrate plus fuel oil), slurries, and emulsions. The blasting holes are loaded so that each one is fired in a predetermined sequence to achieve the desired rock break. After that, the explosives are detonated in the drill holes (Revuelta, 2018).

4. Mineral Processing

Mineral processing separates useful minerals from waste rock or gangue after extraction, resulting in a more concentrated material for further processing. Beneficiation or concentration refers to the process of concentrating valuable minerals in run-of-mine material. The goal is to reduce the bulk of the material by using low-cost, low-energy physical methods to separate valuable minerals from waste rock (Revuelta, 2018).

5. Closure and Reclamation

Mine closure is the final stage in the mining cycle because mining is a temporary activity with operating lives ranging from a few years to several decades. When a mineral

resource is depleted or operations are no longer profitable, the process of closure begins. Most regulatory agencies around the world require mine closure plans before issuing a mining permit. Many countries require financial assurance as a guarantee that the funds required for mine closure will be available if the responsible company is unable to complete the process as planned. Reclamation involves earthwork and site restoration, including revegetation of disposal areas, and occurs at all stages of mine life (environmental analysis begins at the earliest stages of mineral resource exploration) (Revuelta, 2018).

2.6 Machine Learning

Machine learning is the study of computer algorithms that can recognize patterns in data with the aim of converting various kinds of data into real actions with as minimum human intervention as possible. In general, machine learning is part of artificial intelligence (AI) (Kurniawan, 2020). According to El Morr and Ali-Hassan (2019), machine learning includes many different algorithms which are divided into two main categories namely supervised and unsupervised. Supervised learning relies on data that has been processed by experts such as labelling while unsupervised learning does not depend on labelling by experts. However, there are more category of machine learning algorithm called semi-supervised learning (Ge et al., 2017) and reinforcement learning (Bonaccorso, 2017). but a more complete theory of the two categories is not explained further in this study.

1. Supervised Learning

In supervised learning, the training data set contains the correct input and output data. It then "learns" how to process this data in a certain way so that it ends up with a solution provided as output. The learning process is about building a model that can make predictions of possible outcomes in the presence of unknown outputs or solutions.

Once the learning is done, the supervised learning software uses the model it learns to provide an output prediction for each new input to data set. Supervised learning algorithms can use classification and regression techniques. A classification technique starts with a data set and predicts if an output belongs to a certain category. Some of the main techniques used for classification are:

- a. Classification trees.
- b. Fuzzy classification.
- c. Random forests.
- d. Artificial neural networks.
- e. Discriminant analysis.
- f. Naive Bayes.
- g. K nearest neighbor.
- h. Logistic regression.
- i. Support vector machine.
- j. Ensemble methods.

Instead of classifying in categories, a regression technique predicts a value of an output or variable of a continuous nature such as a number. Simply understood, a regression technique approximates a function f of a data input x that produces an output $y = f(x)$; x and y are known, and f is approximated. Some of the main techniques used in regression are:

- a. Linear regression.
- b. Generalized linear model.
- c. Decision trees.
- d. Bayesian networks.
- e. Fuzzy classification.
- f. Neural networks.

- g. Gaussian process regression.
 - h. Relevance vector machine.
 - i. Support vector regression.
 - j. Ensemble methods.
2. Unsupervised Learning

In unsupervised learning, there is no known output for the data input, but we have no idea about the possible results we are searching for. The algorithm tries to detect hidden patterns or structures within the data set; once learning is performed, the algorithm will then be able to predict the possible output or solution from a new data set in the future. Two main techniques used in unsupervised learning are clustering and dimension reduction. Clustering aims to find hidden patterns and groupings within the data. It takes a data set as an input and partitions it into clusters. Some of the main algorithms used in regression are:

- a. K-means clustering.
- b. Hierarchical clustering.
- c. Genetic algorithms.
- d. Artificial neural networks.
- e. Hidden Markov model.

Dimension reduction is mainly interested in reducing the number of variables needed to represent the data input, thus projecting the data to fewer dimensions. The reduction of the data dimensions will allow simpler representation of the data and faster processing time. Some of the main algorithms used in regression are:

- a. Principal component analysis.
- b. Linear discriminant analysis.
- c. Multidimensional statistics.
- d. Random projection.

