USING DATA MINING TECHNOLOGY TO ANALYZE TRADITIONAL GOLD MINING IN SUDAN

2023
MASTER THESIS
COMPUTER ENGINEERING

Elamin SALAHELDIN

Thesis Advisor
Assist. Prof. Dr. Yasin ORTAKCI

USING DATA MINING TECHNOLOGY TO ANALYZE TRADITIONAL GOLD MINING IN SUDAN

Elamin SALAHELDIN

Thesis Advisor
Assist. Prof. Dr. Yasin ORTAKCI

T.C.
Karabuk University
Institute of Graduate Programs
Department of Computer Engineering
Prepared as
Master Thesis

KARABUK
February 2023

I certify that the master's thesis Elamin SALAHELDIN presented, "USING DATA MINING TECHNOLOGY TO ANALYZE TRADITIONAL GOLD MINING IN SUDAN" is, in my view, flawless in terms of quality and breadth.

Assist. Prof. Dr. Yasin ORTAKCI ....
Thesis Advisor, Department of Computer Engineering

This thesis is approved as a Master of Science thesis by the examining committee at the Department of Computer Engineering with a unanimous vote. 23/02/2023

Examining Committee Members (Institutions)	Signature
Chairman : Assoc. Prof. Dr. Rafet DURGUT (BANU)	....
Member : Assist.Prof.Dr. Sait DEMİR (KBU)	....
Member : Assist. Prof. Dr. Yasin ORTAKCI (KBU)	....

The Administrative Board of the Institute of Graduate Programs at Karabuk University has authorized the awarding of the Master of Science degree based on the submitted thesis.

Prof. Dr. Müslüm KUZU ....
Director of the Institute of Graduate Programs

“I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well.”

Elamin SALAHELDIN

ABSTRACT

M. Sc. Thesis

USING DATA MINING TECHNOLOGY TO ANALYZE TRADITIONAL GOLD MINING IN SUDAN

Elamin SALAHELDIN

Karabük University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Yasin ORTAKCI

February 2023, 56 pages

The traditional gold mining sector employs over two million people and spans the majority of Sudan. This study employed data mining methodology to aid decision-making by employing five models: Logistic Regression (LR), Decision Tree (DT), naive bayes (NB), Support Vector Machine (SVM) and K-Nearest Neighbors (K-NN), A fair comparison of their performance was made. These models divide firms into two categories: active, highly productive firms and inactive, low-productivity firms. The learning process was divided into four stages: raw data processing, training, testing, and validation. The results demonstrated that the proposed models successfully classified the state of the company's work. The models achieved high accuracy for DT, LR, and NB classifiers of 1.00, 0.91, and 0.81, respectively. When compared to the other models, the SVM and K-NN models decreased by 0.57 and 0.53, respectively.

Key Words : *Decision Tree; Support Vector Machine; K-nearest neighbor; Naive Bayes; Logistic Regression; Traditional Mining.*

Science Code : 92431

ÖZET

Yüksek Lisans Tezi

SUDAN'DA GELENEKSEL ALTIN MADENCİLİĞİNİ ANALİZ ETMEK İÇİN VERİ MADENCİLİĞİ TEKNOLOJİSİNİ KULLANMA

Elamin SALAHELDİN

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Yasin ORTAKCI

Şubat 2023, 56 sayfa

Geleneksel altın madenciliği sektörü, iki milyondan fazla kişiyi istihdam etmekte ve Sudan'ın büyük bölümünü kapsamaktadır. Bu çalışma, beş model kullanarak karar vermeye yardımcı olmak için veri madenciliği metodolojisi kullanmıştır: Destek Vektör Makinesi (SVM), lojistik regresyon (LR), naive bayes (NB), karar ağacı (DT) ve en yakın Komşular (K-NN), Performanslarının adil bir karşılaştırması yapıldı. Bu modeller firmaları iki kategoriye ayırır: aktif, yüksek verimli firmalar ve aktif olmayan, düşük verimli firmalar. Öğrenme süreci dört aşamaya ayrıldı: ham veri işleme, eğitim, test etme ve doğrulama. Sonuçlar, önerilen modellerin şirketin çalışma durumunu başarıyla sınıflandırdığını gösterdi. Modeller sırasıyla 1.00, 0.91 ve 0.81'lik DT, LR ve NB sınıflandırıcıları için yüksek doğruluk elde etti. Diğer modellerle karşılaştırıldığında, SVM ve K-NN modelleri sırasıyla 0,57 ve 0,53 azalmıştır.

Anahtar Kelimeler : Karar ağacı; Destek Vektör Makinesi; K-en yakın komşu; Naif bayanlar; Lojistik regresyon; Geleneksel Madencilik.

Bilim Kodu : 92431

ACKNOWLEDGMENT

We must constantly acknowledge and appreciate those who have assisted us and provided a helping hand when required, and we must always show our delight in their presence and our gratitude for their support.

I want to thank the sincere ones who exerted all their efforts in helping to accomplish this work. Especially mention Assist. Prof. Dr. Yasin ORTAKCI**,** who has great credit for directing and assisting in the suggestions and directives he gave to make this work a success.

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

ABBREVITIONS INDEX

ABBREVITIONS

KDD	: Knowledge Discovery in Databases
DT	: Decision Tree Algorithm
SVM	: Support Vector Machine
KNN	: K-Nearest Neighbor
LR	: Logistic Regression
NB	: Naive Bayes

PART 1

INTRODUCTION

The fast advancement of technology across a wide range of human endeavors has led to the everyday generation of massive amounts of data. In order to process this volume of data, we require instruments that speed up our analysis and, if possible, improve its precision. Due to this demand, information mining, a new branch of science, has emerged. Exploration of Existing Databases and Data Mining Knowledge is a branch of science that focuses on the discovery of patterns and models that were not previously recognized or obvious and that are buried deep within enormous databases. The ultimate goal of this endeavor is to gain a greater comprehension of the information at our disposal so that we can make more informed choices [1].

The mining industry is one of the most important industries and acts as the economic fulcrum for a number of countries all over the globe. Laws have been enacted to ensure the health and safety of employees, as well as the protection of the environment, in mining operations. This is done with the goal of preventing environmental damage caused by inefficient mining methods. Because of this, there are quarries that are run by automated systems to cut down on the dangers faced by employees. In Sudan, many projects surfaced in the field of gold mining as well as other minerals by national and multinational companies, and these companies' acquired contracts for mining in a structured fashion. The projects included both gold mining and mining for other minerals. Traditional gold mining also emerged and developed into a significant barrier that posed a risk to the mining industry as well as the economics of the country. It also left a negative impact on the environment and put employees in potentially hazardous situations. In addition, conventional methods of gold mining began to gain traction and developed into a significant obstacle that posed a risk to the mining industry as well as the economics of the country. In addition, it had an adverse effect on the natural world and placed the safety of the workforce in jeopardy.

In order to legitimize traditional mining and transform it into official mining under government control, the Ministry of Minerals in Sudan keeps a large number of records spanning a variety of time periods. These records contain information on the productivity of traditional mining as well as the quantity of trash that is produced and the proprietors of the organizations that engage in traditional mining. The question of how to make conventional mining lawful while still making a profit from this data kept coming up. As a result, it was necessary to perform analysis on these data in order to reap the benefits of doing so in terms of producing effective decisions and ensuring the continuation of this work in the future at a reduced cost and weight. As a result, the model based on gathering large amounts of data was the most effective model for resolving this issue. Data mining is described as the process of examining data from a variety of perspectives in order to discover disruptions, patterns, and relationships discovered in data sets that are informative and helpful for forecasting outcomes that assist us in making the best decision. Data mining can also be thought of as the practice of mining for data. Data mining is one of the subfields of computer science that is expanding at the fastest rate, and its prominence and growth have corresponded with the increasing demand for tools that can assist with the analysis and comprehension of enormous quantities of data [1].

This study's main goal is to determine if formalizing current mining practices is feasible in order to boost output while also lowering the possibility of unfavorable environmental effects. This will be accomplished through the application of computational methods that offer a high level of accuracy and sound judgment. In addition to this, it contributes to the growth of the national economy and safeguards both human resources and natural resources against the threat posed by the contamination of the environment by mining refuse. As researchers, it was our responsibility to devise strategies and an assortment of instruments in order to promote economic growth and development.

The study's importance lies in its use of data mining tools to analyze historical mining data and provide findings that may be used to turning these operations into legitimate enterprises and ending historical mining. The study's conclusions can be utilized to stop conventional mining. A large amount of information, up to 20,000 documents worth, was gathered through the collection of data. This information includes the proprietors of the businesses, their utilization of electricity and water, as well as the amount of trash extracted and the quantity of gold produced. This study used data mining methodology to help make decisions by comparing five models: Support Vector Machine, Logistic Regression, Naive Bayes, Decision Tree, and K-NN, in order to achieve the best results and high accuracy in classifying these companies in order to develop and legalize them as formal mining companies.

1.1. RESEARCH PROBLEM

Concerns have been raised regarding the mobility of people from different cultural backgrounds and educational levels as a result of the recent uptick in gold prospecting in Sudan, which has recently seen an increase in activity. In regions that traditionally supported mining, there have been documented to be a great number of adverse societal, economic, environmental, and security repercussions. This brings us to the most important question posed by the study: Is there a correlation between the implementation of development strategies by the government and corporates and the reduction of risk in the conventional gold mining sector? If so, what kind of significance does this relationship have?

The following points summarize the research problem:

- Traditional mining hinders the development of the mining industry in Sudan.
- Gold extracted through traditional mining is smuggled out of the country;
- Mine collapse and environmental pollution.

1.2. THE IMPORTANCE OF RESEARCH

In this segment, we will discuss the people who will benefit from the paradigm that was established. In addition, by enabling the legalization of high-production firms and turning them into official operations, the adoption of such a model in the mining sector will aid decision-makers in ending traditional mining.

The suggested model will be beneficial to the Ministry of Minerals and pertinent government businesses because it will provide them with the ability to know the real scale of official mining and access to the companies' mineral production. This will add accountability to statistical data while also boosting the economy of the country.
In addition, future academicians and researchers in the field will benefit from the research because it will serve as an instrument that will assist in the development of further research and creativity.

1.3. RESEARCH GOALS

The goal of this investigation is to legitimize informal mining and predict operating and non-operating projects to reduce non-operational projects and develop operational projects, which leads to the development of investment in the mining industry by understanding the true size of the mining industry.

1.4. BACKGROUND

A widespread commercial and societal movement has been ignited as a result of gold extraction activities in many regions of Sudan. Gold mining, in both its conventional and structured forms, is one of the solutions that the state depends on now as it searches for replacements to the funding it receives from oil after it has been removed from the state budget. Gold is one of the significant choices and alternatives, but it won't make up for all the losses that have been incurred. In 2011, gold sales brought in 1.5 billion dollars, and in January of 2012, 50 tons of gold were shipped out, which is reflective of the rising prices of gold and the accompanying increase in gold sales. Furthermore, the movement in gold exploration demonstrates that it is not only an economic return; rather, it has a significant social and developmental impact for those involved in its exploration operations, as approximately 200,000 citizens are now working in traditional mining, with 150 Sudanese and foreign companies also working; One of them is French in eastern Sudan, and they are all granted concessions to explore for gold. In the subsequent years, it is anticipated that earnings from gold will increase to $3 billion from the current level of $1.5 billion [2].

Traditional miners were provided with important environmental services for their care and protection, and they were also organized into unions and organizations at the same time that attempts to measure traditional miners in collaboration with the Central Statistical Bureau got launched. The Central Bank of Sudan's branch offices both purchase gold from mining and work to battle contraband. This fight against smuggling is an area in which significant efforts are being made in collaboration with the appropriate authorities. The sale of gold to the Central Bank is how it became the only source of gold while simultaneously prohibiting no one from transferring any gold. In addition, a gold refinery that is capable of producing one hundred tons of gold and fifty tons of silver each year was built. The Central Bank purchases two tons of gold each month at a cost of $50 million, bringing in a total of $1.5 billion in yearly earnings as a result of this activity.

One of the difficulties faced by mining in the various areas of the north and west is the issue of traditional gold mining and its significant danger to the environment due to the improper exploitation of mining, as well as its danger to sector workers as it exposes them to the risk of death by using chemicals and also the use of children in mining. Because of this, the state wants to make official mining lawful while outlawing traditional mining [2].

1.5. RESEARCH QUESTIONS

Taking into account the data acquired regarding the mining sector and the aforementioned problems. Our research was guided by the following questions.

- How will the data mining-based model classify companies with the highest and lowest productivity?
- How can the work of the data mining model be evaluated and compared to the traditional method of analysis?

1.6. RESEARCH SCOPE

The research was applied to the Sudanese Mineral Resources Company, which is the regulatory body that oversees mineral resources in Sudan. The information that was gathered from previous works that was relevant to the research questions served as the foundation for this research. This was done in order to collect facts and perspectives on the various things that can be done to enhance the mining business, the environment, and the welfare of the workers. Review and analysis of relevant scientific research in this field have been used as the basis for the model that has been suggested. In addition to this, the research offers a framework for the analysis and validation of data regarding gold productivity by making use of categorization algorithms and validating the findings.

1.7. SUMMARY

The mining business in Sudan was covered in this chapter on a more general level. The following is a brief introduction to data mining. Following that, a discussion of the challenges that lie ahead for the gold extraction industry in Sudan is presented. In addition to that, research questions and objectives are outlined here. After that, a discussion of the implications of the findings will take place. In conclusion, the chapter provides a brief overview of the overall purview for paper.

