T.C.
KIRSEHİR AHI EVRAN UNIVERSITY
INSTITUTE OF SCIENCE
ADVANCED TECHNOLOGY DEPARTMENT

USING DATA MINING TECHNIQUES TO DETERMINE THE EVIDENCE OF INFECTION WITH CORONAVIRUS "COVID-19" IN IRAQ

RUSLAN SALIM NASEEF

MASTER'S THESIS

KIRŞEHİR / 2022

T.C.
KIRSEHİR AHI EVRAN UNIVERSITY
INSTITUTE OF SCIENCE
ADVANCED TECHNOLOGY DEPARTMENT

USING DATA MINING TECHNIQUES TO DETERMINE THE EVIDENCE OF INFECTION WITH CORONAVIRUS "COVID-19" IN IRAQ

RUSLAN SALIM NASEEF

MASTER'S THESIS

Supervisor

Assistant Professor Dr. Murat IŞIK

KIRŞEHİR / 2022

DECLARATION

I declare that all the data in this thesis was obtained by myself in academic rules, all visual and written information and results were presented in accordance with academic and ethical rules, there is no distortion in the presented data, in case of utilizing other people's works they were referenced properly under scientific norms, the data presented in this thesis has not been used in any other thesis in this university or in any other university.

Ruslan Salim NASEEF

20.04.2016 tarihli Resmî Gazete ’de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, Kırşehir Ahi Evran Üniversitesi’nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü’nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

ACKNOWLEDGEMENT

Yüksek lisansa başlamamda ve yüksek lisans ders ve tez yazım sürecinde kendisini tanıdığım günden bu yana gösterdiği sakin ve sabırlı hali ile her zaman bana örnek olmasının yanı sıra bir bilim adamının nasıl çalışması gerektiğini kendisinden öğrendiğim değerli danışmanım Dr. Öğr. Üyesi Murat IŞIK 'a teşekkür ederim. Çalışmalarım boyunca maddi manevi destekleriyle beni hiçbir zaman yalnız bırakmayan, beni bugünlere getiren aileme de sonsuz teşekkürler ederim.

Haziran, 2022 Ruslan Salim NASEEF

INDEX

Page Number

FIGURE LIST

Page Number

TABLE LIST

Page Number

ABBREVIATIONS LIST

Abbreviation	Described
DM	**:** Data Mining
DMT	**:** Data Mining Techniques
CSV	**:** Comma Separated Values
KDD	**:** Knowledge Discovery in Data
K-NN	**:** K-Nearest Neighbour
LR	**:** Logistic Regression
DT	**:** Decision Tree
LDA	**:** Linear Discriminant Analysis
Gaussian NB	**:** Gaussian Naive Bayes
SVM	**:** Support Vector Machine
RFC	**:** Random Forest Classifier
MLP	**:** Multi-layer Perceptron classifier
GBC	**:** Gradient Boosting Classifier
CV	**:** Cross-Validation

ABSTRACT

M.Sc. THESIS

USING DATA MINING TECHNIQUES TO DETERMINE THE EVIDENCE OF INFECTION WITH CORONAVIRUS "COVID-19" IN IRAQ

RUSLAN SALIM NASEEF

Kırsehir Ahi Evran University
Science and Engineering Institute
Advanced Technologies Department

Supervisor: Assistant Professor Dr. Murat IŞIK

As the novel coronavirus epidemic spreads over the world, countries are developing new strategies to tackle the virus and keep it under control. Information on diagnostic techniques, infection and symptomology, and the most recent therapy and vaccine research have all been updated as part of the effort to battle the virus. It is the primary goal of this study to lessen the significant burden on the healthcare system by offering the best method for diagnosing patients and accurately predicting infection with Covid 19.

Data mining is an important field at today. This sector has started to expand fast as a result of explosive expansion of created data volume. Computerization of our culture and the fast development of strong data gathering and storage systems, Generate Big Data and the rapid development of technology, all these causes were adequate to sustain this expansion.

This study focuses on the most important and demanding medical-appropriate data mining algorithms and aims to explore how data mining can assist physicians in diagnosing Covid-19. In this regard, it is important to note the use of data extraction to assist in the rapid and accurate diagnosis and treatment of Covid-19, taking into account the time lost by analysts

in manual data analysis. The main motivation of the study was to observe a large flow of data through official reports, together with the number of epidemiological and scientific studies in the field of Covid-19 and the coronavirus family. A database created specifically for (Covid-19) disease, based on one clinical diagnosis sheet organized by the Iraqi Ministry of Health, was used to detect cases of infection with the virus, and the number of samples was 727.

The database included 28 features that were actually associated with the disease to determine the infection Because there are many algorithms and data mining tools available, we consider them only a few tools to evaluate these applications and develop classification rules that can be used for forecasting Nine algorithms were utilized to determine the most efficient algorithm to build the model. The Random Forest algorithm (RF) received the overall best performance compared to other models in terms of the classification accuracy 86.30 % to determine the incidence of the virus.

June 2022, 72 Pages

Keywords: Coronavirus, Pandemic, Data mining, Patients' recovery, Clinical diagnosis, Machine learning.

ÖZET

YÜKSEK LİSANS TEZİ

VERİ MADENCİLİĞİ TEKNİKLERİ İLE IRAK'TA KORONAVİRÜS "COVID-19" HASTALIĞININ TEŞHİS EDİLMESİ

RUSLAN SALIM NASEEF

Kırşehir Ahi Evran Üniversitesi
Fen Bilimleri Enstitüsü
İleri Teknolojiler Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Murat IŞIK

Ülkeler, virüsle mücadele etmek ve yeni koronavirüs salgını tüm dünyaya "Covid-19'un yayılmasını sınırlamak için farklı yollar geliştirmeye devam ediyor. Virüsle mücadele için alınan önlemlerin bir parçası olarak, teşhis yöntemleri, enfeksiyon ve semptomlar ve tedavi ve aşı ile ilgili en son araştırmalar hakkında bilgiler güncellenmiştir. Bu makalenin temel amacı, hastaları teşhis etmenin ve Covid-19 enfeksiyonunu etkili bir şekilde tahmin etmenin en iyi yolunu sağlayarak sağlık sistemi üzerindeki muazzam yükü azaltmaktır.

Veri madenciliği şu anda önemli bir alandır. Bu alan, üretilen veri hacminin patlayıcı büyümesi sonucunda hızla büyümeye başlamıştır. Toplumumuzun bilgisayarlaştırılması ve güçlü veri toplama ve depolama araçlarının hızlı gelişimi, Büyük Veri Üretme ve donanımın hızlı gelişimi, tüm bu nedenler bu büyümeyi desteklemek için yeterliydi.

Bu çalışma en önemli ve zorlu tıbbi uygun veri madenciliği algoritmalarına odaklanmaktadır ve veri madenciliğinin doktorlara Covid-19 teşhisinde nasıl yardımcı olabileceğini araştırmayı amaçlamaktadır. Bu bağlamda, manuel veri analizinde analistlerin kaybettiği zamanı dikkate alarak, Covid-19'un hızlı ve doğru teşhis ve tedavisine yardımcı olmak için veri çıkarma kullanımını not etmek önemlidir. Çalışmanın ana motivasyonu, Covid-19've koronavirüs ailesi alanındaki epidemiyolojik ve bilimsel çalışmaların sayısı ile birlikte resmi raporlar aracılığıyla büyük bir veri akışını

gözlemlemektir. Irak Sağlık Bakanlığı tarafından düzenlenen bir klinik tanı tablosuna dayanan (Covid-19) hastalığı için özel olarak oluşturulan bir veritabanı, virüsle enfeksiyon vakalarını tespit etmek için kullanıldı ve numune sayısı 727'idi.

Veritabanı, enfeksiyonu belirlemek için aslında hastalıkla ilişkili 28 özellik içeriyordu. Çünkü birçok algoritma ve veri madenciliği aracı mevcut, onları bu uygulamaları değerlendirmek ve tahmin etmek için kullanılabilecek sınıflandırma kuralları geliştirmek için sadece birkaç araç olarak görüyoruz. Modeli oluşturmak için en verimli algoritmayı belirlemek için dokuz algoritma kullanıldı. Rastgele Orman algoritması (RO), virüsün insidansını belirlemek için %86,30 sınıflandırma doğruluğu açısından diğer modellere kıyasla genel olarak en iyi performansı aldı.

Haziran 2022, 72 Sayfa

Anahtar Kelimeleri: coronavirüs, pandemi, veri madenciliği, klinik teşhis, makine öğrenimi.

Chapter 1

1. INTRODUCTION

Since the various applications of artificial intelligence techniques have appeared, mankind has been searching for a way to enable it to benefit from these technologies in achieving and maintaining its prosperity and obtaining comfort, happiness and well-being through them.

No one expected that humanity would be in dire need of it in light of the crises and disasters it will witness from demise. The world was and still suffers from continuous humanitarian crises caused by natural disasters, and epidemics that humanity is exposed to many times, which imposes on us the need for the applications of artificial intelligence and data science to have a clear role in alleviating the suffering of the sick and injured, helping humanity [1], and developing ways that It predicts disasters before they happen and develops ways and means to deal with them before, during and after their occurrence. In this sense, Lucas Juba, who heads the “Artificial Intelligence for Earth” program at Microsoft, says: “We believe that artificial intelligence can be a game-changer in the face of pressing societal challenges. and create a better future for her.” [2]

As the new Covid-19 pandemic spreads across the globe, countries continue to develop various methods to combat and reduce the spread of the virus. There has been a significant increase in knowledge about diagnosis, infection, and symptoms as part of the actions done to battle the virus [3].