2.7 Relationship of Data Mining and Machine Learning

Data mining and machine learning have an intersecting relationship. Machine learning is more about building systems that do specific jobs automatically. While data mining is more of an analytical job to look for hidden patterns in the data to help decision making. In a broader scope, data mining is part of data science. Data science is a multidisciplinary science that studies efforts to gain a deeper understanding of various kinds of data. in order to obtain conclusions from the information contained in the data, so that people can take the right decisions and actions (Kurniawan, 2020).

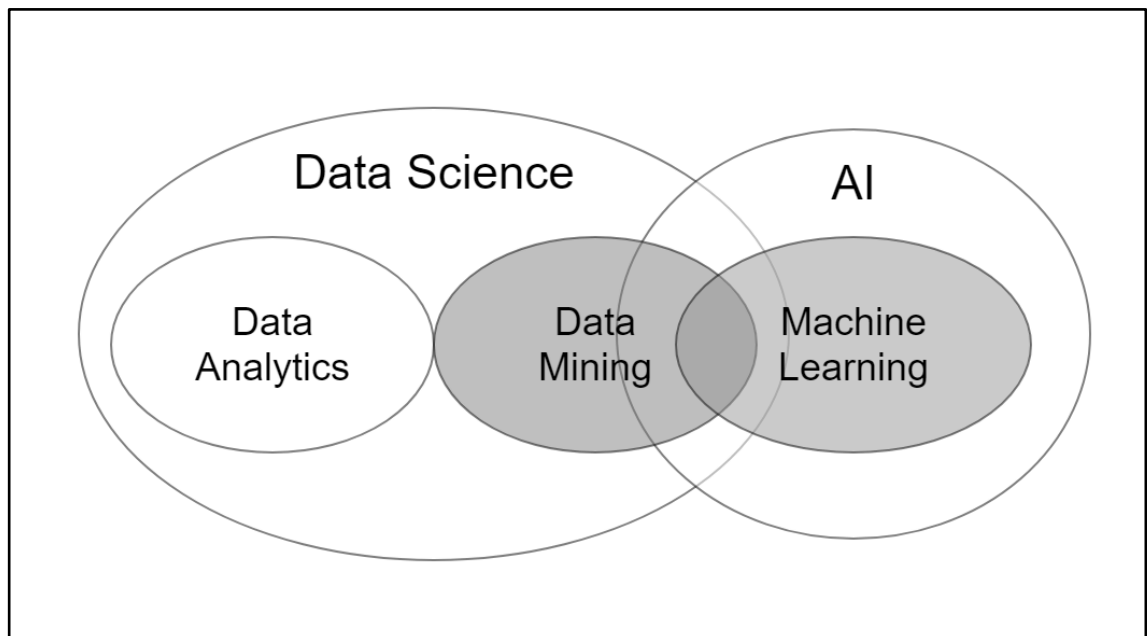


Figure 2.9 Data science, data mining and machine learning (Kurniawan, 2020)

2.8 Clustering in Machine Learning with K-Means Algorithm

Clustering is a process to group a set of data objects that have a high similarity into one group or cluster and each group that is formed has no resemblance to other groups (Han et al., 2011). Similarity is assessed based on the attribute value that describes the data object. Clustering or also referred to as segmentation is one of the

unsupervised learning methods, because there are no attributes that are used as a guide or there are no labels on the data in the learning process. Techniques in clustering are carried out to find knowledge from datasets (Chiu et al., 2009). Approach in cluster based on suggestion from Fraley and Raftery (1998), divided the clustering method into two main groups, namely the hierarchical method and the partition method.

1. The hierarchical method is a method that forms clusters by repeatedly partitioning from top to bottom or vice versa. The results of the hierarchical method are in the form of dendrograms that represent groups of objects and the level of similarity where there is a change in grouping. A grouping of data objects is obtained by cutting the dendrogram at the desired level of similarity.
2. The partition method is a method that initializes k partitions at the beginning, where the k parameter is the number of partitions to form. Then iteratively uses the relocation technique by trying to repeatedly move objects from one group to another to get the optimal partition. This type of partition method is like K-Means, K-Medoids and CLARANS (Han et al., 2011).