PART 2

LITERATURE REVIEW

2.1. INTRUSION DETECTION METHODS CLASSIFICATION

A significant rise may be seen as a consequence of the widespread adoption of internet usage and the subsequent meteoric rise in the number of devices that are linked to the internet. E-mails and social networking sites are two examples of the kinds of online platforms that have benefited significantly from the proliferation of internet usage throughout the globe. As a direct consequence of the rise in the total quantity of data that has been generated, a new notion known as "big data" has come into being [29]. When one considers the perspectives of authors as well as when one examines the findings of the analyses that are released by Domo every year, the terrifying proportions of the quantity of data created every minute are exposed [30].

A number of the benefits brought about by internet technologies make it possible for businesses to utilize this technology in order to improve the experience they provide to their customers and to present themselves to the rest of the world. The growth of the Internet has brought about a variety of advantages, but it has also ushered in a new set of challenges for maintaining network security. This circumstance has the potential to result in the loss of data as well as economic damages [5,6].

According to study in [7] intrusion detection systems may be separated into two categories: the first category is determined by the location of the installation inside the network, and the second category is determined by the detection technique. Figure 2.1 illustrates this distinction.

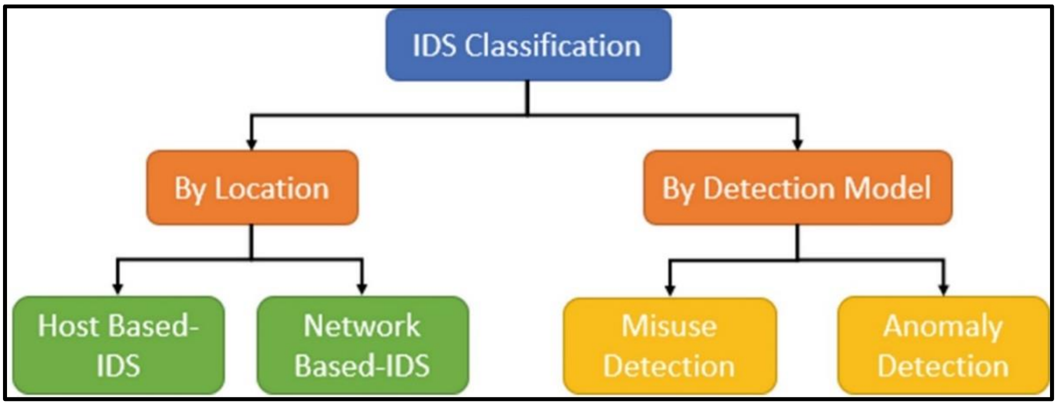

Figure 0.1. IDS system categorization [33].

In Figure 2.2 shown the deployment of host intrusion detection system.

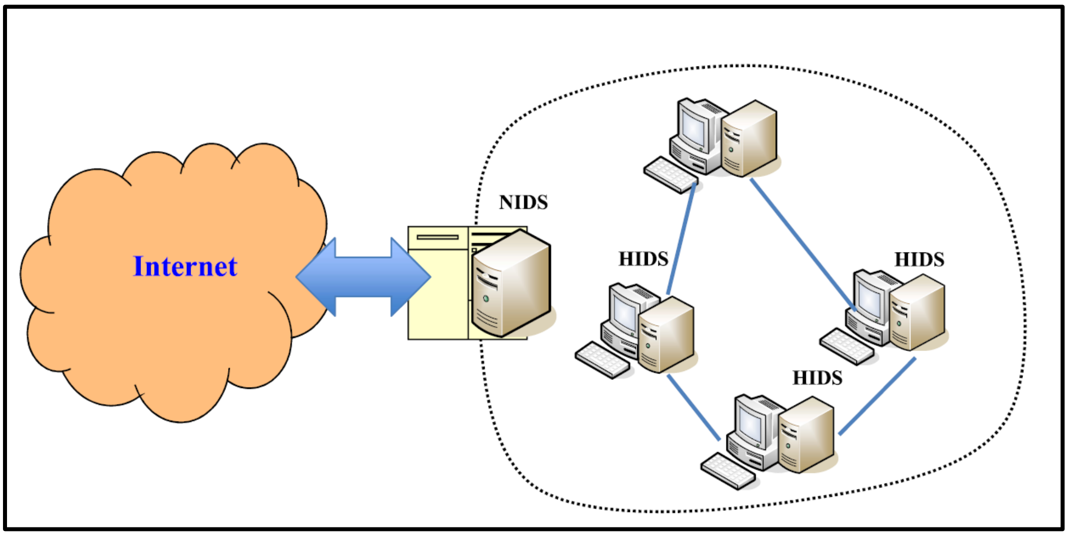

Figure 0.2. The deployment of host intrusion detection system with network intrusion detection system [6].

The first group may potentially be subdivided into two different kinds of techniques, host-based and network-based, respectively. The host-based intrusion detection system operates directly on the client PC and also processes the data that is stored on it, such as log documents, operating operations, and clients that are logged in. An alert is issued to the administrator if there is a change to an essential user file or the operating system itself. This allows the administrator to respond appropriately [8]. On the other hand, network-based intrusion detection systems will monitor and analyses packets as they go across a network in order to identify troublesome behaviors such as denial of service attacks [9].

This segment includes a compilation of studies and conversations based on the use of data mining models to achieve solid and effective decision-making in less time. These models are used to extract information from large amounts of data. A model that analyzes insurance risks and was suggested by Akinsola Adeniyi and colleagues and uses a data mining strategy. A decision tree was created to forecast the degree of risk in the car insurance industry based on the ID3 algorithm and using the following independent variables: previous claims, driver's license, age, drinking habits, visual impairment, safety means, vehicle use, physical disability, phone use, and vehicle garage. The research's results show that the system can predict the level of hazard with accuracy [3].

A model that helps to forecast student achievement at the conclusion of the semester was suggested by Baradwaj and Pal. The model employs data mining methods. The research utilized a database consisting of the information of fifty-one master's degree candidates in computer applications who attended Purvanchal Jaunpur University in India between the years 2011 and 2017. The technique known as the decision tree was used. The researchers came to the conclusion that a teacher could benefit from the decision tree algorithm by using it to anticipate - and in advance - the possibility of a student failing the end-of-semester test and, as a result, being required to receive additional care and attention [4].

Batista et al. They analyzed the advantages and disadvantages of the data processing methods, and he analyzed the behavior of three of the processing methods and used the K-NN algorithm to process the missing data, and concluded that the best method is NNI-10, which gave very good results even if the training records were in large quantities [5]. Duy-Hien Vu. presented a strategy that was developed in order to address the issue of anonymity. It has been suggested that the first component of a novel model called Semi-Distributed Numbers be committed to memory by the data consumer, the second component by the miner, and the third component by the general public. He used the NB categorization algorithm because it had a high level of precision and was applicable at the lowest possible expense [6].

Lyras et al. propose a model for the most common evaluation criteria in terms of educational effectiveness through predictive techniques, testing traits and relationships, as well as studying the results qualitatively and quantitatively, and came to reveal strong correlations in many features and distinguish the educational process as well as students and understanding of educational content [7]. In their study on the illness hepatitis C, Reza and coworkers developed a model by applying data mining techniques to a collection from the University of California. It makes use of six different algorithms: the supporting vector machine, the naïve Bayes algorithm, the decision tree algorithm, the random forest (RF) algorithm, and the K-Nearest Neighbors (KNN) algorithm. The findings were 0.921, 0.963, 0.953, 0.972, 0.896, and 0.998, respectively, and the RF model had the highest precision of 97.29 [8].

The design of Mohamed Amr was arrived at by contrasting four different versions. Assess the precision of each model, then evaluate and contrast their respective levels of performance. These models are decision trees (DT), K-NN, support vector machines (SVM), and artificial neural networks (ANN), and they are being applied to a dataset taken from the database collection at UCI. The evaluation of credit applications in Germany makes use of these facts. The numbers that were obtained were as follows: - 0.763 - 0.709 - 0.739 - 0.713. precision, memory, and precision were all evaluated based on predetermined criteria to determine which categorization method yielded the best results [9].

Anthony and the others the traffic accident analysis model is presented, and the research sample, which consists of 3,330 records of traffic accidents, was selected because there was a possibility of improved analysis of the findings and because it was reasonably simple to obtain the data. The algorithms were chosen to cover a wide range of knowledge discovery processes in the database and form an integrated package that starts from the internal discovery of the relationships between the data in the database and the discovery of similarities between the phenomena covered by this data until the classification techniques that allow the division of data into predefined categories and give a kind of insight future data [10].

Using the Classification Bayesian Naive and algorithms based on Decision Trees C4.5, Al-Radaideh et al. suggested a method for identifying vehicle insurance fraud and analyzing fraud patterns from data that was submitted. This method could also be used to analyze fraud patterns. In addition to evaluating the resulting models, it offers a concise explanation of the algorithms along with their implementations to identify plagiarism and deception in vehicle insurance. The accuracy of the decision tree model was 78.0 [11], and this article also evaluates the models.

2.2.SUMMARY

This chapter examines businesses and solutions that use data mining technology to solve problems and discover new patterns or knowledge. The companies and solutions are discussed in detail. In conclusion, this chapter discussed the present status of these works, in addition to some of the problems and deficiencies that are intrinsic to them.

PART 3

THEORETICAL BACKGROUND

3.1. THE ORIGINS OF DATA MINING

The emergence of data mining can be attributed to the inevitable progression of information technology. The progression of data management is depicted in Figure 3.1, beginning with the acquisition stage and continuing through production, administration (which includes storing, retrieving, and processing), and finally advanced analysis. In the beginning, we worked on developing mechanisms for accumulating and keeping data, and then we moved on to developing mechanisms for retrieving, processing, and searching data. The mechanisms of analysis and data mining have inevitably become the next level that we need to search for [1], after the majority of database management systems have provided the fundamental operations of processing and enquiry by default. The large amounts of data collected and stored in different types of warehouses have exceeded our capacity to absorb them and make meaningful use of them without the appropriate tools, leaving us in a situation known as data-poor but rich due to the abundance of data and the lack of tools required to analyze it. Because of this, it is extremely important to make use of these facts and transform them into information that can be used to assist in making decisions.

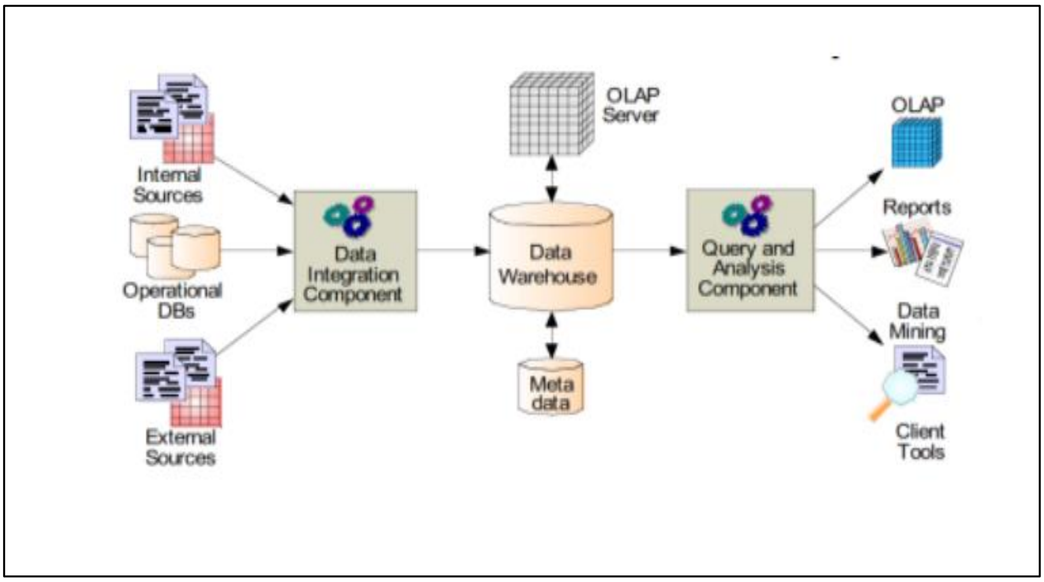

Figure 3.2. Illustrates online analytical processing and data mining.

3.2. DATA MINING

The term "data mining," which is synonymous with "knowledge mining," refers to the practice of collecting and generating various types of information from large quantities of data. The process of extracting small amounts of precious material from vast quantities of unprocessed material is referred to as mining. In any event, the word "data mining" is included in a variety of other titles, such as knowledge mining, which has become the most well-known of the bunch, knowledge extraction, and data/pattern analysis [1].

The process of data mining is frequently referred to as knowledge discovery. Data mining can be defined as the process of analyzing data from various angles, find relationships between them, and then present them into different pattern of information, such as information that can help increase or decrease costs, also its the process of finding useful information implementing of complex tools, such as normal statistics tools, artificial intelligence and computer-generated graphs. It required a significant amount of time and effort for research and development to reach the stage of preparedness and acceptance, which contributes to the impression that the life cycle of data mining is some kind of mysterious technology. It is a type of technology that employs a variety of algorithms in conjunction with traditional data processing techniques to analyze a wide variety of data kinds and derive knowledge from vast quantities of data [1].

In light of the enormous quantity of data that is available, there are two distinct kinds of information: the first is Online Analytical Processing (OLAP On-Line). The second type of analytical instrument is known as Data Mining (DM), and both of these kinds rely on data repositories to function properly. However, data mining is referred to as the inaccessible model of the recipients in the deepest levels of the data and is done in an automated manner without people's submission. The analytical analyzing of data utilizes a variety of points of view, showing the efficiency and speed of responding to the needs of the customers. The graphic that follows illustrates the key distinctions between online analytical processing and data mining in data repositories.