The virus can lead to death in patients with chronic diseases such as stress, heart, and diabetes, and those with high temperatures and low respiratory conditions should receive medical support. All the information circulating about the virus on social media, official websites and hospital reports, generated a huge amount of data about the virus. [4]

When compared to the global pandemic of 1918, known as the "Spanish flu," which spread across the globe, Epidemic disease events were still largely uncontrolled and unexplained

[5] and scientific studies on virology and epidemiology in general and the Covid-19 virus family in particular, as well as the MERS and SARS viruses.

Massive amounts of user-generated Big Data are now readily available in the form of posts and searches on social media platforms, communication networks, and search engines. You may watch the amount of searches for symptoms of a disease on Google every day [6], which could indicate the beginning of a certain epidemic in a particular location, as an example.

There are three basic components to data mining: statistics and mathematics, programming abilities, especially artificial intelligence and machine learning, and knowledge of the nature of the subject whose data is being watched and analyzed. The 2011, US health care system reports that data alone, up to 150 exabytes. Sponsorship big data at this growth rate will soon reach the zettabyte scale (1021 GB), and soon after, the yotabyte scale (1024 GB) [7].

We need to increase the quality of service for data management. It includes data management and improvement of work mechanisms in the health sector. By using computerized data management systems in medical offices to manage data, organize and control its work [8].

1.1. Aim

The thesis' major goal is to use data mining techniques by employing artificial intelligence algorithms to facilitate and speed up the patient registration process to quickly access the correct archive data during treatment [9], and statistical data more easily, to predict a person's infection with Covid-19 infection. The medical staff helps Improving the speed and accuracy of diagnoses and disease detection, assisting in clinical care, reducing medical errors, helps support the decision to treat the patient.

We have dealt with problems in hospitals and institutions that have identified large amounts of data on infected patients that need to be organized and coordinated. Predictive analysis of tendencies is challenging [10] but ultimately it will facilitate creating the right choices with timely information about the patient and medication.

The purpose of this thesis lies in predicting the patient's diagnosis correctly, accurately and quickly, relying on the clinical diagnosis of Covid-19 before relying on laboratory analyzes

by building a mathematical model and using machine learning prediction algorithms, where nine algorithms will be relied on to create the model so that the best outcomes can be compared and the best algorithm can be used.

1.2. Literature Review

Mariam Laatifi et al. [11], 337 samples were used from Sheikh Zayed Hospital in the UAE, four algorithms X_GBoost, AdaBoost, Random Forest and Extra Trees were used. Laboratory examinations were relied on in building the database, where it was found that the Extra Trees classifier is the best among the classifiers and achieved 100% accuracy, Differences with our study The location of the sample differs in Morocco, the study determines the severity of infection with the virus, while the similarity with this study is the use of clinical examinations and some algorithms similar to our study were used.

Mohammad Pourhomayoun et al. [12], Six methods were employed to analyze 117,000 samples from around the world: Support Vector Machine (SVM), Artificial Neural Networks (ANN), Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbor (KNN). The Random Forest classifier was determined to be the best among the classifiers, with an accuracy of 97%. The difference in this study is the reliance on laboratory tests in constructing the data set. As for the location of the sample, it was obtained from internet sites, and it was relied on a separate data set to determine the characteristics that lead to death. The similarity with our study is the reliance on the same algorithms in building the model.

Xiangao Jiang et al [13], 53 samples were used that depended on 11 features to build the database that were taken from Chinese hospitals, six algorithms were Logistic Regression, KNN, Decision Tree (based on Gain Ratio), Decision Tree (based on Gini Index), Random Forests and Support Vector Machine The accuracy was 70-80%, where KNN and SVM were the best among them, the difference in this study is the reliance on laboratory tests in building the data set. As for the location of the sample, it was obtained from hospitals in China. The goal of this study is to find out incidence of acute respiratory distress syndrome (ARDS), which has a significant impact on reaching dangerous stages when infected with the Corona virus. The similarity is the use of five algorithms that were used in building models for our studies, and the same method of selecting data was used for training and testing (k-fold).

Li Yan et al. [14], 485 blood samples were used in a hospital in Wuhan, China, to build a database that was taken from Chinese hospitals. A model was built based on the XGBoost algorithm, and the model's accuracy was 90%, The goal in determining infection with the Corona virus is the same as the goal of our study and the use of the decision tree algorithm, and the difference is the type and location of the sample, where blood tests were relied on and three elements in the blood were identified as the main cause of infection.

A classification algorithm based on XGBoost was built by Li et al. [15] and included clinical data from 413 patients. We got 92.5 % sensitivity and an accuracy rate of 97.1 %. Clinical exams were used to generate the data set, which is identical to our study, in order to determine the presence of the Corona virus. According to their findings, age, CT scan results, temperature, lymphocyte concentrations, fever and cough were the most relevant aspects of their prediction model in a different data set.

Models for predicting COVID-19 diagnosis have been developed by Brinati et al. [16]. Italy's IRCCS Hopedale San Raffaele used 279 samples for this research project. Numerous machine learning techniques were used in the study ranging from DT to ET to KNN to SVM to tree-weighting RF (TWRF). The RF classifier had the greatest performance, with an AUC of 84%, an accuracy of 82%, a sensitivity of 92%, and a specificity of 65%. There were two sets of data, one from RT-PCR laboratory testing and the other from clinical tests. A key distinction between our analysis and that of the researchers is that AST has a significant influence on model performance. As for the similarities in the reliance on clinical examinations in building the data set, and the adoption of six algorithms that were used in our study.

Using clinical blood test data, Jiangpeng Wu et al. [17] identify COVID-19 patients. A total of 253 samples were collected throughout China as part of this collection. The model has an AUC of 99.26%, a sensitivity of 100%, and a specificity of 94.44 percent thanks to the use of 49 characteristics. We note that the difference in this study is the location of the sample and the type of data set (laboratory tests for blood), while the similarities are by using the random forest algorithm to build the model.

Tschoellitsch et al. [18] developed a model that uses an RF algorithm to predict COVID-19 diagnosis based on standard blood testing. The model was built using a dataset of 1528 samples and has an accuracy of 81 percent. Leukocyte count, red blood cell distribution

width (RDW), hemoglobin, and serum calcium were determined to be the most relevant indicators in predicting diagnosis in the research.

Laboratory analysis based on polymerase chains was used to build the data set, and this is one of the differences. The similarity is the use of the random forest algorithm to build the model.

Atta-ur-Rahman et al. [19], The proposed model is built in the MATLAB software tool. The dataset contains 547 samples that were categorized with an accuracy of 98.4% and the machine learning algorithm was SVM.

The usage of the SVM technique to develop the model is similar; the difference is a method that uses iterative cross-validation 5-fold, 10-fold, 15-fold, and 20-fold to split data between training and testing. Sina Ardabili et al. [20], 500 samples, using Artificial Neural Network (ANN) algorithms. The data is divided into 70% of the total data They were selected for the training phase, 30% of the total data were selected the test accuracy was obtained 99.4%. The difference is the use of neural networks using the gray wolf optimization algorithm to build the model. The similarity is the reliance on clinical examinations of covid-19 patients.

LipingSun et al. [21], 336 sample in Shanghai, use Support Vector Machine (SVM), Clinical and laboratory features and demographics were used to build the data set, and the accuracy rate was 77.5%. Clinical examinations were used as the data set and the SVM algorithm was adopted to build the model, and these are the most important points of similarity with our study. As for the difference, it was found that age is the feature affecting the model's performance, while the results of the study were to determine the separation between critical and non-critical cases.

Artificial neural network (ANN), random forest (RF), support vector machines (SVM), decision trees (DT) which comprise classification and regression trees (CART), and gradient boosted trees (GBM) were utilized in the dataset bulled by Rabia Al Mamlook et al. [22]. For our investigation, we used 90% of these methods to develop a model.

All of these methods have been taken into account in the classification and the selection of the most appropriate model. A conventional cross-validation approach was used to evaluate the performance of each model. CART is the most accurate model for predicting classes of children with COVID-19 based on these results for classification performance

and accuracy prediction. The most significant disparities between these model results and those of our investigation were found to be in the concentrations of leukocytes, monocytes, potassium, and eosinophils. With the CART model, 92.5 percent of the time, researchers were able to accurately classify the data.

Xiangao Jiang et al. [23], 53 clinical examinations of disease and predictive indicators were based on a case series from Wenzhou, Zhejiang, China. Use 11 feature, and six algorithms Logistic Regression, KNN, Decision Tree, Decision Tree, Random Forests, Support Vector Machine, the beast accuracy KNN and SVM IT 80%. Gender and age are the features that affect the performance of the model, and this feature differs with our study. The similarity is the use of 70% of the algorithms that we have adopted to build our models.