In this study, the clustering method used is the partition method because it aims to create groups of mining contractors where each contractor is only in one particular group. The K-Means algorithm is one of the clustering algorithms that aims to divide the data into several groups. This algorithm accepts input in the form of data without class labels (unsupervised learning). The K-Means clustering process is carried out by a computer by grouping the data into its own input without first knowing the target class. In each cluster there is a central point (centroid) that represents the cluster. The algorithm to perform K-Means clustering is as follows (Tan et al., 2014):

1. Determine the value of k as the number of clusters formed

- Determines the initial value of the centroid or the centre point of the cluster. At this stage the centroid value is determined randomly, but for the next stage using the Equation 2.5.

$$V_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} X_{kj} \dots\dots\dots (2.5)$$

Where:

V_{ij} = cluster centroid (i) for variable (j)

N_i = amount of data in cluster (i)

i, k = cluster index

j = variable index

X_{kj} = Data value (k) in cluster for variable (j)

- Calculating distance between the centroid point and the point of each object, can be done using Euclidean Distance with the Equation 2.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_t - y_t)^2} \dots\dots\dots (2.6)$$

Where:

D_e = Euclidean Distance

i = data index

(x, y) = data coordinates

(s, t) = centroid coordinates

- Grouping the data to form clusters with the centroid point of each cluster being the closest centroid point. Determination of cluster members is to take into account the minimum distance of the object.
- Updating the centroid value of each cluster.
- Repeat step b to the end until the value of the centroid point no longer changes.

2.9 Determination of the Optimal Number of Cluster

Determination of the optimal number of clusters can use the Elbow method. The Elbow method provides ideas by choosing the cluster value and then adding the cluster value to be used as a data model in determining the best cluster. This method will produce information in determining the best number of clusters by looking at the percentage of comparison results between the number of clusters that form an angle at a point (Bholowalia and Kumar, 2014). Determination of the optimal number of clusters using the elbow method is done by looking at the SSE (Sum of Squared Error) value. SSE is shown as in Equation 2.7 and Equation 2.8 (Thinsungnoen et al., 2015).

$$SSE = \sum_{i=1}^K \sum_{x_j \in C_i} |X_j - M_i|^2 \dots\dots\dots (2.7)$$

$$M_i = \frac{\sum_{i=1}^n X_i}{n} \dots\dots\dots (2.8)$$

Where:

K = Number of clusters

X = {x1, x2, ..., xi, ..., xn}

C = {C1, C2, ..., Ci, ..., Cn}

M = Centroid of cluster (Ci)

In addition to SSE, silhouette score can also be used to select optimal values from clusters with a suitable ratio scale to define clearly separated clusters. Clustering is considered as the average closeness between clusters that have similarities and differences. Silhouette score is calculated as in Equation 2.9 (Thinsungnoen et al., 2015).

$$S = \frac{b-a}{\max \{a, b\}} \dots\dots\dots (2.9)$$

Where:

a = The average distance from one data to another in the same cluster

b = The average distance from one data to the next closest data in the nearest cluster

The result of the Silhouette Score calculation will be expressed in $-1 < S < +1$ (Wang et al., 2017). According to Kaufman and Rousseeuw (1990), the subjective interpretation of the Silhouette Coefficient (SC) can be described as in Table 2.1.

Table 2.1 Subjective interpretation of silhouette score (Kaufman and Rousseeuw, 1990)

Silhouette Score	Interpretation
0.71 – 1	A strong structure has been found
0.51 – 0.70	A reasonable structure has been found
0.26 – 0.50	The structure still weak
≤ 0.25	No substantial structure has been found

2.10 Python Programming Language

Python is a high-level programming language that is widely known for being easy to learn. In addition, the wide Python community makes many tools and libraries available that are very interesting in working on data science, machine learning and scientific computing. Python still holds the top spot for the most widely used language in analytics, data science, and machine learning (Raschka et al., 2020). According to (Ceder, 2018), the reasons why Python is the best choice for data analysis and superior to other programming languages are as follows:

1. Python is a modern, high-level language with dynamic typing and simple, consistent syntax and semantics.
2. Python is multiplatform, highly modular, and suited for both rapid development and large-scale programming.
3. It is reasonably fast and can be easily extended with C or C++ modules for higher speeds.

4. Python has built-in advanced features such as persistent object storage, advanced hash tables, expandable class syntax, and universal comparison functions.
5. Python includes a wide range of libraries such as numeric processing, image manipulation, user interfaces, and web scripting.
6. It is supported by a dynamic Python community

One of the main advantages of Python is the ability to interact directly with code using the terminal or other tools such as Jupyter Notebook. Machine learning and data mining are essentially iterative processes so it is important to use tools that allow fast iteration and easy interaction. Python has many libraries for data processing, visualization, statistics, to text and image processing. Some important libraries that are often used in data analysis in Python include Numpy, Scipy, Matplotlib, Pandas, Scikit-learn and many more (Muller and Guido, 2016; Nongthombam, 2021).