3.3. BENEFITS OF DATA MINING

According to the previous definitions, the data mining process is necessary to complete the tasks related to finding and extracting specific information from large amounts of data. However, the most prominent benefits of this process are:

- This process enables the beneficiaries to access information that other methods could not access.
- The decision-makers were able to find a deductive pattern by understanding the past in order to reach a prediction about the future of the required issue.
- This process enables the beneficiaries to know and discover the links, trends, and patterns prevailing in the work of a particular institution.

3.4. STAGES OF DISCOVERY OF KNOWLEDGE

Knowledge discovery in databases (KDD) is considered a step in discovering knowledge from databases. It extends to analysis and predictions of what will happen in the future the Figure 3.2 shows the knowledge discovery stages [11].

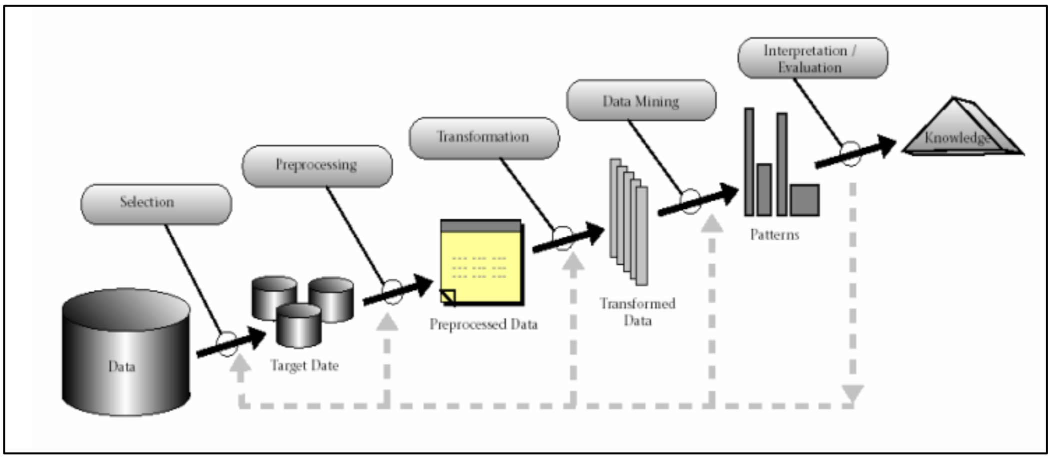

Figure 3.3. Knowledge discovery stages [11].

Finding information includes data mining, which is the most thorough method. The stages in the knowledge discovery process are as follows:

- Data discovery:
- It is the data collection stage and includes the detection, identification, and characterization of the available data [1].
- Data filtering and purification:
- At this stage, the annoying noise that is of no importance is removed, and the conflicting and inconsistent data are also deleted, and it includes the following steps:
  i. Data Cleaning
  ii. Variables Removing
  iii. Data Transformation:
  iv. Data Segmentation
- Data integration:
- At this stage, comparable and related data from various sources are gathered and integrated.
- Data selection:
- In this phase, the required data are now located in the dataset and retrieved.
- Data transformation:

- The data is now transformed into personalized formats appropriate for retrieval and search operations utilizing accomplishment summaries or aggregation algorithms.
- Data mining:
- Useful models should be extracted as many ways as possible to extract data patterns.
- Pattern evaluation:
- At this stage, important patterns are defined, and measures are used to ensure data accuracy.
- Knowledge representation and presentation:
- The recipient only sees the last level of discovering knowledge in databases. The recipient is assisted in understanding and interpreting the data extraction findings at this fundamental step using the visual technique.

3.5. THE STAGES OF DATA MINING

The term data mining is used as a synonym for Data from Discovery Knowledge or KDD, while others believe that data mining is only a stage of KDD [1]. These phases are shown in Figure 3.4.

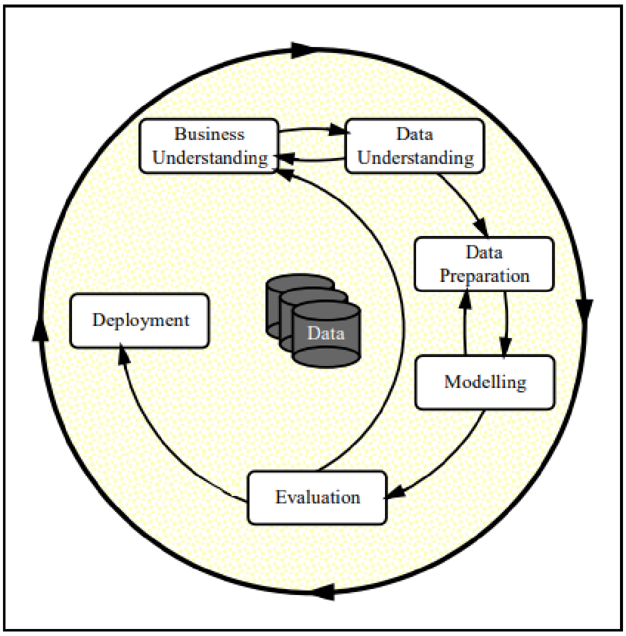

Figure 3.3. Illustrate the phases of KDD in discovering knowledge [1].

3.5.1. Business Understanding Phase

Understanding the nature of the business is the first step toward understanding the problems and issues that it faces. Simply, it's how to maximize the value of data mining, which need a single formula for specific business goals [11][28].

3.5.2. Data Understanding Phase

Understanding the data and its nature is essential for data mining and knowledge discovery. Additionally, having a thorough understanding of the data make the designers able use the algorithms or tools required for solving particular problems successfully [11]. This increases the likelihood of success and improves the knowledge discovery system's effectiveness and efficiency. Although it is not necessary for the data mining process to gather data in the data warehouse, it is recommended to avoid directly monopolizing the warehouse for data mining if it is housed within the enterprise.

The following steps required for data understanding process:

- Data collection: This establishes the source of the data used in the research, which may involve exterior data that is accessible such as tax information and other sources.
- Data description: concentrate on characterizing of the data. A single file's contents are either files or tables.
- Data Quality and Verification:
  As a great model demands excellent data, the data must be correct and clear, therefore this decides if some useless or low-quality data which may not be used in the research need to be reduced or dismissed.
- Exploratory Data Analysis: Preliminary data analysis is done using techniques like visualization, visualization, or using the Direct Assessment Procedure (OLAP). This stage is important since it focuses on creating theories relevant to the research subject.

3.5.3. Data Preparation Phase

In this phase of intensive work in where raw data are transformed into a complete data collection that will be used in all succeeding phases, where an appropriate records and variables for exploration are chosen, and certain variables are carried out if necessary and transformations are also performed to prepare the data in order to be ready for analysis tools. It consists of the following phases:

- Selection: meaning that selecting the expected variables and sample size.
- Build and Transform Variables: To create efficient models, new variables must constantly be created.
- Data Integration: Because data are grouped in an exploration study, in order to combine data from many databases for different purposes, multipurpose databases may be used.
- Data Formating: The goal of this stage is to rearrange the data so that it meets the requirements of the data mining form.

3.5.4. Modeling Phase

At this stage, appropriate modeling techniques are identified and applied, as multiple techniques can often be used to solve the same problem. The model settings are then adjusted to improve the results. Return to the data preparation stage if necessary to modify the data format to meet the specific requirements of the data mining technique. There are two types of models in data mining, either descriptive or predictive, as shown in Figure 3.4.

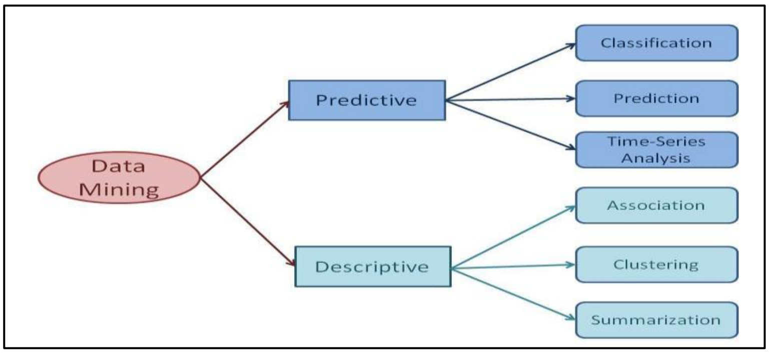

Figure 3.4. Data mining modeling type.

Different modeling techniques are chosen and applied at this stage, and their parameters are determined against optimal values. Usually, many techniques are used for a single problem in data mining. Some techniques have specific data-shaping requirements [11]. Therefore, a return to the data preparation stage is often necessary. Some famous algorithms are Decision Tree algorithm (DT), Random Forest algorithm (RF), K-Nearest Neighbors algorithm (KNN), Naïve Bayes algorithm (NB), and logistical regression algorithm (LR).

3.5.5. Evaluation Phase

At this stage, the quality and effectiveness of the models developed by the modeling stage are assessed before they are published and used, as well as whether the model actually achieves the goals defined for it in the first stage, and whether some critical aspects in the stage of understanding the work or the research problem has not been sufficiently considered, and then make a decision based on the results of data mining [11].

3.5.6. Deployment Phase

The development of the model is typically not the project's conclusion instead, the knowledge acquired must be presented in a way that enable client to perform. According to the objectives, the deployment step might be as simple as creating an overview or as complicated as creating an iterative data mining process. The procedure for deploying is often carried out by the user rather than the data analyst. In any event, it's critical to comprehend the procedures to follow in order to take advantage of the produced [28].

3.6. DATA MINING TECHNIQUES

Data mining techniques consist of several methods to extract the knowledge hidden behind many data and information and the ability to predict and make decisions, to help us solve many problems in different fields. Classification, estimation, and prediction techniques are examples of direct data mining, as the goal is to obtain the class value. At the same time, association, clustering, and description rules are indirect techniques whose purpose is to reveal the structure of the data.

3.6.1. Artificial Neural Network

They are not linear, training-based prediction models. Despite being effective methodologies for predictive modeling, part of the modeling power decreases their usability and adoption. Recognizing fraud and activities that may be considered fraud is one field that inspectors may quickly employ while analyzing data. Given its complexity, its power is most evident in cases where the model is used frequently, such as reviewing credit card transactions every month to check for anomalies.

3.6.2. Decision Trees (DT)

There are two different kinds of decision tree assessments, among those most popular visually data mining approaches. First is used for classification operations,. Classification is based on logic and uses a variety of conditions until all relevant data has been identified. The second type is used for regression operations, which is used when the target decision is numeric the, figure 3.5 shows the decision tree technique [17].

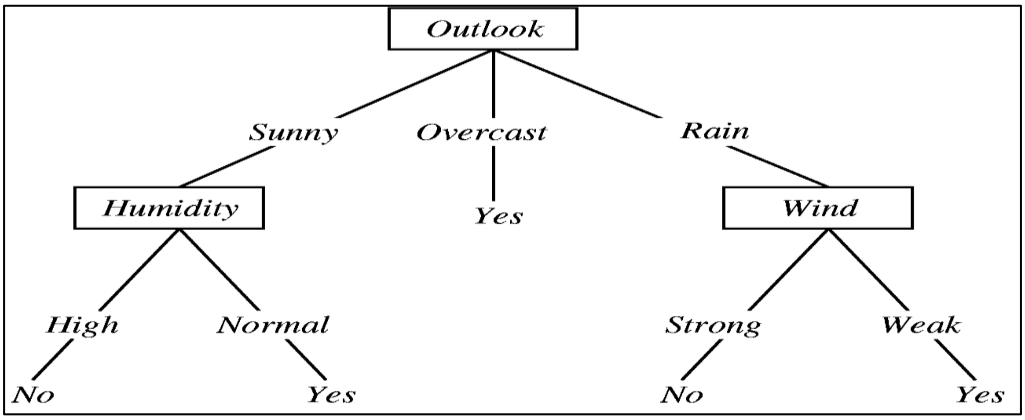

Figure 3.5. Decision Tree Technique

3.6.3. Classification

A machine learning algorithm is trained to categorize data using the data mining approach of classification. The categorization determines the class using statistical techniques like decision trees and the closest neighbor. The algorithm is preprogrammed with well-known data classifications in each of these techniques to make educated guesses about the new data item's quality. For instance, researchers may use photos of apples and mangoes to train a data-mining engine. The program can rather accurately identify if a fresh image is an apple, mango, or other fruit. [28].

3.6.4. Estimation

It is similar to classification, except that the main variable is mathematical rather than categorical. [34].

3.6.5. Prediction

It is the same as classification and estimation, except that it is done to predict the results while keeping the time factor in mind. For example, forecast a stock's price three months ahead [35].

3.6.6. Clustering

The process of clustering involves separating data into small groups (clusters) depending on details about the objects and the connections between them. The aim is for Entities in the same cluster to be similar while distinct from Entities in other clusters. Clustering differs from classification in that it does not rely on a known class; instead, records are grouped based on similarity [31].The diagram in Figure 3.6 shows an example of the assembly process. In this example, a sample of sales data compares customer age to sale volume.