1.3. Organization of Thesis

First chapter includes an introduction to the emergence and spread of the epidemic in the world and the infection of a large number of people, which led to the generation of large amounts of data about the virus. The need to use data mining techniques and artificial intelligence algorithms to study this virus. Also, there were fifteen literature studies cited, the amount of samples, the method utilized to create each model and its correctness were all determined in this thesis.

Second chapter defines the term data mining and defines the fields on which it depends. The five techniques for exploration and exploration stages are identified, which are six stages, and each stage is explained. The data set and its types and method of collection were also identified. As for the mining algorithms, the nine algorithms that will be adopted in building our model were explained, and the K-fold Cross-Validation was explained to identify the mechanism that was used to select samples when conducting training and testing. As for the method of testing the performance of the algorithms, it was necessary to identify the Confusion Matrix, to identify all its details and explanation, and to identify the mathematical formulas that are used for calculating, and another measure was used to perform the trained model, which is the Classification Report.

Third Chapter, in this chapter, we will talk about the methodology that was followed to build the model from the beginning of data collection and knowing its details and the number of features that it consists of in addition to knowing the number of samples and

clarifying its extents in the form of diagrams. Clarifying the formula in which the data was collected, knowing the features of this formula, discussing the most important tools that were used in building the model (programming language) and explaining the reasons for choosing this language.

Fourth chapter was introduced to the built model, its structure and the stage that the data set goes through to obtain clean data. The process of data unification and selection of features and the process of determining the importance of features and measuring the degree of their impact on the performance of the model were also identified. At the end of the chapter, the analysis and evaluation of the model that was made was done. constructively and compare the results obtained from the nine algorithms when training and testing.

Fifth chapter contains the results obtained from this thesis and the discussion of these results and their importance in future studies and the most important obstacles that we encountered during the construction of the model. A number of recommendations for the development of work in the future were also identified.

Chapter 2

2. FUNDAMENTAL CONCEPTS

2.1. Data Mining (DM)

There are many different ways to mine data for useful information, such as finding patterns and correlations and identifying anomalies and key structures. Data mining is one of the most often used methods. Data mining has become increasingly popular in the information industry in recent years as a result of the easy accessibility of enormous amounts of data in electronic form and the pressing need to turn this data into useful information and knowledge for a variety of purposes, including market analysis, healthcare, and decision support [24, 25].

It is a contemporary, multidisciplinary field, relying on topics such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing. Figure 2.1.

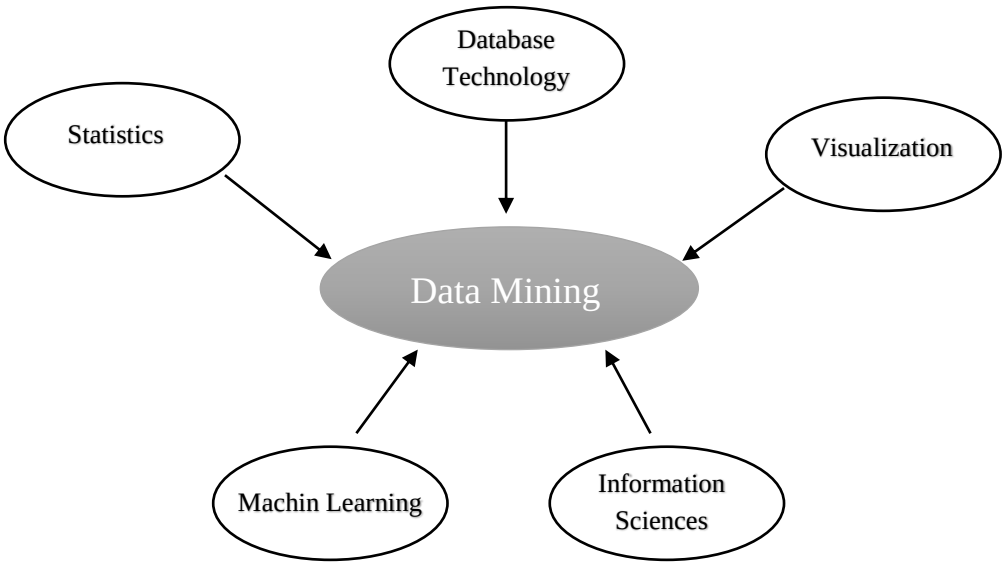


Figure 2.1: intersection of multiple disciplines as tools and basics in data mining

As well as the aforementioned domains of neural networks and pattern recognition, geographical data analysis and image databases and signal processing are a number of application areas such as business, economics and biology. Extracting previously unknown, accurate, relevant, and hidden patterns from enormous data sets is the goal of data aggregation.

In a nutshell, data mining is the process of extracting useful information from a big amount of raw data. Using advanced mathematical techniques, data mining is used to segment data and assess the likelihood of future events. [26] data mining is the core in the 'Knowledge Discovery in Data' (KDD) [27] as shown in Figure 2.2

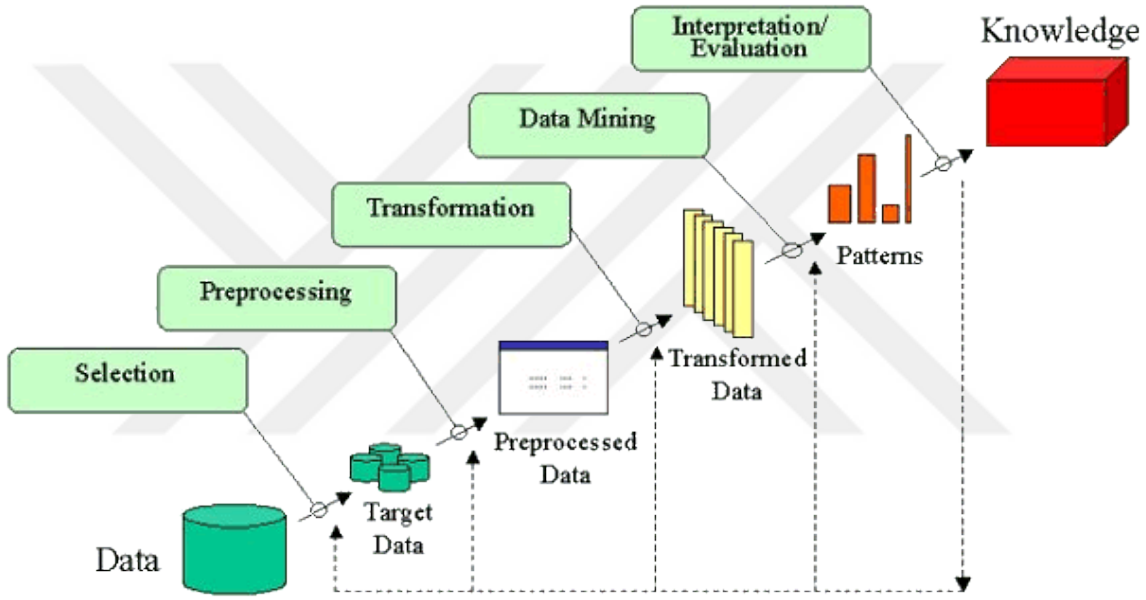


Figure 2.2: Knowledge Discovery Databases (KDD) Process

2.2. Data Mining Techniques (DMT)

Each of the following data mining methods targets a specific issue and offers a unique perspective. The sort of data mining approach that produces the greatest results depends on the type of business challenge you're seeking to address. We live in a digital age when the amount of data available to us is increasing at an exponential rate.

It's a paradox that we're awash in information, but knowledge is scarce. Why? We've created a lot of amorphous data, but our big data projects have failed because we can't exploit it effectively. The information is tucked away somewhere deep within. Without effective mining tools or procedures, we cannot reap any benefits from this data.

2.2.1. Association

It is one of the most frequently used data mining methods. Using a transaction and the connections between its components, this approach seeks out patterns. This is why it is referred to as a relational technique [28]. In order to discover all the goods that clients often purchase together, market basket analysis makes use of it.

Retailers may learn a lot about their consumers' purchasing patterns by utilizing this method. In the past, retailers can research sales data and then hunt for things that people purchase in combination. To make customers' lives easier and boost sales, retailers can group similar items together in their physical stores. [29]

2.2.2. Clustering

Clustering is a way of grouping together things that share a common trait. Finally, clustering is aimed at putting together groups of data that share many of the same characteristics. Data mining relies on the division of items into homogenous clusters as a basic activity. A variety of data mining jobs rely on this procedure. Data mining relies heavily on clustering or data grouping as its primary method. In this endeavor, one is looking for a finite number of categories, known as clusters, to represent the data. The notion of maximizing interclass similarity and decreasing interclass similarity guides the clustering of data. The purpose of clustering is to identify the natural groupings in a set of unlabeled data [30]

Although it's easy to mistake it with categorization, if you know how they both function, you won't have any problems. Instead of classifying items based on established categories, clustering groups them according to the criteria it specifies. As the name suggests, data clustering is a database management strategy in which data that is conceptually related is grouped together. The amount of disk visits in database administration is to be reduced in order to improve search and retrieval performance.