1. NumPy: NumPy is a library written in Python, used for numerical analysis in Python. It stores the data in the form of nd-arrays (n-dimensional arrays).
2. Pandas: Pandas is mainly used for converting data into tabular form and hence, makes the data more structured and easily to read.
3. Matplotlib: Matplotlib is a data visualisation and graphical plotting package for Python and its numerical extension NumPy that runs on all platforms.
4. Seaborn: Seaborn is a Python data visualisation package based on matplotlib that is tightly connected with panda data structures. The core component of Seaborn is visualisation, which aids in data exploration and comprehension.
5. Sklearn: Scikit-learn is the most useful library for machine learning in Python. It includes numerous useful tools for classification, regression, clustering, and dimensionality reduction.

2.11 Data Visualization using Tableau

Data visualization is visual communication that makes data easier to understand. Data visualization shortens the time to convert data into knowledge thereby enabling better data driven decisions. Tableau is one of the best visualization tools with the following important features (Acharya and Chellappan, 2017).

1. Provide a variety of graphs and chart forms in a clear and specific way.
2. A number of Tableau products allow reports and dashboards to be created more quickly and easily.
3. There are several features available to understand the data, interpret the data and analyse it statistically.
4. Tableau can connect to various data sources from traditional ones like Excel and text files to non-traditional ones like social media data, Microsoft Azure, etc.

Tableau in this study will be used in the manufacture of mine production dashboards. Visualizations in the dashboard will be integrated with the machine learning model created.

2.12 Mining Production

Mining production, especially coal mining, is divided into two classes namely overburden (OB) and coal getting (CG). Overburden is the rock or soil layer that needs to be removed in order to access the ore being mined. Even though it has no economic value, OB is considered as production because it takes a certain OB stripping target to be able to extract the coal. Mining production is calculated with Equation 2.10 (Basuki et al., 2020):

$$\text{Production} = PA \times UA \times P_{dty} \times MOHH \dots\dots\dots (2.10)$$

Where:

- Production = Total work result per unit of time (Ton/day or BCM/day)
- PA = Physical Availability
- UA = Use of Availability
- Pdty = Productivity (Ton/hour or BCM/hour)
- MOHH = Machine on Hand Hour (24 hours)

PA and UA are parameters that shows the availability of the equipment used. PA shows the time of physical availability of equipment while UA shows the effective time of equipment that can be used to operate in an undamaged condition. PA and UA are calculated using Equation 2.11 and Equation 2.12 (Rivai and Octova, 2021).

$$PA = \frac{W+S}{W+R+S} \times 100\% \dots\dots\dots (2.11)$$

$$UA = \frac{W}{W+S} \times 100\% \dots\dots\dots (2.12)$$

Where:

- PA = Physical Availability
- UA = Use of Availability
- W = Working hours
- S = Standby hours
- R = Repair hours

Productivity is production result per unit of time (production/hour). There are two types of equipment that are commonly used, namely loader and hauler. Productivity calculation usually depend on type of equipment used. But in general, productivity of mining equipment has the same calculation principle. The commonly used productivity formula is shown as in Equation 2.13 (Sokop et al., 2018).

$$Pdty = q \times N \times Ek \dots\dots\dots (2.13)$$

Where:

- Pdty = Productivity (Ton/hour or BCM/hour)
- q = Equipment capacity (BCM/m3)
- N = Number of cycles per hour = (60/CT)
- CT = Cycle time (minutes)
- Ek = Working efficiency

In this study, data used is the average production parameters of several units for each equipment type. The production parameters referred to include physical availability (PA), use of availability (UA) and productivity (Pdty). Therefore, the calculation of total production for each equipment type uses Equation 2.14.

$$\text{Total Production} = \text{PA} \times \text{UA} \times \text{Pdty} \times \text{Qty} \times \text{MOHH} \dots \dots \dots (2.14)$$

Where:

- PA = Average Physical Availability
- UA = Average Use of Availability
- Pdty = Average Productivity
- Qty = Equipment quantity
- MOHH = Machine on Hand Hour (24 hours)