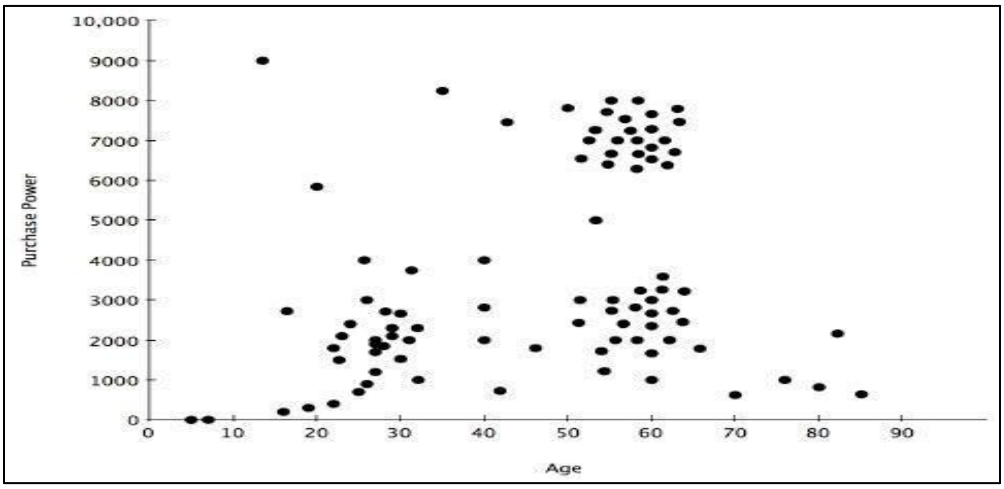

Figure 3.6. Using clustering technology to mine the data

In this example, we may split the population into two categories: those aged 20–30 ($2000) and those aged 50–65 ($7000–8000).

Individual consumers can be identified based on their closeness to one another on the chart. Customers in the same group are likely to share additional characteristics, and this prediction may assist in guiding, categorizing, and analyzing other individuals in the dataset.

3.6.7. Association

Finding relationships between two different and unrelated datasets is the goal of mining with association rules. Statements that are conditional suggest that there could be a connection among two pieces of data. Data scientists use support and confidence metrics to assess the correctness of outcomes [30]. Confidence demonstrates how often a conditional assertion is true, whereas support indicates the frequency with which related items exist in the collection. For instance, after purchasing one item, a consumer often purchases a similar item. Retailers may estimate the interests of potential customers by correlation mining prior purchase data. Retailers employ the findings from data mining to fill up the suggested parts of online shops.

3.6.8. Regression Analysis

Regression is a data mining technique used to predict numerical values (continuous values). For example, a regression can be used to predict the cost of a product or service, given other variables [20]. The regression task is to find a function that models the data with the fewest errors. A straight line can be drawn to show how each variable relates to the others. Suppose a business aims to make a forecast based on the effect of one variable on others. In that case, it is used in many business planning and marketing areas, including financial forecasting, environmental modeling, and trend analysis.

3.7. DATA MINING APPLICATIONS

Due to the tremendous expansion in data, particularly in databases and data warehouses, and the intense rivalry in the market, which forces companies to change their systems and engage in data mining, data mining has great impacts and its applications have grown in commercial firms.

Beginning with logistics corporations, data mining swiftly spread to banks, insurance companies, telecom companies, water and energy utilities, and most recently, air and rail transport industries, among others. And its first uses were in the area of managing customer services by examining consumer behavior in order to connect with them, win their loyalty, and provide them with goods depending on their needs [31].

In general, data mining may be used in a number of fields in commercial organizations, including:

- Marketing: Target market analysis has used artificial neural networks, including market shares. These marketing strategies have aided in segmenting clients based on basic data such as gender, age, group membership, and purchase habits.

- Retail: Data mining techniques have been utilized successfully in sales forecasting. Many variables have been used in studies, such as different business variables and customers' capacities based on purchasing behaviors. Methods such as purchasing basket analysis and market basket also assisted in determining which goods customers can purchase together.
- Banks: Forecasting business and financial outcomes has shown to be an efficient approach to data mining applications. These methods have been utilized to determine assured prices, forecast future prices, and analyze stock performance. These methods have also proven successful in developing computer-based measurement systems to assess the dangers of loans and financial crime.
- Insurance: Data mining approaches are also widely used in the insurance sector, especially in allocating client groups in order to calculate rates and variations in predictions of future claims, as well as to detect bogus claims.
- Communications: To limit the dangers of network infiltration and the volume of information, it deployed data mining approaches such as neural networks.

3.8. SUMMARY

In a nutshell, this chapter describes the important progress of data mining to tie everything together thus far. After that, the meaning of the exploration was discussed in the data, as well as the stages and activities. After that, we presented the applications in which the effectiveness of data mining science appears largely and successfully.

PART 4

METHODOLOGY

The research strategy used to carry out the study is described in this section. A full scientific justification for the designed model used in the research must be provided, along with answers to the questions raised in chapter one's introduction. Along with implementation methods, the methodology and processes necessary to achieve the primary goal of building and constructing a model that analyzes conventional mining data are also presented. The procedure and actions used to carry out the objectives shown in Figure 4.1, together with the final depiction of how the system is validated, recorded, and assessed for performance.

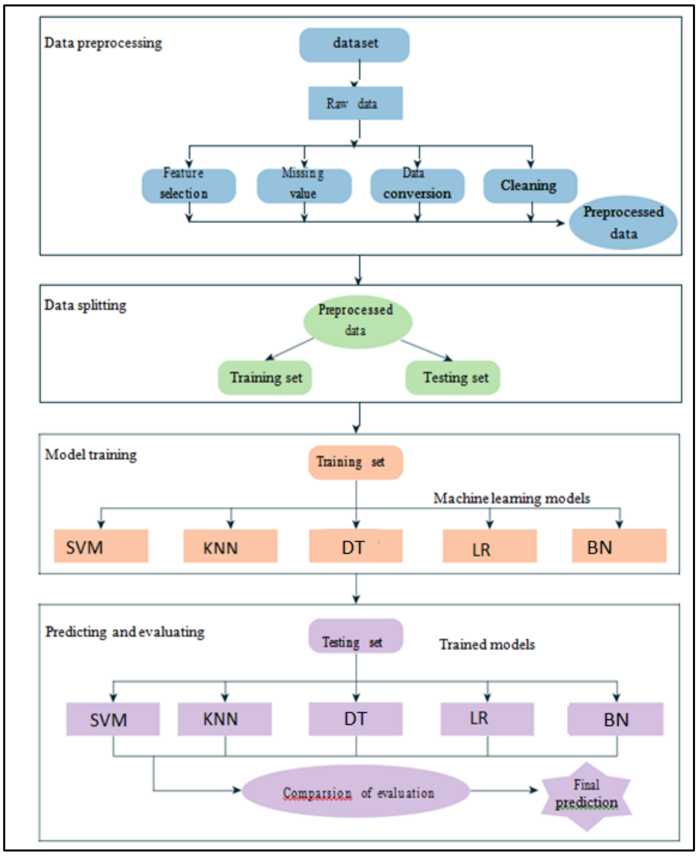

Figure 4.1. Schematic Diagram of the Proposed Method

4.1. DATA PREPROCESSING

For each algorithm, data must be presented in a specific format. Data collection comes first. Converting attributes is another step performed during pre-processing and replacing all missing values or incomplete data. Raw data can be stored in several formats, including text Excel and CSV.

4.2. MODEL TRAINING AND TESTING

We train and test the model using five machine-learning methods to get the maximum possible accuracy. The training and testing data were prepared for these algorithms. The algorithms used in the training and testing processes are as follows:

4.2.1. Logistic Regression (LR)

One of the most important statistical models for estimating the likelihood of a certain class or event, like success or failure, is the LR model. Multiple predicted variables—some of which may be categorical or numeric—are presented by LR. If the picture comprises a cat, tiger, fish, or other animal, for example, it may be modified to depict other classes of other occurrences. Each detected object in the image will be given a probability between Zero - 1, such that the total equals one. Other names, such as the Logit model or general Entropy classified. LR falls under supervised machine learning algorithms assigned to “classification” tasks. The following mathematical equation describes the algorithm. [20].

$$h\theta(x) \ = \frac{1}{1 + e^{-(\beta 0 + \beta 1x)}} \tag{4.1}$$

4.2.2. Decision Tree (DT)

A DT enables each node to contrast potential courses of action based on their advantages, disadvantages, and probabilities. In general, it saves the collection of potential outcomes for connected decisions. A DT often starts with only one node and branches into possible results. Every one of these findings causes more nodes to be divided into different states. As a result, a tree-like form emerges. Take into account a binary tree where the parent node is split into the right child and the left child nodes. The left child parent node and the right child contain Pd data (parent data), LCd data (left child data), and RCd data (right child data), respectively. Given the characteristics x, the impurity measures I, how many samples there are in the parent node (Pn) the quantity of samples taken from the left kid (LCn) and the right child (RCN); Maximizing the information acquired in the following equation is the objective of (DT) [17].

$$Pd, x = I(Pd) - \frac{LCn}{Pn} I(LCd) - \frac{RCn}{Pn} I(RCd) \tag{4.2}$$

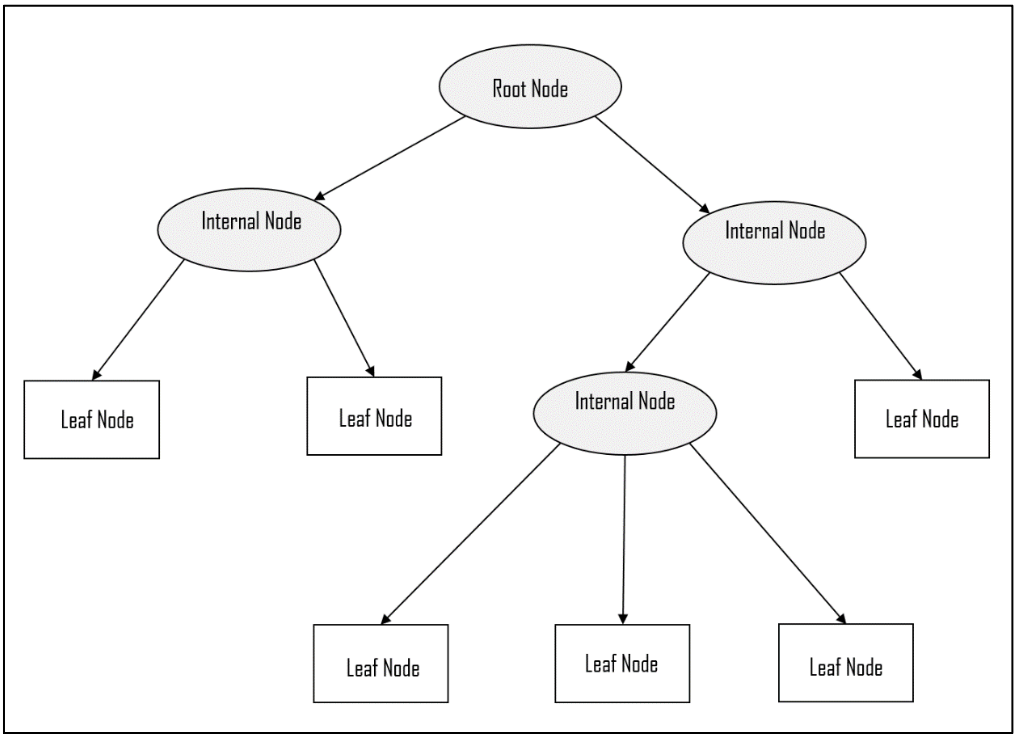

Figure 4.2. Showing the Structure and Functioning of a Decision Tree.

4.2.3. Support Vector Machine (SVM)

The most typical supervised approach separates the data into two groups, looks for points that are closest to an element, and uses those points as support vectors. The margin is the measurement of how near the dividing line is to the support vectors.

The algorithm looks for the maximum margin to produce the best line possible between the two groups. To do this, determine how far apart the two points are from the dividing line perpendicularly. [10]. The data points that make up SVM are those that are nearest to the top plane and could cause the super plane, which splits the data, to move if they were removed from the data set. These details might be regarded as significant data set components. Support vector machines are used for text classification tasks such as topic classification, spam recognition, and sentiment analysis. It is also used in image recognition challenges, particularly in classification based on colors or properties. It also plays an important role in handwritten number recognition areas such as postal service automation.

4.2.4. Naive Bayes (NB)

A simplified machine learning algorithm adds an independence condition to Bayes' theorem. NB can learn how to scan a set of well-categorized test documents, generate a list of words and their appearance and compare the content in all categories. [19].
Strengths of the Bayesian Algorithm:

- Strong base algorithm in the case of distorted and incoherent data.
- Deal with missing values by ignoring during the probability estimation process.
- Other technologies can be used such as Networks Belief

The following equation expresses the Bayes equation:

$$P(C|D) = \frac{P(D|C) * P(C)}{P(D)} \quad (4.3)$$

- D is data
- C is a class

The basic idea of the NB algorithm is to classify unseen (or unclassified) states into their nearest states within a given volume. Which helps in finding solutions to new problems by observing previous problems that have been solved, and it can also be relied upon to predict the existence of fundamental errors through a group, and he expresses that by saying that he has 70% confidence, for example, of a certain value that he explored, and the degree of confidence is determined Based on the following

- The distance between the explorer log and the nearest neighbor.
- How homogeneous the neighborhood group is and whether it leads to the same value explored.

4.2.5. K-Nearest Neighbor (KNN)

It is considered one of the methods that aim to predict by comparing records similar to the record to be predicted and estimating the unknown value of this record based on the information of those records. It is a method that depends on distance measures, paying attention to its quick, simple, and easy-to-apply aspects [5]. It is based on the assumption that samples in one data set will be close to samples with similar characteristics in another data set. The principle behind (K)is nearest neighbors. Here, the feature space data points nearest to our new data point are referred to as its closest neighbors. In addition, K (is the number of data points that we consider while using the approach. As a consequence, while using the KNN approach, the amount of K and the measure of distance are both very important. The most used unit of measurement is the Euclidean distance. Use the Minkowski, Manhattan, and Hamming distances as necessary. To forecast a continuous class or value for a new data point, it takes into account all the training data points. Find the "K" closest neighbors of the new data point(s) from the space specific to the decal categories or continuous magnitudes.