A single disk access is all that is needed to obtain the full class of objects in a clustered database. The issue space can be condensed by focusing on a single representative from each subgroup of the population once it has been clustered in some abstract algorithmic space. In the end, clustering is a way to break down a large amount of data into smaller, more manageable chunks. These piles may be made up of "similar" things for the sake of simplifying cognitive and computational processes. [31]

2.2.3. Classification

In this case, machine learning is the source of inspiration. Predefined groups or classes are used to organize the variables and elements in a dataset. In data mining, linear programming, statistics, decision trees, and artificial neural networks are only some of the tools used. Programming that can be described in such a manner that it is able to categorize things in a data collection into multiple categories is developed using classification techniques. As an example, we can use it to divide a candidate pool into two groups: those who were accepted and those who were denied.

2.2.4. Prediction

Recognizing missing or unavailable data for a new observation method is what this is about. The validity of a prediction is determined by the accuracy with which a particular predictor can forecast the new value of a predicted property [32].

The connection between independent and dependent variables, as well as the relationship between independent variables alone, may be predicted using this approach. Based on the sale, it may anticipate the company's future earnings. It is not uncommon for IT professionals to use the terms "predictive data mining" and "predictive analytics" interchangeably. When data is used to predict future outcomes, predictive analytics is the method of choice. Going through system databases, looking for relevant data, and then analyzing it is known as data mining.

2.2.5. Sequential Patterns

The goal of this method is to look at a large amount of transaction data and see if any trends, patterns, or occurrences emerge. The sales data from the past may be utilized to identify things that customers bought in groups at various periods of the year [33]. Using this data, the business can urge clients buy things at times when past data suggests they won't. This information can make sense. It's possible for companies to utilize lucrative offers and discounts to promote this suggestion.

2.3. Data Mining Process

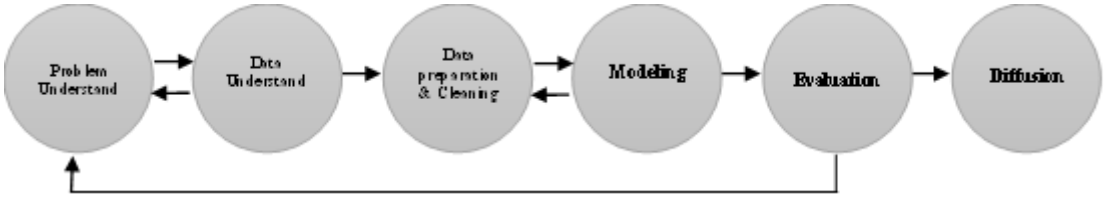


Figure 2.2: Data Mining Process

Figure 2.2 shows the data mining process steps. Let's examine each process step in detail.

2.3.1. Understand Problem

Everything Depending on your objectives, the mining procedure may accommodate anything. The formulation of a problem statement necessitates an awareness of and articulation of organizational needs and requirements. Once we have a clear understanding of the problem, we can begin to collect the data we need to solve it

Among the problems identified in this research that must be solved using data mining:

- The Covid-19 virus spread extensively due to the difficulty to correctly establish the reasons of the disease's dissemination.
- Not knowing how to take advantage of the large amount of existing data (keeping pace with technical development)
- Difficulty and imprecise planning to diagnose and treat disease in a short time.
- The waste of time in the process of searching, analyzing and extracting knowledge from a dataset

2.3.2. Understanding the Data

Our major goal is to collect original data, interpret it, and then use it to identify problems with quality, consistency, and readability so that we can get back to the very beginning of the concepts we're exploring [34].

- What you need to do at this stage is define:
- "Where did the data come from"?

- "Who compiled it, and did its group follow standard methods"?
- "What do the different columns and rows of data mean"?
- "Describe data, verify its volume and examine its aggregate characteristics".
- "Accessibility and availability of features. Trait types, attachments, and hobbies".
- "Understand the meaning and value of each feature in the business terms".

2.3.3. Data preparation

In this section, you'll find all of the steps that were used to transform raw data into a finished dataset. The vast majority of the project time is spent transforming raw data into an analytical dataset. It's time to go to work on gathering and defining the data sources so that we can develop and style it appropriately. The work of sifting through the data must be done thoroughly and meticulously in order to discover patterns that might help businesses make sense of their data [35]. The model's ultimate performance will be influenced by the quality of the data displayed or the final data.

- Data preparation tasks are likely to be completed several times rather than in a set order and comprise a variety of actions for example.
- Combining many sets of data
- Reducing the number of variables in a data collection that are relevant to a given data problem (feature engineering).
- Data cleaning and clarification "imbalances such as outliers, missing data, reformatting", and (data cleansing)
- Check for inconsistent data.

2.3.4. Data Cleaning

It is imperative that we go through this procedure in order to eliminate the possibility of missing data, irregular data (Outliers), repetitive data, duplicates, inconsistent data, incorrect capitalization, and other issues. Therefore, it was imperative that the data be cleaned before it could be used to develop the model.

Benefit Data Cleaning

It's difficult to avoid errors and inconsistencies when gathering data from many sources, and this is especially true when attempting to reduce errors. Reduce Data Noise Recognize the information The first step in gaining a complete grasp of your data is to begin cleaning it. This is where the real learning begins [36]. And when you understand the data you have, it goes beyond this understanding to remove from it what you see as inappropriate for it, and this leads you to define your goals as well as understand the nature of the variables and others.

2.3.5. Modeling

In data mining, it is an arithmetic representation of real word observation. Models are applications of an algorithm to find, identify, and display any patterns or message in your data. There are two types of models in data mining as shown Figure 2.3

- Descriptive
- Predictive

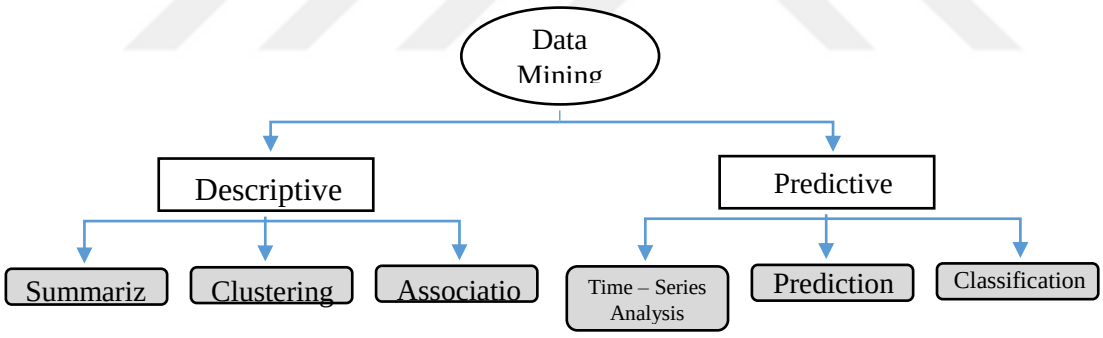


Figure 2.3: Processes that occur within descriptive and predictive models

The parameters of various modeling approaches are determined to the optimal values at this step, where they are picked and implemented. In data mining, many strategies are often employed to solve a single problem. Depending on the technology, there may be restrictions on the data format required.

When this happens a return to data preparation is frequently required. Decision trees, random forests, KNNs, Naive Bayes, K means, linear regression, and logistic regression are some of the most prominent methods. When this happens a return to data preparation is

frequently required. Decision trees, random forests, KNNs, Naive Bayes, K means, linear regression, and logistic regression are some of the most prominent methods.

2.3.6. Evaluation

In this step, the results are analyzed in relation to the company's goals. Determine whether there are any issues that have not been effectively addressed [37]. At this point, a choice must be taken whether or not to proceed with publication.

2.3.7. Diffusion

The last step in the data mining process, is to deploy the models that performed the best to a production environment [38]. This is where we'll talk about how to put the data to use. Stakeholders need to be able to make use of the information that has been gathered. Publishing might be as easy as creating a report or as sophisticated as implementing data mining that can be duplicated, depending on the requirements.

2.4. Dataset

Scientific study or an individual project may need the collection of a dataset. Tables, databases, and other representations of this data can be found. It is common for this group to be presented in a tabular format. Each column is devoted to a single metric. In the data set, each row represents a distinct individual [39,40]. Managing data is an element of this.

2.4.1. Types of Datasets

For various forms of information, we have a variety of data sets at our disposal. These are their names:

- In the case of numerical data, it must be expressed numerically.
- Bivariate: It is termed a bivariate dataset when two variables are included in the data collection, such as ice cream sales and the temperature on that day.
- Multivariate: One that has several variables. It is considered a multivariate data set when it comprises three or more types of data (variables). Multivariate data is a collection of measurements based on at least three independent factors. Example A

rectangular box's length, breadth, height, and volume may all be measured using this method.

- Categorical: A person's or an object's qualities are reflected in categorical data sets. In a categorical data collection, there is just one categorical variable, which is also known as a qualitative variable. As a result, a dichotomous variable is used to describe it. An illustration of a person's gender (male or female), Unmarried/married status.

- Correlation: Correlation data sets are the set of values that demonstrate some correlation to one other. In this case, the values are shown to be interdependent.

- As a general rule, the statistical relationship between two ideal variables is what is meant by correlation. [41]

2.4.2. Data collection methods

There cloud be many way to collect the data, but some of them can be listed as follows: [42,43]

Interview: It is possible to conduct an interview with the target group either directly or via a method of communication. Focused and open-ended questions should be asked in the interview.