For sorting, the majority of KNN from the trained dataset are labeled with the expected group for the new data point. Is a continuous magnitude predicted for our new data point using the training data set's K nearest neighbors' mean continuous magnitudes and the regression mean.

4.2. PERFORMANCE MATRICES FOR EVALUATION

The data miner generates several data mining models, and needs to determine the best model to present to the data owner. The owner of the data uses his experience and knowledge in the field to evaluate the model, and the data miners can do the same, but since their knowledge is not as deep as the user, they had to carry out the evaluation process by using some measures.

There are several metrics to evaluate classification performance, including Kappa statistic, absolute error rate, root mean square error, classification accuracy, correct classification ratio, Recall, Measure-F, and ROC. By explaining the different classification cases as shown in the Table 4.1, which is the same as the Matrix Confusion matrix that will be used in calculating the value of some rating performance metrics.

Table 4.1. Different classification cases when predicting the values of two classes.

actual values	**Predicted Values**	
	+	-
+	The correct positive state	false negative case
-	The false positive case	The false positive case

The confusion matrix provides the necessary information to determine the quality of the performance of the classification model, and this task can be accomplished using a number of measures, including:

4.2.1. Accuracy

It is the proportion of correctly classified records. Most classification algorithms look for models that achieve the highest accuracy [26], and are calculated as follows:

$$ACCURACY = \frac{TN + TP}{T + F} \tag{4.4}$$

- TP is true positive
- TN is true negative

- T is total true (true positive and true negative together
- F is total of error (false positive and false negative).

4.2.2. Precision

It measures the proportion of records in the set which the classifier reported to be positive and which are really positive [25], and the more thorough the inquiry, the fewer false positives the classifier made. It is computed as following:

$$\mathrm{PRECISION} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \qquad (4.5)$$

4.2.3. Recall

It measures the proportion of positive examples correctly predicted by the classifier. If a classifier has a high recall value, it means that there are very few positive examples that are incorrectly classified as being of the negative class. The call value is equivalent to the positive true rate [13]. It is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \qquad (4.6)$$

It should be noted that building a model that maximizes both recall and inquiry measures is a fundamental challenge facing classification algorithms.

4.2.4. F-Measure

The F1 is represented mathematically by equation as a weighted median of a model's accuracy and recall.

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \qquad (4.7)$$

The F1 Score value displays the harmonic average of the recall and precision ratings. It is a mean that is harmonic instead of a simple mean while we shouldn't ignore the extreme circumstances. If the F1 Score were determined by a straightforward average calculation, a model with a Precision value of 1 and a Recall value of 0 would result in a deceptive F1 Score of 0. [23]. Choosing the incorrect model in data sets with uneven distribution is avoided by using the F1score value instead of accuracy. Additionally, the F1score is crucial to us since we want a measuring tool that will account for all mistake costs in addition to false positives and false negatives. [24].

4.3. SOFTWARE REQUIREMENTS

Python is very popular in data science and working with big data. It is one of the best languages used in data science. Depending on libraries specialized in dealing with data, such as NumPy, Pandas, matplotlib, SciPy, and others, all these libraries and many others deal with helping the data scientist get tangible results in reality, which makes Python the primary language in analysis and data science. Additionally, the subject of artificial intelligence is vast, encompassing several other topics such as deep learning, machine learning, the Internet of Things, and many others... In this major, the Python language has come to the fore. Moreover, it has become the best and most powerful language used in the field of artificial intelligence as a whole. It offers many libraries, such as PyTorch, Theano, Pandas, and many others. Python is used heavily in the field of artificial intelligence and is almost the only one there [28].

4.3.1. Scikit-Learn Library

A Python-based machine learning library is called scikit-learn. In addition to its use in the data processing and model assessment phases, it incorporates a variety of algorithms and techniques used in machine learning, including classification, clustering, and regression. Scipy, Numpy, Matplotlib, and several more libraries were used in its construction.

4.3.2. Data Analysis Using Python

After the data has been cleaned up and given the necessary transformations, analysis can start. Through statistical operations, specialized algorithms, and data visualization techniques, information is then extracted.

4.3.2.1. The First Stage (Read Data)

We created a file and put the data for traditional mining in it, and we also created a file that takes the extension (.py) so that the Python interpreter can recognize it. To read the data and save it in the data frame, we use the (read_csv) function in the Pandas library, and to get a sample of the data after reading it, we use the head function of the data frame, which will return the first five records of the data. We can use the sample function to get a random sample. Scaling numerical features:

```
train_data = pd.read_csv("Train_data_mining.csv")
print(train_data.head(5))
```

Extract the numeric data, scale it so that the mean is zero and the variance is small, and then return the result to the data frame.

```
colms = train_data.select_dtypes(include=['float64','int64']).columns
sctrain = scaler.fit_transform(traindata.select_dtypes(include=['float64','int64']))
sc_traindf = pd.DataFrame(sctrain, columns = cols)
```

Choosing the models that will be used in the learning and testing processes

```
# K-Neighbors Model
KNN_Classifier = KNeighborsClassifier(n_jobs=-1)
KNN_Classifier.fit(Xtrain, Ytrain);
# Logistic Regression Model
LGR_Classifier = LogisticRegression(n_jobs=-1,max_iter=25000, random_state=0)
LGR_Classifier.fit(Xtrain, Ytrain);
# Naive Baye Model
BNB_Classifier = BernoulliNB()
BNB_Classifier.fit(Xtrain, Ytrain)
# Decision Tree Model
DTC_Classifier = tree.DecisionTreeClassifier(criterion='entropy', random_state=0)
DTC_Classifier.fit(Xtrain, Ytrain)
# SVC Model
SVC_Classifier = svm.SVC(gamma='auto')
SVC_Classifier.fit(Xtrain, Ytrain)
Models_list = []
Models_list.append(('Naive Baye Classifier', BNB_Classifier))
Models_list.append(('Decision Tree Classifier', DTC_Classifier))
Models_list.append(('Kneighbors Classifier', KNN_Classifier))
Models_list.append(('Logistic Regression', LGR_Classifier))
Models_list.append(('Support Vector Classification', SVC_classifier))
```

Coding for taxonomic features:

Extracting categorical features from the training and test sets and encoding the categorical features, separating the target column from the encoded data.

```
mining_train = train.select_dtypes(include=['object']).copy()
train_cat = mining_train.apply(encoder.fit_transform)
enc_train = traincat.drop(['class'], axis=1)
catogry_Ytrain = traincat[['class']].copy()
trainx = pd.concat([sc_traindf,enctrain],axis=1)
trainy = train['class']
trainx.shape
```

Evaluation of the model after the training process by taking the values from the inconsistency matrix and also using the F1 score, accuracy, precision, recall and AUC.

```
for m, fu in Models_list:
val_scores = cross_val_score(fu, Xtrain, Ytrain, cv=10)
model_accuracy = metrics.accuracy_score(Ytrain, fu.predict(Xtrain))
matrix = metrics.confusion_matrix(Ytrain, fu.predict(Xtrain))
class_report = metrics.classification_report(Ytrain, fu.predict(Xtrain))
print('Evaluation  - '.format(m))
print ("Cross Validation  Score:" "\n", scores)
print ("Cross Validation Mean Score:" "\n", scores.mean())
print ("Cross Validation Max Score:" "\n", scores.max())
print ("Cross Validation Min Score:" "\n", scores.min())
print ("Model Accuracy:" "\n", model_accuracy)
print("Confusion matrix:" "\n", matrix)
print("Classification report:" "\n", class_report)
```

Model validation is the process performed after training the model and is done using the test data.

```
for m, fu in Models_list:
model_accuracy = metrics.accuracy_score(Ytest, fu.predict(Xtest))
matrix = metrics.confusion_matrix(Ytest, fu.predict(Xtest))
class_report = metrics.classification_report(Ytest, fu.predict(Xtest))
print(' Test ======='.format(m))
print ("Model Accuracy:" "\n", model_accuracy)
print("Confusion matrix:" "\n", matrix)
print()
```

4.4. SUMMARY

This chapter covered the research activities in detail, focusing on locating useful business data for formal mining and explaining the model development and design processes. After this, it was mentioned how to train the proposed model and test the algorithms used in the model, which were highlighted. Finally, clarify how to evaluate the model to reach the desired results.

PART 5

EXPERIMENTAL RESULT

Spread on a large scale in several fields, which is the data mining method, in an attempt to study the possibility of benefiting from it in the development of traditional gold mining and stopping it in Sudan. It became clear to us the importance of this method and the possibility of benefiting from it, as it contains multiple tools. Each of them is a useful tool for analysis. Each of them plays a positive role in developing the process of the mining industry and increasing efficiency and effectiveness. Using the SVM, LR, NB, DT, and K-NN models on the training data, the models were evaluated for ratings with performance criteria prepared in advance. Due to the measure of the final performance of the test data set, a confusion matrix was created for all models used to measure efficiency, accuracy, precision, recall, F1 score, and AUC curve.

5.1. DATASET PREPARATION

The study is conducted on real data for gold exploration in Sudan. Gold exploration operations are conducted by Sudanese mineral resources companies in 6 gold-producing states in Sudan. In order to obtain gold, some basic resources are used, such as water, electricity, and human resources. Machinery is operated in mining operations, and some waste is also exported. A data set containing 20,000 company records was prepared, including the information that was referred to. Field controllers are the human resource components of mining companies. This data represents statistics on the volume of work, the number of workers, productivity, and the various tools in gold production. An Android application is used to collect data; the processes included five years procedures starting from 2017 as shown in Table 5.1. We are using the dataset with five classification algorithms. In order to build and test the model, the data is divided into two parts: 70% for training and 30% for testing. Training is performed for each module with the records in the training set.

Table 5.1. Datasets characteristics

Attribute	Value	Description	Type
owner_name	Ahmed Mohamed	Company owner	Nominal
type_gold_ext	Windmill – jeweler- watermill	Type of gold extraction	Nominal
State	river Nile – kassala – gdarif – north state	Description of the geographical location in the states	Nominal
company_name	Delgo Mining Company	Description of the company name	Nominal
number_year_mining	5- 10 -20 – 40	Describe the number of years in the mining sector	Numeric
number_gilogest	10 -20	Describe the number of geologists in the company	Numeric
number_laborer	50 – 60 -70	The number of workers in the mines	Numeric
Units	14 -18 -20	The number of machines working in the extraction of gold	Numeric
Kerta	2.4 – 5.6 -	The percentage of gold in the waste	Numeric
Production	10 – 30 -100	Description of gold productivity during a month	Numeric

5.1.1. Data Cleaning

It is the process of separating out data from the data collection that has noise or inclusions.

5.1.2. Data Integration

This step is often the origin of the computation of variable components and can be integrated in a single source that has been connected between the sources to guarantee the data's integrity.

5.1.3. Data Transformation

It is the procedure of converting the chosen data into a format that is suitable for search and retrieval operations, where the data has been transformed to *.csv for simplicity of handling.

5.2. MODEL PERFORMANCE ANALYSIS

In this section, we examine how well each classifier performs and how it can distinguish between companies that are actively producing high levels of output and those that have ceased to do so and are producing low levels of output. In the following table 5.2 we are listing the overall evaluation on training and testing data with all used algorithms.

Table 5.2. Overall evaluation and comparison of the five models.

Training	**Model**	**Precision**	**recall**	**f1-score**	**Accuracy**
	SVM	0.93	0.94	0.93	0.94
	LR	0.96	0.95	0.94	0.96
	NB	0.9	0.91	0.89	0.9
	DT	1	1	1	1
	KNN	0.72	0.71	0.71	0.72
Testing	Model	Precision	recall	f1-score	Accuracy
	SVM	0.57	0.58	0.55	0.58
	LR	0.91	0.92	0.91	0.91
	NB	0.82	0.82	0.82	0.82
	DT	0.99	0.98	0.99	0.98
	KNN	0.53	0.54	0.53	0.54

The accuracy of each model was calculated, the DT classifier had the best performance and accuracy between classifier as shown in Figure 5.6.

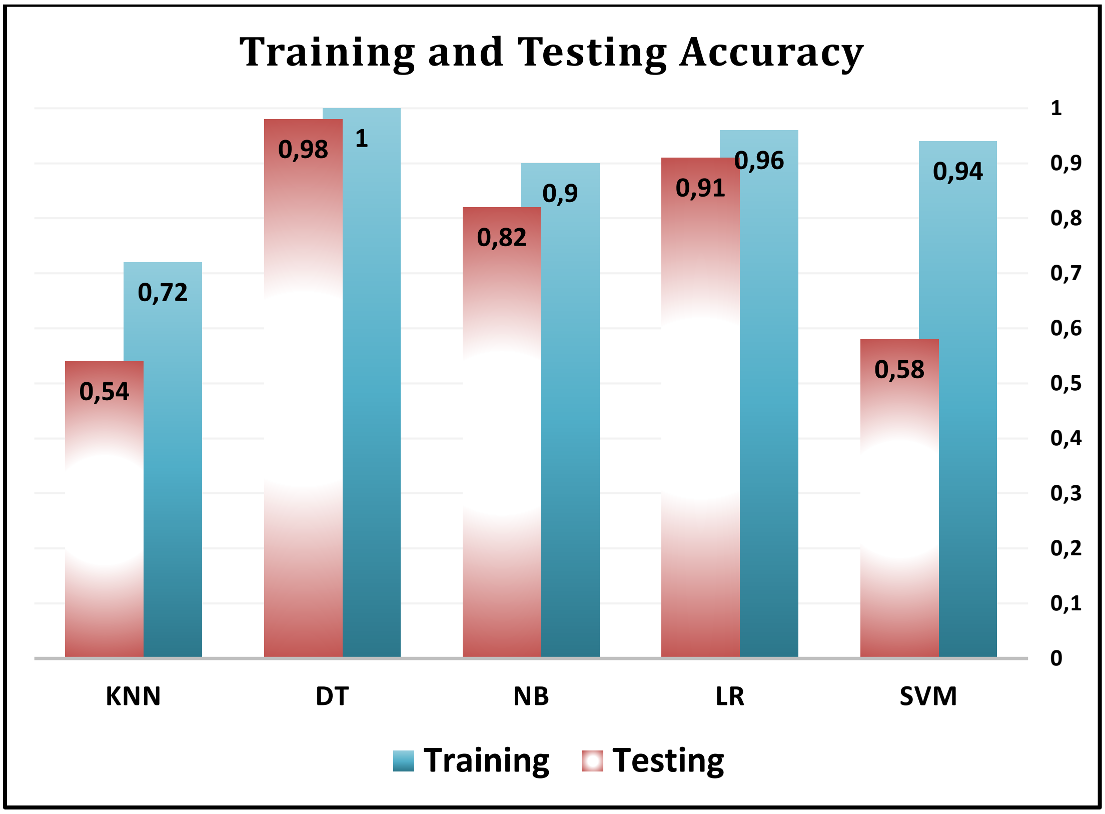

Figure 5.1. Compare accuracy of models in the classification process

The likelihood that the classifier will score a randomly chosen positive example greater than an unconsciously selected negative sample is represented by the area that lies of the curve (AUC). The classifier can discriminate across classes and the ROC curve summary using the area that is under the curve. The model is predicated to identify across positive and negative categories with the highest efficiency.

According to Figure 5.7, DT had the greatest accuracy score of 1.00, followed by LR and NB.

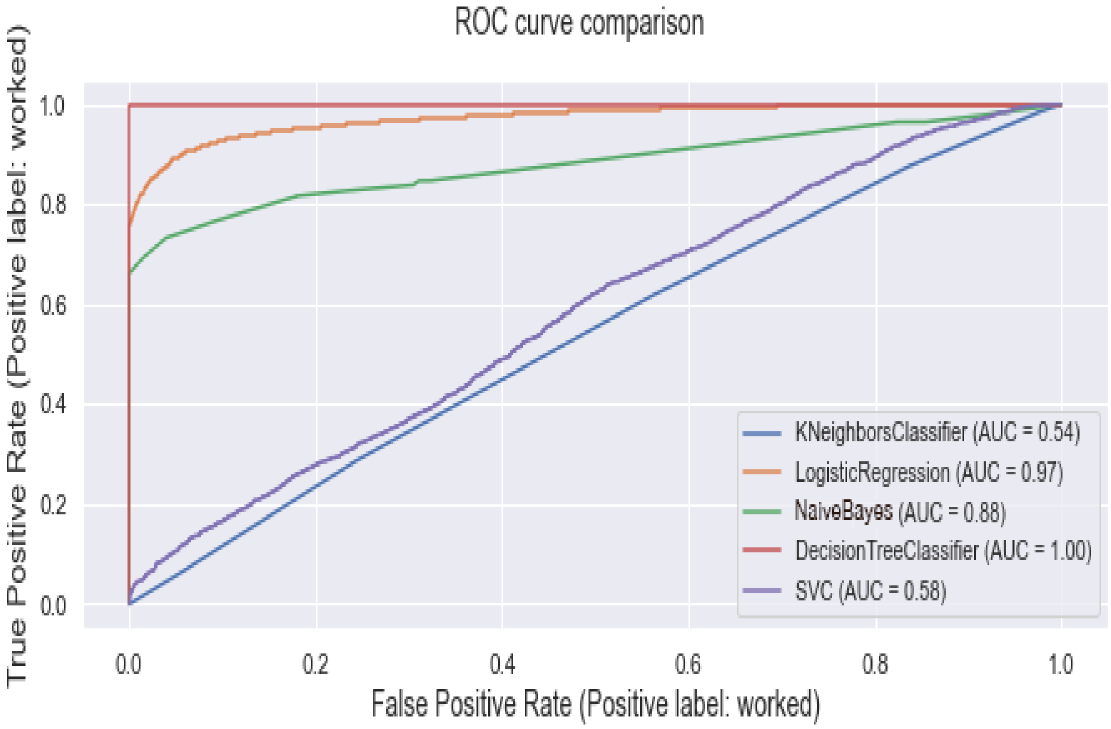

Figure 5.2. A curve showing the performance of each algorithm in the classification process.

As shown in Figure 5.7 that the curve of used machine learning models (SVM, NB, DT, SVM, and LR) classifiers, indicate that the DT classifier is proving the high AUC of classification about 99 % compared with other models, and the LR is proving a good AUC too about 97 %, while the other classifiers are less than 90 % of AUC.

5.2.1. DT Analysis

The DT model achieved high accuracy in classifying companies for each category. The accuracy for this model is shown in Figure 5.1. The accuracy of the model lies in the evaluation of the employed and unoccupied categories. An accuracy rate of 99% was achieved in classifying high-yield firms for 3,710 cases. A rate of 100% was achieved for 3010 accurate cases in the classification of companies with low productivity. Hence, it was the best model, we are using the default parameters of this algorithm. We are using "entropy" as the criterion value and random_state equal to 0.

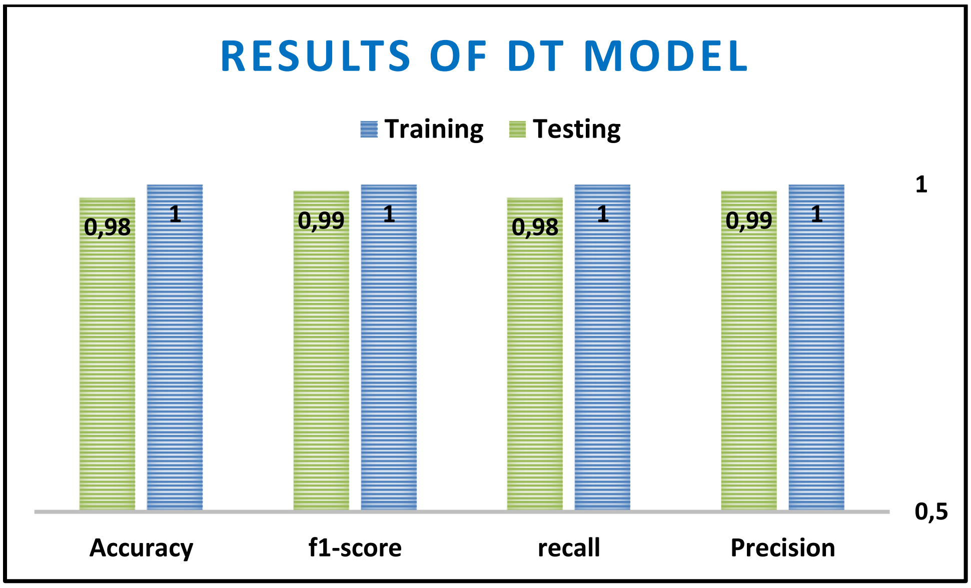

Figure 5.3. DT model for precision, recall, and F1 for test data.

Table 5.3. Parameters used in DT Model.

parameter	value
criterion	gini
splitter	best
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	None

5.2.2. LR Analysis

The LR model achieved the best accuracy of 0.91 in classifying companies. An average score of 0.92 was achieved for the 3,710 cases in the classification. A rate of 0.91 correct cases was achieved in the classification of companies with low productivity for 3010 cases. In Fig. 5.2, the LR model of accuracy, recall, and F1 score are shown for the dataset and companies cases. The training state of this algorithm included max_iter by 25000, the other parameters are used with the default values.

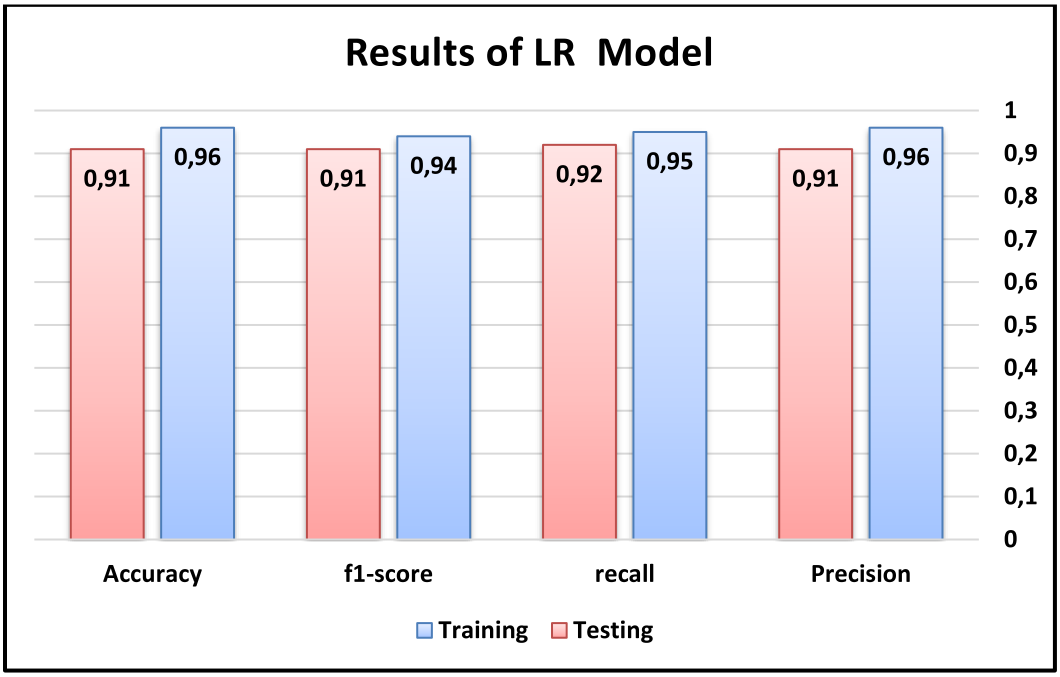

Figure 5.4. LR model for precision, recall, and F1 for test data.

Table 5.4. Parameters used in LR Model

parameter	value
penalty	‘l2’
dual	Treu
C	1.0
fit_intercept	True
max_iter	27000
multi_class	ovr
class_weight	‘balanced’

5.2.3. KNN Analysis

In the KNN model, the mean accuracy was low compared to the rest of the algorithms, with a rate of 0.54. The average accuracy of correct cases in the classification of operating companies was achieved by 0.60 for 3710 cases. The average number of correct cases in the classification of companies with low productivity was 0.46 for 3010 cases. In Figure 5.3, the accuracy, recall, and F1 of the KNN model are shown as the results with the dataset. The number of neighbors which is represented by (k) is set to be ten while training the model.

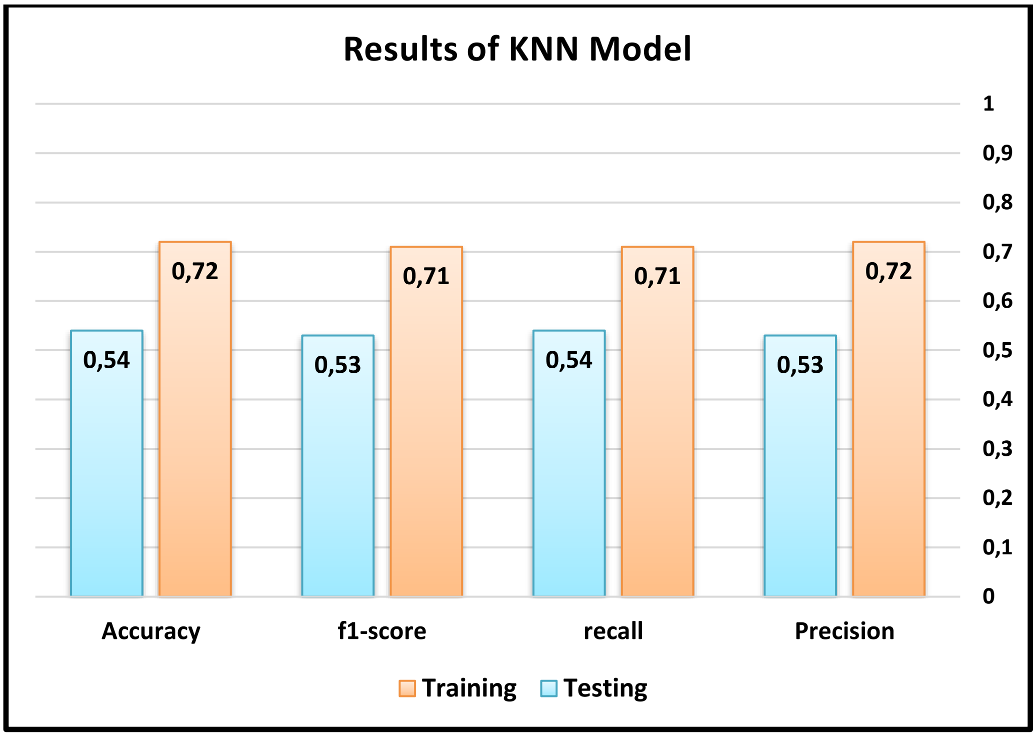

Figure 5.5. KNN model for precision, recall, and F1 for test data.

Table 5.5. Parameters used in KNN Model.

parameter	value
n_neighbors	10
weights	’uniform’
leaf_size	30
p	2
metric	’minkowski’

5.2.4. NB Analysis

The NB model achieved an average accuracy of 0.82 in classifying companies, which is better than KNN. An accuracy rate of 0.85 was achieved in classifying operating companies for 3710 cases. An average accuracy of 0.78 was also achieved in classifying companies with low productivity for 3010 cases. The accuracy, recall, and F1 values of the dataset are listed in Figure 5.4 for the NB model. The parameters of this algorithm used as the default states.