Questionnaire: Data collection is done by creating forms that contain the questions needed to target a specific group, for example "women's groups" if the goal of this data is to understand breast cancer awareness, and then collecting the questionnaire results and answers from a sample of the participants after they have completed the questionnaires.

Observations: In order to examine or observe changes in the circumstance or the recurrence of actions, taking notes is essential. An excellent source for extra information about a certain organization.

Focus groups : An interview with a group of people who share a common interest in order to acquire information about their thoughts and opinions. Frequently, answers are categorized and examined according to the subject matter.

"Documents and Records": data from databases, meeting minutes, reports (including attendance logs and financial statements), newsletters, etc. While this is a low-cost method of data collection, it is possible that the data collected is incomplete.

2.5. Data Mining Algorithms

Most data mining techniques have been created and deployed in real-world operations to unearth information hidden within large databases. There are a variety of approaches that may be used to solve various business issues. To get the greatest results from data mining, you need to know what kind of business problem you're seeking to address:

2.5.1. K-Nearest Neighbour (K‑NN)

Classification is a common usage [44] for this technique. Two types of probabilities are generated by the algorithm. This algorithm impacts classification accuracy in many fields, the most important of which are the medical field and various digital data.

Nearest Neighbour is a classification algorithm used to classify a class from unknown groups. An algorithm that uses the entire sample set has the potential to outperform algorithms that use individual data [45]. It is treated as a supervised learning algorithm.

2.5.2. Logistic Regression(LR)

Logistic regression is a sort of regression analysis in statistics that uses a collection of predictors or independent variables to predict the result of a categorical dependent variable (a dependent variable with a finite number of values) [46]. Logistic regression once in a while called the logistic version or logit model, analyzes the connection between multiple unbiased variables and an express based variable and estimates the possibility of prevalence of an event with the aid of fitting facts to a logistic curve. There are two models of logistic regression, binary logistic regression and multinomial logistic regression [47]. Binary logistic regression is normally used when the based variable is dichotomous and the unbiased variables are both non-stop or specific. Whilst the based variable is not dichotomous and is constituted of more than categories, a multinomial logistic regression can be hired.

2.5.3. Decision Tree (DT)

This is a classification method and one of the most widely used data mining approaches. A decision tree is a tree-like structure with root nodes, branches, and leaves. Each inner node refers to an attribute test, each branch to the test result, and each leaf node to the class label [48]. The root node is the topmost node in the tree. Each base reflects the journey from the root to each leaf in a unique way [3]. It is possible to extract knowledge and reflect it in categorization rules (if-then).

2.5.4. Linear Discriminant Analysis (LDA)

LDA is a classic method for supervised dimensionality reduction. It is a method of dimensionality reduction that finds the optimal linear transformation that increases the reparability of classes. [49,50] The idea of the LDA is to display the samples in a straight line. Sample points of the same type are as close as possible, and sample points of different types are as far away as possible. In other words, each sample of its data set has an output of the class. The idea of LDA can be summed up in one sentence, which is "the smallest inter-layer variance and the largest inter-layer variance" after projection.

2.5.5. Gaussian Naive Bayes (Gaussian NB)

To identify between instances of a data collection based on certain attributes or qualities, one sort of data mining classification technique is utilized [51]. Naive Bayes classifiers are really basic, yet they've shown to be effective in a variety of real-world applications, including the cancer/non-cancer classification. In comparison to other ML classifiers, it takes extremely little training data and may be quite quick. On the other hand, while naïve Bayes is regarded as a good classifier, she is regarded as a poor estimator [52]

2.5.6. Support Vector Machine (SVM)

SVM stands for Support Vector Machines and is a strong and complex algorithm based on the Vapnik-Chervonenkis theory. The marshaling properties of SVM are excellent. The term "regularization" refers to the model's generalization to new data [53]. Because of its efficacy and great accuracy in most utilized data, one of the Supervised Learning Algorithms, which may be employed in Classification and Regression issues, is frequently used in Classification problems. There are a variety of different hyperplanes that might be used to split the two classes of data points. Finding a plane with the greatest margin—that

is, the greatest separation between data points from both classes—is our goal. Maximizing the margin distance adds some support, increasing the confidence with which future data points can be categorized. [54] Our aim with the SVM method is to discover the optimal cutoff level for categorizing the data, but we'll have more than one separator level when we start training the algorithm, so we'll have to pick only one cutoff level from all of them.

2.5.7. Random Forest Classifier

The supervised learning approach includes the widely used machine learning algorithm known as random forest. Using several classifiers to solve a complicated issue and increase model performance is at the heart of the notion of group learning. In ML, they may be applied to issues involving classification and regression.

The Random Forest Algorithm, as the name indicates, is a classifier that incorporates a number of decision trees on different subsets of a given data set and takes the average to increase predicted accuracy of that data set. For a random forest, each forecast is taken from all of the trees, a majority of votes are used to make predictions, and a final output is predicted. In order to minimize overfitting, a larger number of trees in the forest is required [55] The flexibility of Random Forest is one of its most significant features. Additionally, the relative value of input characteristics is quite clear to see. You may also utilize Random Forest because the default hyper settings typically lead to a successful prediction. Hyper parameters are simple to grasp, and there aren't many of them.

Overfitting is a common issue in machine learning, however the random forest classifier almost eliminates it. It is impossible for the classifier to work with a model that has too many trees [56].

2.5.8. MLP Classifier

With an arbitrary number of synapses and weights connecting its neurons, MLP is an artificial neural network that feeds back information to its hidden layers. Because it is unable to anticipate the intended output in the intermediate layers, the layer is referred to as hidden. Back propagation is a supervised learning technique commonly used to construct this sort of network [57].

2.5.9. Gradient Boosting Classifier

Using ensemble approaches, a single best-fitting predictive model may be created by combining many different machine learning models. Stacking, mixing, bagging, and boosting are just a few examples of ensemble techniques. Gradient It's a way of boosting, as the name implies. Multiple basic models are combined into a single composite model, which is known as "boost" [58]. Improved accuracy of the model as a whole.

In machine learning, clustering is a strategy that integrates multiple different models in order to build the best possible prediction model. Stacking, mixing, packing, and boosting are all examples of stacking techniques. Gradient boosting, as the name suggests, is a technique for enhancing performance over time. Boosting is the process of creating a more complex model from a collection of smaller, simpler models. With the introduction of simpler models, the general model becomes a more accurate predictor.

2.6. K-fold Cross-Validation

To evaluate if a model can predict data that hasn't been fed into it during training, a technique known as cross-validation (or CV) is employed as seen in Figure 2.4. In situations where we have a small sample size, a CV might be beneficial. A CV may be done in a variety of ways. CV typically divides the training data into k blocks before analyzing it. After k-1 blocks have been trained and verified, each iteration the model is run again until it is ready to be used. Reduced variability can be achieved by using many CV iterations. The model's performance is evaluated by calculating the average of the errors in each iteration.

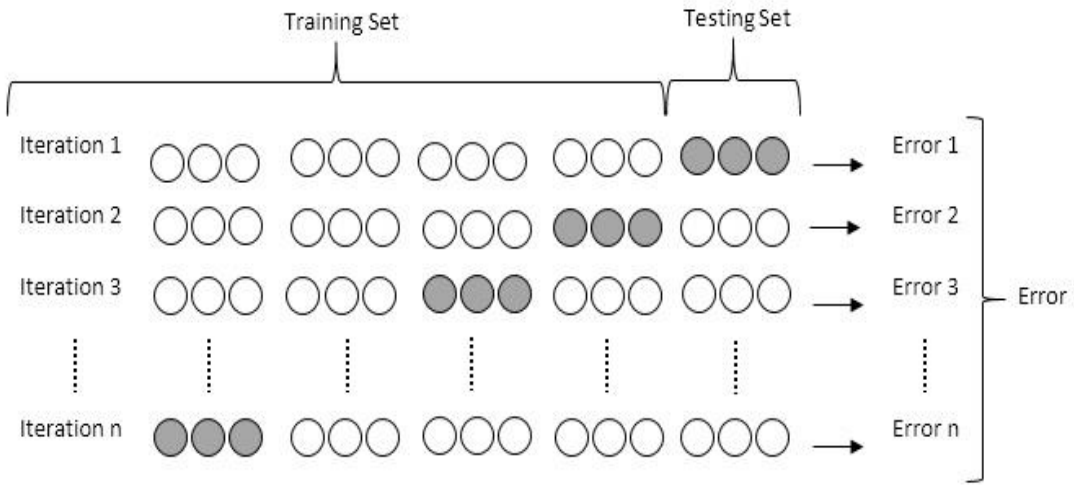


Figure 2.4: K-fold Cross Validation Technique

2.7. Confusion Matrix

The algorithms were tested using the Confusion Matrix. Tables like this one include information and details regarding the data set's actual ratings and the classifier's predicted ratings, helps us understand distances in high-dimensional datasets [59]. Predicted classes are represented by the array's columns, while real-world classes are represented by its rows. The data in the matrix is used to evaluate the classifier's performance, and the matrix size was 2x2 as shown in Figure 2.5

The following table shows the shape of the matrix and the meaning of each column of description in it and each cell in the matrix, the meaning of which we will explain immediately.

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Figure 2.5: Confusion Matrix

P, N represents the existing categories. Each entry (cell) in this array has a meaning, and we will explain it now:

True Positive (TP): The number of states correctly predicted by the classifier of class p and actually belonging to class p .

$$TP = \frac{TN}{FP+TN}$$ Equation 2.1

False Negative (FN): The number of states incorrectly predicted by the classifier of class p and actually belonging to class n.

$$FN = \frac{FP}{FP+TN}$$ Equation 2.2

True Negative (TN): The number of states correctly predicted by the classifier of class n and actually belonging to class p .

$$\mathrm{TN} = \frac{TP}{TP+FN}$$ Equation 2.3

False Positive (FP): The number of states incorrectly predicted by the classifier of class n and actually belonging to class p.

$$FP = \frac{FN}{FN+TP}$$ Equation 2.4

2.8. Classification Report

It's one of the measures used to assess the success of a classification-based machine learning model. It allows us to have a better grasp of our trained model's overall performance. It shows the accuracy, recall, F1 score, and support of your model. [60]

To fully comprehend the categorization report, you must be familiar with all of the parameters listed there:

Precision : refers to how many true positives are in comparison to all of the false positives, including both true and false ones.

Recall : True positives are defined as the total of true positives and false negatives divided by the amount of true positives.

F1 Score : Weighted harmonic mean of accuracy and recall is the F1. The model's projected performance improves when the F1 score gets closer to 1.0.

Support : may be used to determine how many times the class appears in the dataset. Only the performance evaluation method is different between models.

2.9. Performance Evaluation of Model

To determine the accuracy of data mining models, evaluation procedures are used. A test's capacity to distinguish between healthy and diseased patients is a measure of its accuracy. True positives and false negatives should be taken into account when determining a test's accuracy. This may be expressed mathematically as follows:

Where TP is the True Positive, TN is the true negative, FP is the false positive, while FN is the false negative .

$$\text{Accuracy} = \frac{\text{TP}}{\text{TP+FP+FP+FN}} \quad \text{Equation 2.5}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{Equation 2.6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FP}} \quad \text{Equation 2.7}$$

$$\text{F1 Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad \text{Equation 2.8}$$

Chapter 3

3. MATERIAL AND METHOD

In this chapter, an explanation will be given to the methodology that was used, where the features that were used in building the database and its types will be identified. The stages of processing it and demonstrating the extent to which these aspects have an impact on the model's performance and importance, as well as the model's operation and the algorithm employed in its construction. The language that was used in the model and the reasons for using it will also be addressed.

3.1. Dataset Collection and Description

The database of Covid-19 patients was obtained from the official Iraqi authorities. The database contains 60 features, depending on the two methods (questionnaire, documents and records). The dataset was obtained from the Iraqi Ministry of Health hospitals; it contains the features (clinical diagnosis) on the basis of which the virus is diagnosed. AP.1, and the form approved by the Ministry for diagnosing clinical signs of Covid-19 was approved. We analyzed data from Iraqi Covid-19 patients. These 727 specimens exhibit 28 different characteristics, including (Sex, Age, Acute flaccid paralysis, Cardiovascular Disease, Shortness breath, Diabetes, High Blood Pressure, Cancer, Liver disease, AIDS, Kidney disease, Smoking, Bronchial respiration, No. of Pregnancy, Fever, Wheezing, Burning pharynx, Nausea vomiting, Chest bang, Cough, Headache, Cyanosis, Runny nose, Convulsions, Frenzy confusion, Slenderness, Diarrhea, Chronic obstructive pulmonary disease).

3.2. Dataset Preparation

There are just the properties from the original dataset that apply to this dataset now that it has been cleaned and prepped. 727 samples with 28 attributes were retrieved from the original data set. Table 3.1 reveals the attributes of the data type, and Table 3.2 shows a sample of some of the data.

We only looked at two patients, one of whom is recuperating and the other is wounded. The isolation condition was eliminated using pointless characteristics, and the inclusion of the missing value in the data set causes skewed estimations, leading to an incorrect conclusion in the dataset [61].

Table 3.1: Features type of values

No	Column	Type
0	Sex	int64
1	Age	int64
2	Acute_ flaccid_ paralysis	int64
3	Cardiovascular_ Disease	int64
4	Shortness _ breath	int64
5	Diabetes	int64
6	High Blood Pressure	int64
7	Cancer	int64
8	Liver disease	int64
9	AIDS	int64
10	Kidney disease	int64
11	Smoking	int64
12	Bronchial _respiration	int64
13	No. of Pregnancy	int64
14	Fever	float64
15	Wheezing	int64
16	Burning pharynx	int64
17	Nausea_ vomiting	int64
18	Chest_ bang	int64
19	Cough	int64
20	Headache	int64
21	cyanosis	int64
22	Runny nose	int64
23	Convulsions	int64
24	Frenzy confusion	int64
25	Slenderness	int64
26	Diarrhea	int64
27	Chronic obstructive pulmonary diseases	int64
28	Outcome	int64

We also note that the features differ in the types of their values (binary, integer, and fractional), while the remaining features are not influencing the model's work (name, address, ... etc.). Table 3.2 shows a sample from the dataset and Table 3.3 show the range value for each feature in dataset.

Table 3.2: A Sample Dataset

No	Sex	Age	Acute_flaccid _paralysis	Cardiovascular_ Disease	…	Chronic obstructive pulmonary diseases	Outcome
0	1	31	0	0	…	40	1
1	0	32	0	0	⋯	55	1
2	1	21	0	0	⋯	72	1
3	1	33	0	0	⋯	30	1
4	1	30	0	0	⋯	58	1
..	..	..	..	..	⋯	..	..
..	..	..	..	..	⋯	..	..

Table 3.3: Features range value

No	Disease	Scale
0	Sex	Mail =1 Female = 0
1	Age	Min > 15 Mix < 70
2	Acute_ flaccid_ paralysis	POSITVE =1 NEGATIVE = 0
3	Cardiovascular_ Disease	POSITVE =1 NEGATIVE = 0
4	Shortness _ breath	POSITVE =1 NEGATIVE = 0
5	Diabetes	50 low 80 – 115 Normal 120-380 Diabetes
6	High Blood Pressure	Low <80 Normal 80-120 High stage1 140-159 High stage2 160 – higher High blood crisis 180
7	Cancer	POSITVE =1 NEGATIVE = 0
8	Liver disease	7-55 Unit/litter normal
9	AIDS	POSITVE =1 NEGATIVE = 0
10	Kidney disease	60-120 Normal 15-59 kidney disease 0 -15 kidney failure
11	Smoking	POSITVE =1 NEGATIVE = 0
12	Bronchial _respiration	Normal 120 % Mildly 60 – 79% Moderately 40 – 59% Severely <40%
13	No. of Pregnancy	Normal 36.4- 37.6 Un Normal > 37.6
14	Fever	POSITVE =1 NEGATIVE = 0
15	Wheezing	POSITVE =1 NEGATIVE = 0
16	Burning pharynx	POSITVE =1 NEGATIVE = 0
17	Nausea_ vomiting	POSITVE =1 NEGATIVE = 0

18	Chest_ bang	POSITVE =1 NEGATIVE = 0
19	Cough	POSITVE =1 NEGATIVE = 0
20	Headache	POSITVE =1 NEGATIVE = 0
21	cyanosis	POSITVE =1 NEGATIVE = 0
22	Runny nose	POSITVE =1 NEGATIVE = 0
23	Convulsions	POSITVE =1 NEGATIVE = 0
24	Frenzy confusion	POSITVE =1 NEGATIVE = 0
25	Slenderness	POSITVE =1 NEGATIVE = 0
26	Diarrhea	>= 80 Normal < 80 Un Normal
27	Chronic obstructive pulmonary diseases	POSITVE =1 NEGATIVE = 0
28	Outcome	Min > 15 Mix < 70

Data cleaning, also known as data purification, is the process of discovering and eliminating mistakes and inconsistencies in data in order to enhance data quality. Individual data groupings, such as files and databases, might have data quality issues, such as spelling mistakes, missing information, or other incorrect data during data input. The necessity for data cleansing grows dramatically when numerous data sources must be merged, such as in data warehouses, unified databases, or web-based global information systems. This is due to the fact that the sources frequently provide redundant data in various forms. A missing value in a data set diminishes prediction power and results in skewed estimations, resulting in an incorrect conclusion [62]. Standardizing multiple data formats and reducing duplicate information becomes required in order to offer access to correct and consistent data.

Figurers (3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 3.10, 3.11, 3.12, 3.13, 3.14) show the frequency of each data set feature.