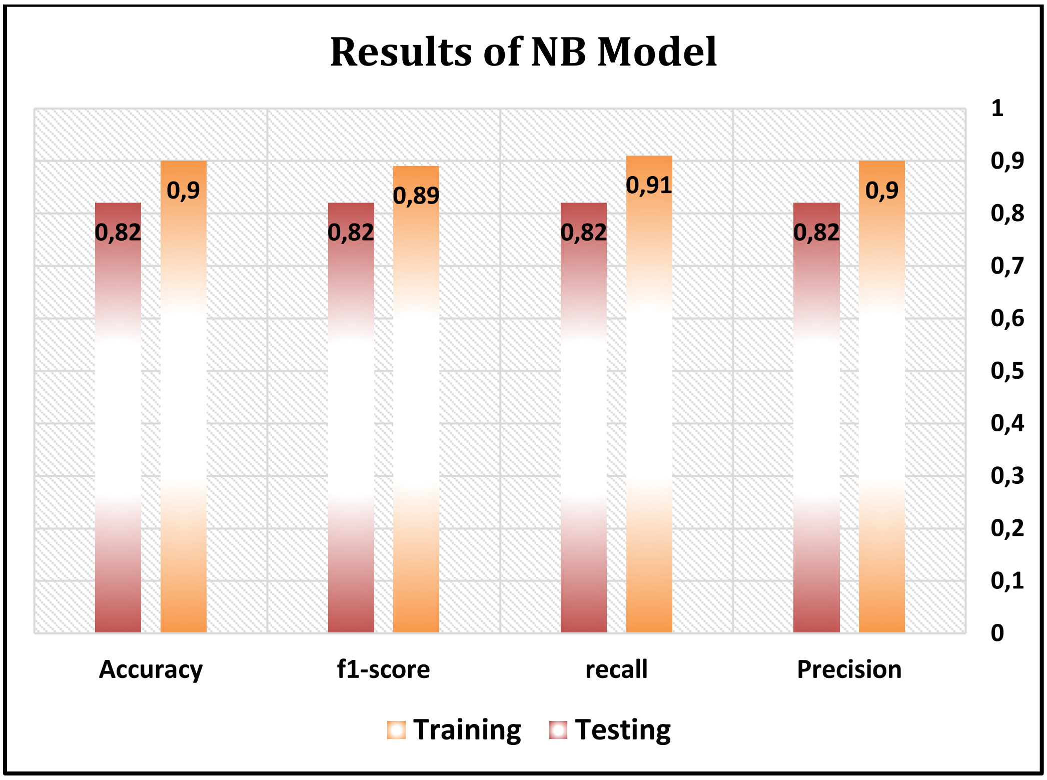

Figure 5.6. NB model for precision, recall, and F1 for test data.

Table 5.6. Parameters used in NB Model.

parameter	value
n_classes	None
var_smoothing	1e-9

5.2.5. SVM Analysis

The accuracy rate achieved in classification with the SVM model was as low as 0.58. The classification of operating companies was carried out with an accuracy of 0.68 for 3710 cases. Accuracy of classification of ISCs with low productivity was achieved at a rate of 0.38 per 3010 cases. The accuracy, recall, and F1 values for this model are shown in Figure 5.5. Gamma value of this algorithm is set to be auto in our model.

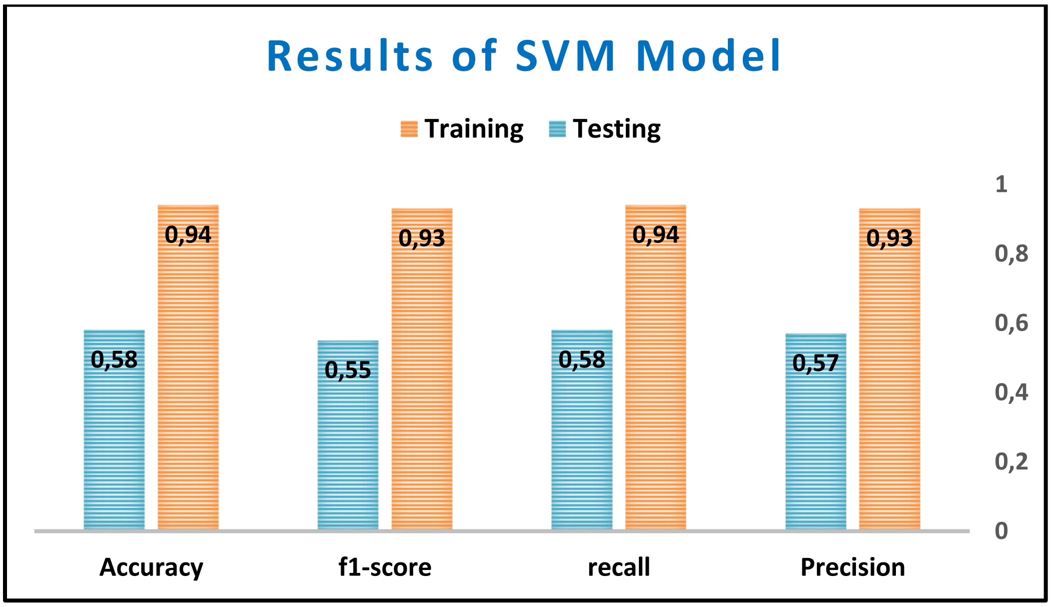

Figure 5.7. SVM model for precision, recall, and F1 for test data.

Table 5.7. Parameters used in SVM Model

parameter	value
C	1.0
Kernel	’rbf’
degree	3
gamma	’scale’
coef0	0.0
shrinking	True
probability	False
tol	1e-3
cache_size	200
verbose	False
decision_function_shape	’ovr’

5.3. DISCUSSION

According to our findings, the DT algorithm outperformed the other algorithms in terms of precision, recall, f1-score, and Accuracy, among other metrics. These study's pairings are examples of supervised learning. For the DT, LR, and NB classifiers, the models attained high levels of accuracy of 0.99, 0.91, and 0.81. In comparison to the other models, the SVM and K-NN models fell by 0.57 and 0.53, respectively. Our study had a lot of shortcomings, including the limited size of the data set and the choice of less precise, less rare algorithms in favor of more frequent ones. Despite these limitations, the study's findings helped legalize highly productive traditional mining companies, which in turn helped decide to convert to formal mining and supported them in their efforts to stop significant environmental pollution and increase production by completely ceasing traditional mining. This is a significant contribution to the mining industry in the Country. Finally, working with data mining algorithms on a huge data set is one of the pivotal bases of the endless search, so we aspire for more studies in this field. One of the fundamental foundations of the endless search.

PART 6

CONCLUSION AND FUTURE WORK

6.1. CONCLUSION

The wide spread of systems and tools and the accumulation of data stored in them, contributed to the development of a large number of important mathematical algorithms, which proved effective in extracting knowledge hidden in databases through data analysis and extraction of new models that were not known, or predicting future events by taking advantage of stored groups of data Known results beforehand. This study used SVM, LR, NB, DT and K-NN and we compared the performance accuracy rate and draw AUC, F1, Accuracy, recall and Accuracy to reveal the classification of traditional mining companies with high productivity and developed. According to Table 5.1, 70% of traditional mining companies have been trained and 30% have been validated and tested.

The DT model scored the highest about 99 %, followed by LR and NB, while the K-NN and SVM models came last. The results in the pre-trained models prove that the DT model can be used in the rationing problems of traditional mining companies.

6.2. FUTURE WORK

In the future, we propose to use other models to increase the accuracy of operating high-productivity companies using big data and at different time periods.

- Developing an integrated data warehouse to provide all of the information and data required by analysts to mine data and discover knowledge.
- Run another algorithm and compare the results to the findings of this study.

- The majority of government institutions store their data in the form of text data on computers in an unorganized manner usually, it is possible to benefit from the model's proposal and develop it in order to classify these documents, reduce their repetition, and benefit from them rather than storing them in vain.
- Conduct research in the field of cloud storage and examine its prospects in the field of knowledge discovery.

REFERENCES

1. De Martino, M., Bertone, A., Albertoni, R., Hauska, H., Demšar, U., & Dunkars, M. (2002). Information Visualisation for Site Planning.

2. "Mining in Sudan-Overview". Mbendi.com. Archived from the original on 29 January 2001. Retrieved 13 June 2015.

3. Akinsola Adeniyi F., Sokunbi M.A., Lawal.O.N., Okikiola F.M-(2015), " A Data Mining Approach To Insurance Risk Analysis. " International Journal of Engineering and Computer Science ISSN: 2319-7242, 5(5), 10255-10258

4. Baradwaj, B. K., & Pal, S. (2011). Mining Educational Data to Analyze Students' Performance. International Journal of Advanced Computer Science and Applications, 2(6), 63-69

5. Batista, Gustavo EAPA, and Maria Carolina Monard. "A study of K-nearest neighbour as an imputation method." His 87.251-260 (2002): 48.

6. Vu, Duy-Hien. "Privacy-preserving Naive Bayes classification in semi-fully distributed data model." Computers & Security 115 (2022): 102630.

7. Lyras, Dimitrios P., et al. "Educational software evaluation: A study from an educational data mining perspective." The International Journal of Multimedia & Its Applications 6.3 (2014): 1.

8. Reza Safdari, Amir Deghatipour, Marsa Gholamzadeh, Keivan Maghooli, Applying data mining techniques to classify patients with suspected hepatitis C virus infection, Intelligent Medicine, 2022, ISSN 2667-1026,

9. Mohamed, Amr E. "Comparative study of four supervised machine learning techniques for classification." International Journal of Applied 7.2 (2017): 1-15.

10. Antonio Comi, Antonio Polimeni, Chiara Balsamo, Road Accident Analysis with Data Mining Approach: evidence from Rome, Transportation Research Procedia,Volume 62, 2022, Pages 798-805,ISSN 2352-1465,https://doi.org/10.1016/j.trpro.2022.02.099.

11. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

12. Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts and Techniques (3rd ed.). Waltham, United States of America: Morgan Kaufmann.

13. Stojadinovic, Uros, et al. "Structure and provenance of Late Cretaceous–Miocene sediments located near the NE Dinarides margin: Inferences from kinematics of orogenic building and subsequent extensional collapse." Tectonophysics 710 (2017): 184-204.

14. Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6 (2020). https://doi.org/10.1186/s12864-019-6413-7

15. Juba, Brendan, and Hai S. Le. "Precision-recall versus accuracy and the role of large data sets." Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.

16. Gunawardana, Asela, and Guy Shani. "A survey of accuracy evaluation metrics of recommendation tasks." Journal of Machine Learning Research 10.12 (2009).

17. Junker, Markus, Rainer Hoch, and Andreas Dengel. "On the evaluation of document analysis components by recall, precision, and accuracy." Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318). IEEE, 1999.

18. Komarek, Paul, and Andrew W. Moore. "Making logistic regression a core data mining tool with tr-irls." Fifth IEEE International Conference on Data Mining (ICDM'05). IEEE, 2005.

19. Hernández, Víctor Adrián Sosa, et al. "A practical tutorial for decision tree induction: Evaluation measures for candidate splits and opportunities." ACM Computing Surveys (CSUR) 54.1 (2021): 1-38.

20. Batista, Gustavo EAPA, and Maria Carolina Monard. "A study of K-nearest neighbour as an imputation method." His 87.251-260 (2002): 48.

21. Ting, S. L., W. H. Ip, and Albert HC Tsang. "Is Naive Bayes a good classifier for document classification." International Journal of Software Engineering and Its Applications 5.3 (2011): 37-46.

22. Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

23. Asadi H, Dowling R, Yan B, et al. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. PLoS One 2014;9(2):e88225. doi:10.1371/journal.pone.0088225.

24. Deng X, Liu Q, Deng Y, et al. An improved method to construct basic probability assignment based on the confusion matrix for classification problem. Inf Sci 2016;340:250–61 NY. doi:10.1016/j.ins.2021.11.018

25. Chicco, D. Jurman, G. (2020). "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification

evaluation", BMC Genomics. Published: 02 January 2020 on https://link.springer.com

26. LAROSE D_ (2005), "Discovering Knowledge in Data an Introduction to Data Mining", John Wiley & Sons. Inc., Canada, 222.

27. D Brzezinski, J. Stefanowski Prequential AUC: properties of the area under the ROC curve for data streams with concept drift Knowl Inf Syst, 52 (2) (2017), pp. 531-562, 10.1007/s10115-017-1022-8

28. Schilling, Melissa A., et al. "Learning by doing something else: Variation, relatedness, and the learning curve." Management science 49.1 (2003): 39-56.

29. Mehmed, k,2011- Data mining: Concepts, Models, and Algorithms, Indian, 234P.

30. BARBE,D,2010-Bayesian Reasoning and Machine Learning. First Edition, Cambridge University Press, London, England, 610p.

31. Soukup, T., & Davidson, I. (2002). Visual data mining: Techniques and tools for data visualization and mining. John Wiley & Sons.

32. Akinsola Adeniyi F., Sokunbi M.A., Lawal.O.N., Okikiola F.M-(2015), " A Data Mining Approach To Insurance Risk Analysis. " International Journal of Engineering and Computer Science ISSN: 2319-7242, 5(5), 10255-10258.

33. Azevedo A, Santos M F-(2008)"KDD, SEMMA and CRISP-DM: a Parallel Overview", International Association for Development of the Information Society IADIS, ISBN: 978- 972-8924-63-8, 182-185

34. Bhowmik R-(2011), "Detecting Auto Insurance Fraud by Data Mining Techniques." Journal of Emerging Trends in Computing and Information Sciences ‹2(4), 156-162

35. Chai t., Daraxler r-(2014)," Root mean square error (RMSE) or mean absolute error (MAE)?" Geoscientific. Model Dev., 7, 1247–1250.