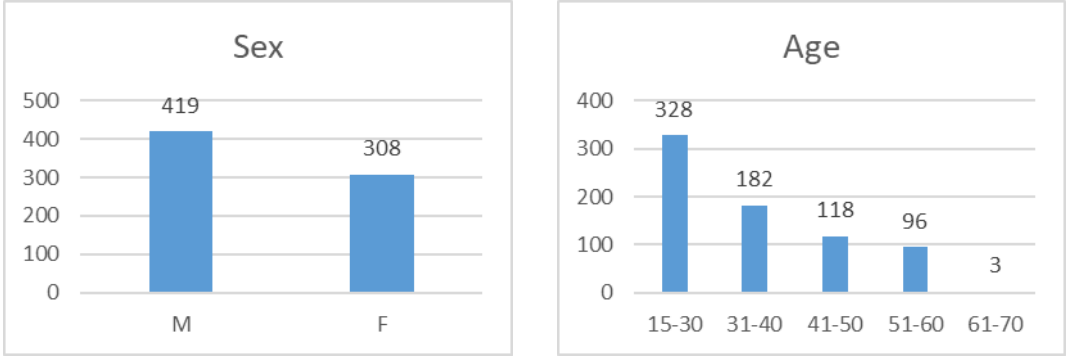
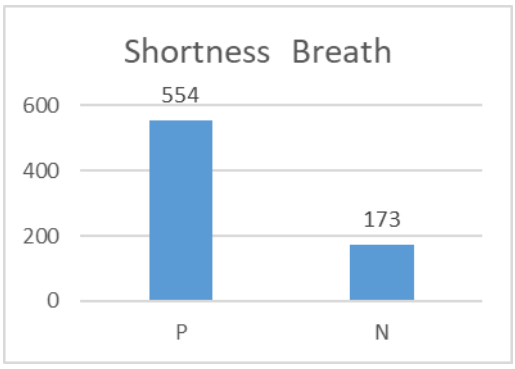


Figure 3.1: Sex and Age Frequency



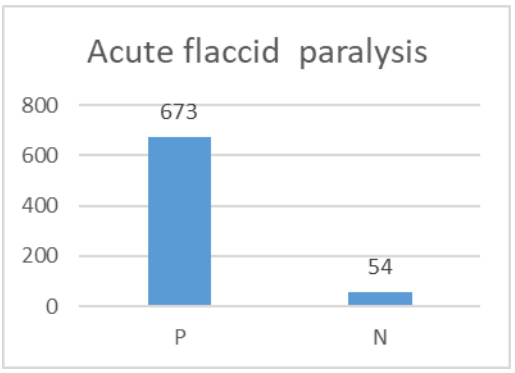
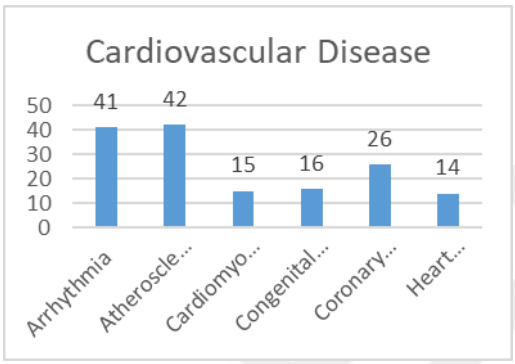


Figure 3.2: Shortness Breath and Acute flaccid paralysis frequency



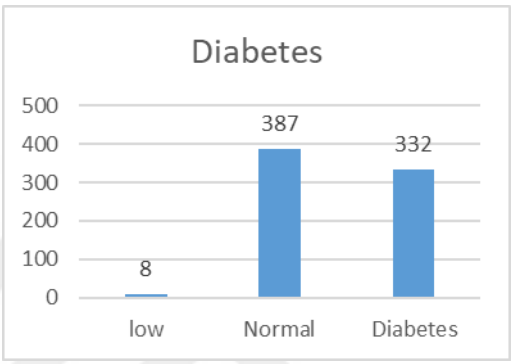
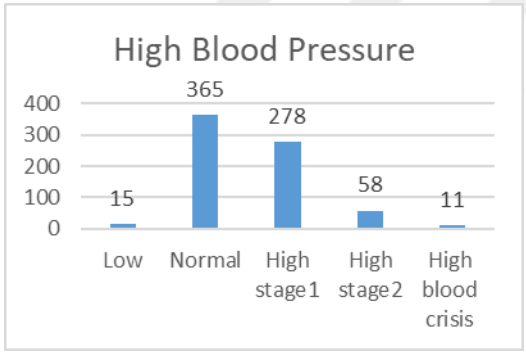


Figure 3.3: Cardiovascular and Diabetes frequency



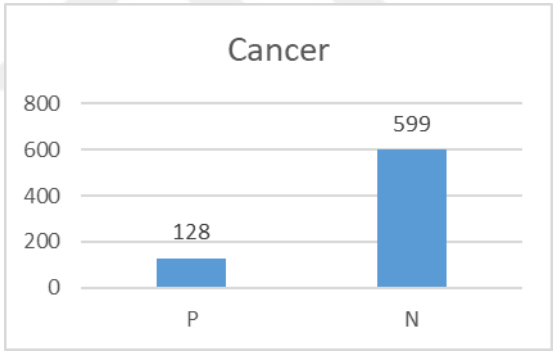
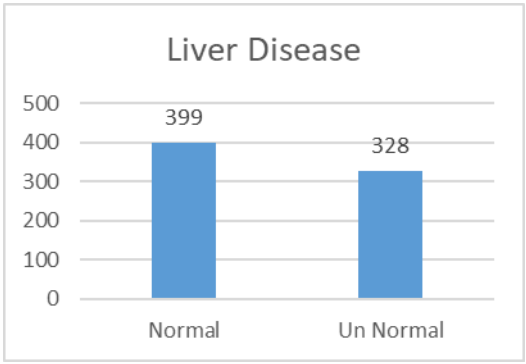


Figure 3.4: High Blood Pressure and Cancer frequency



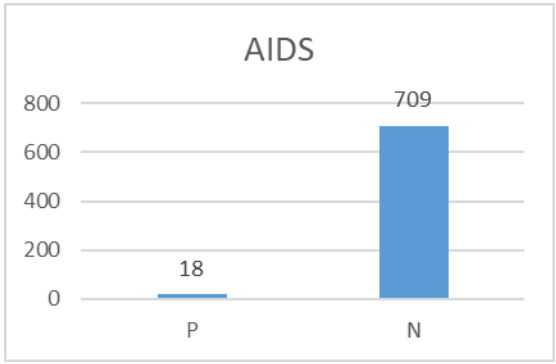
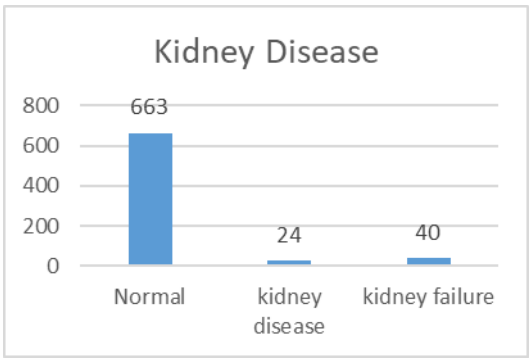


Figure 3.5: Liver and AIDS frequency



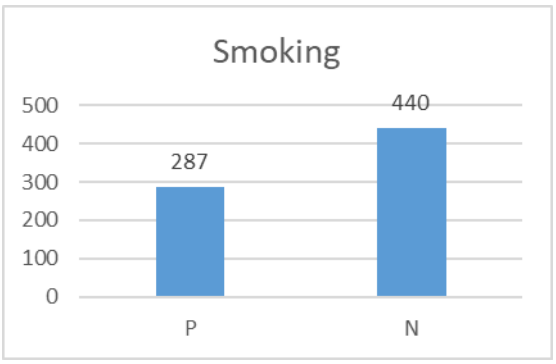
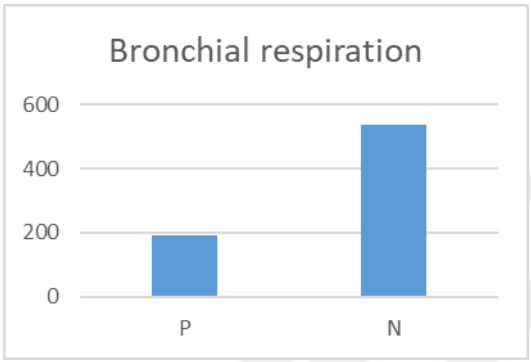


Figure 3.6: Kidney and Smoking frequency



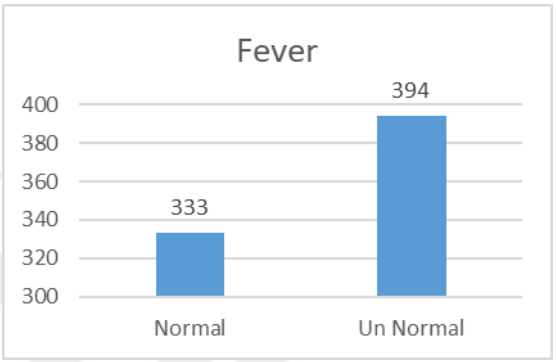
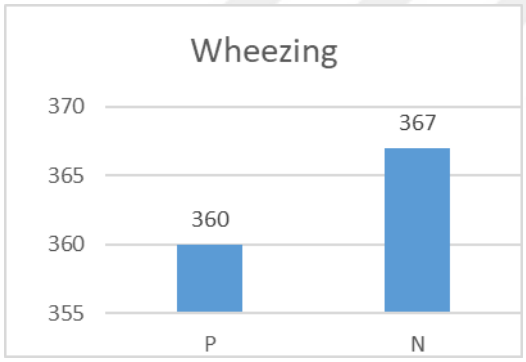