36. Easa S, Hasan M, Hamad M – (2005), "Traffic Collision Analysis Models: Review and Empirical Evaluation", Arab Journal of Administrative Science, University of Kuwait, 12(3) ‹473-497

37. "2012 Minerals Yearbook: Sudan [Advance Release]" (PDF). U.S. Department of the Interior: U.S. Geological Survey. March 2014. Retrieved 13 June 2015.

38. "The Field of Mineral Potential of the Sudan". Government of Sudan. Archived from the original (PDF) on 4 March 2016. Retrieved 13 June 2015.

APPENDIX A.

LARGER VIEWS OF MSTs

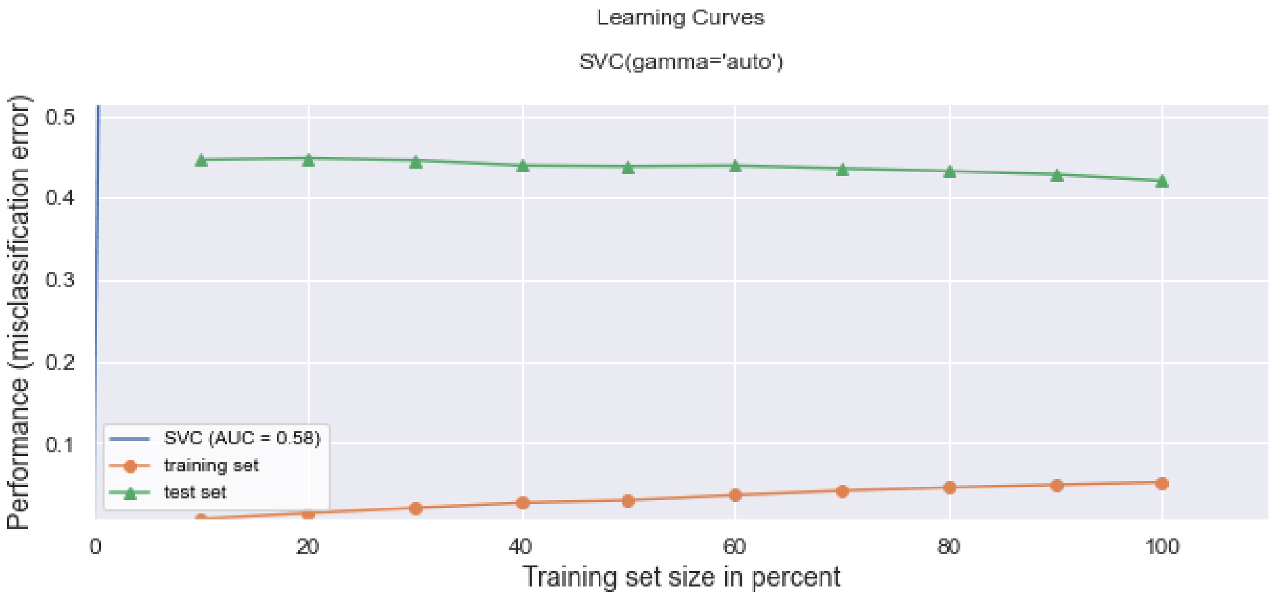

Figure Appendix A.1. Learning curve from SVM model

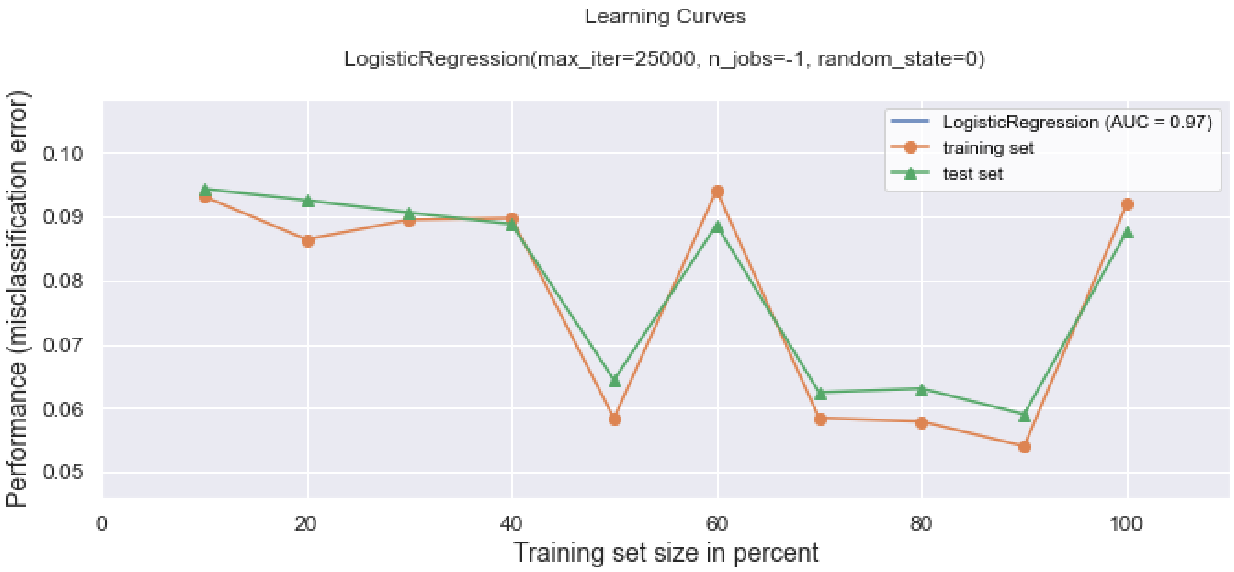

Figure Appendix A.2. Learning curve from LR model

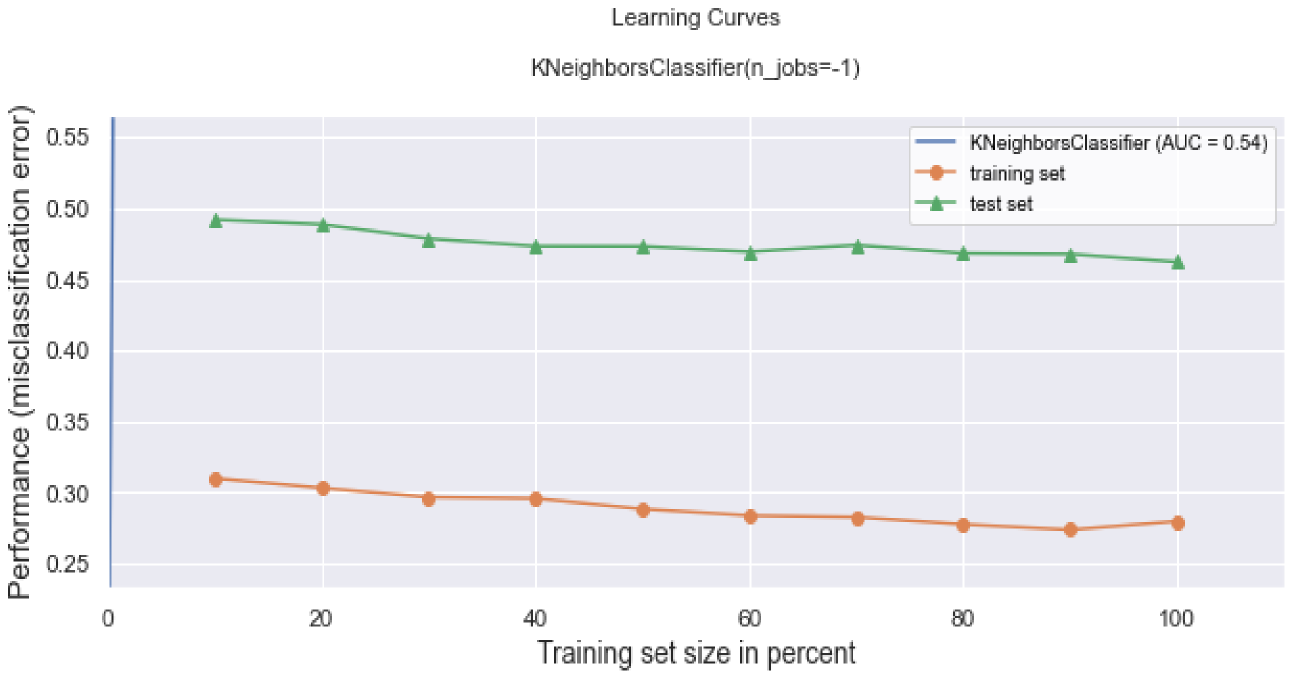

Figure Appendix A.3. Learning curve from KNN model

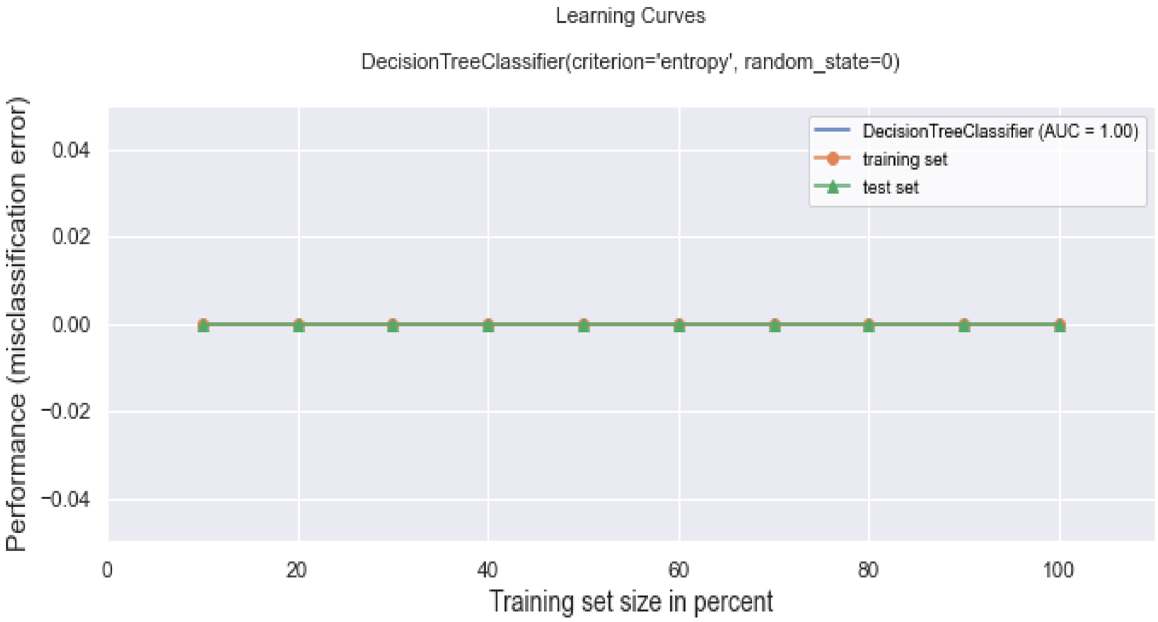

Figure Appendix A.4. Learning curve from DT model

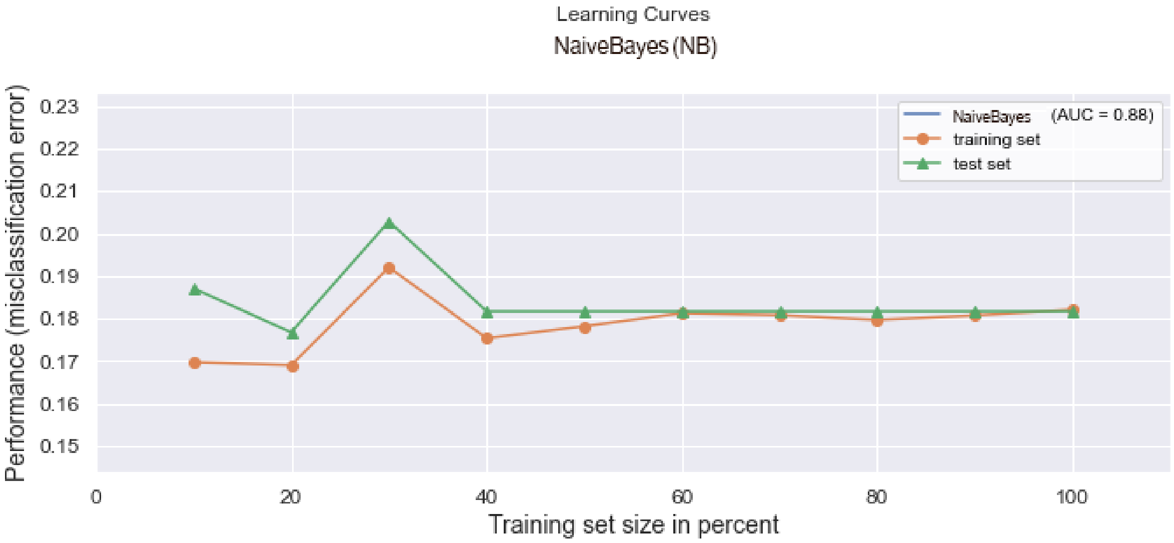

Figure Appendix A.5. Learning curve from NB model

RESUME

Al-Amin Salah EL-DIN completed his primary, preparatory and secondary education in Khartoum schools, Sudan, and obtained a Bachelor's degree in Computer Science from Omdurman Islamic University - Sudan in 2014.

After graduation, I worked as an information systems developer in the Ministry of Investment for two years, after which I moved to work in the Ministry of Minerals in the Information Technology Department, and I am still working now. During that period, I took a number of training courses in the field of computer science. Then, in 2019, he moved to Turkey to study at Karabuk University to obtain a master's degree in computer engineering.