Figure 3.7: Bronchial respiration and Fever frequency



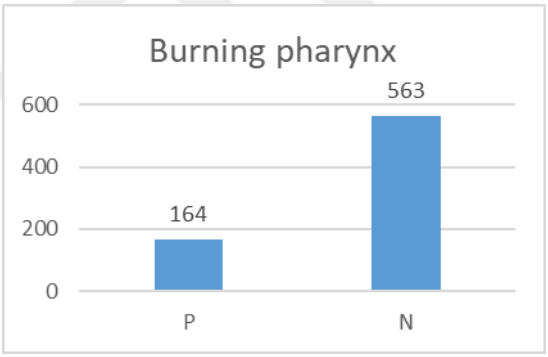
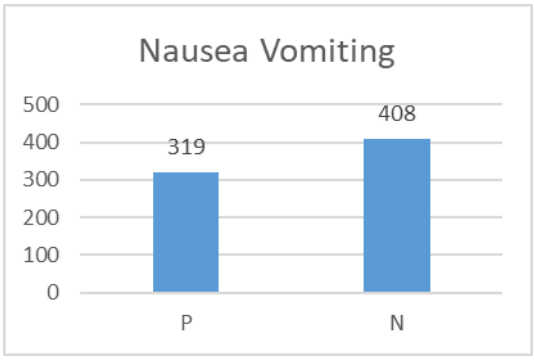


Figure 3.8: Wheezing and Burning pharynx frequency



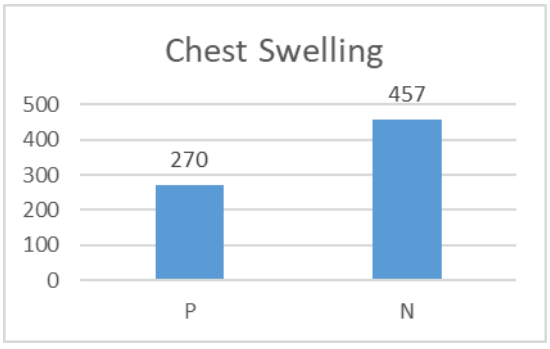
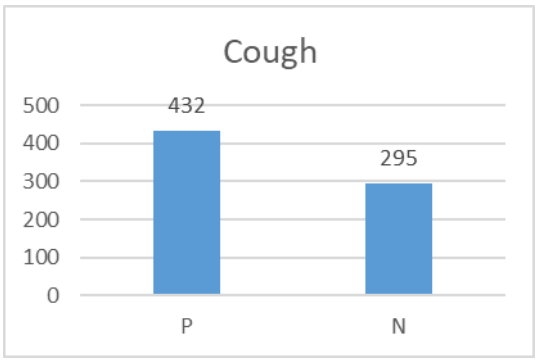


Figure 3.9: Nausea Vomiting and Chest Swelling frequency



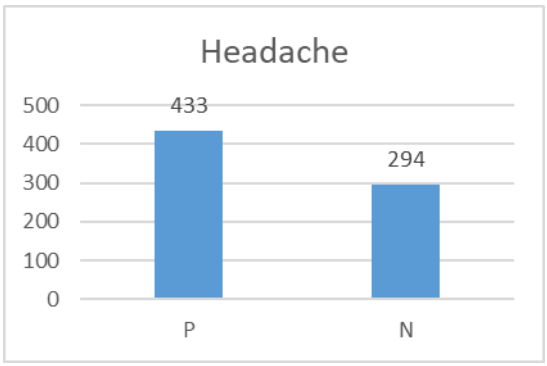
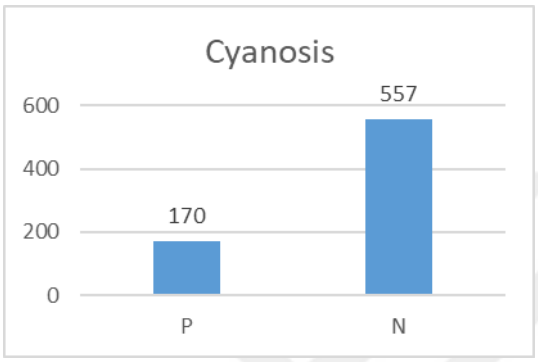


Figure 3.10: Cough and Headache frequency



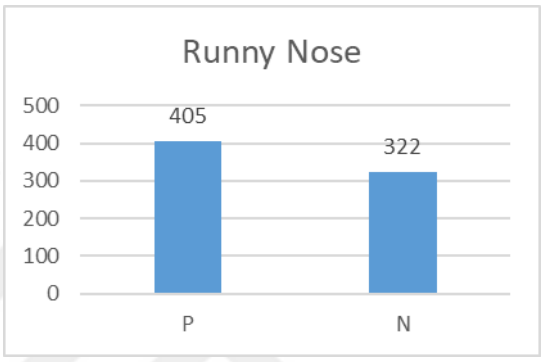
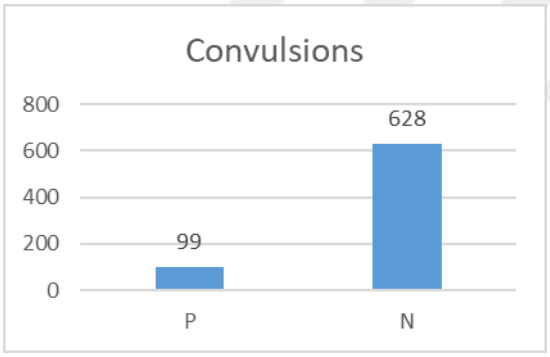


Figure 3.11: Cyanosis and Runny Nose frequency



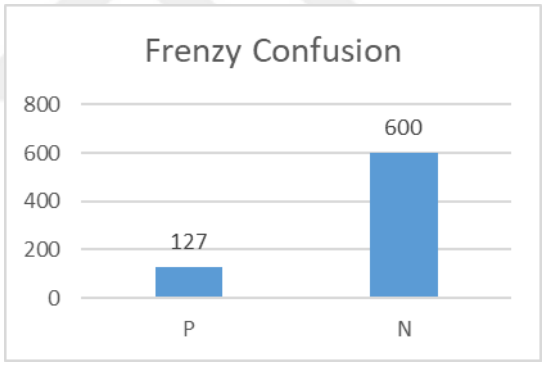
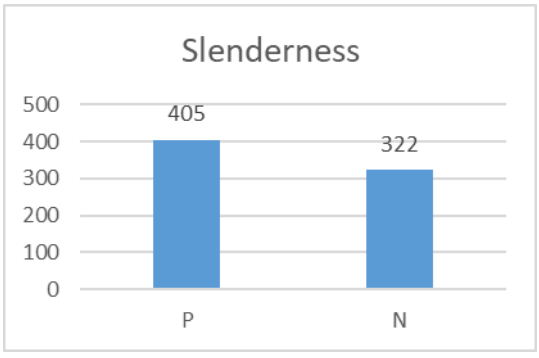


Figure 3.12: Convulsions and Frenzy Confusion frequency



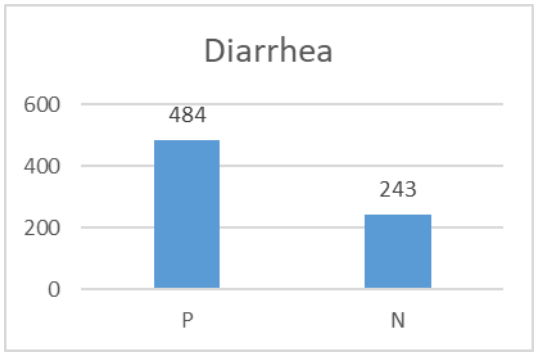


Figure 3.13: Slenderness and Diarrhea frequency

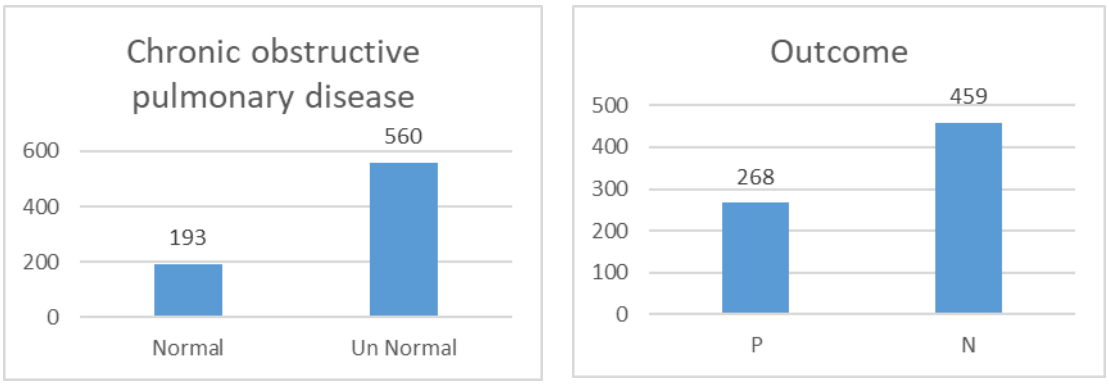


Figure 3.14: COPD and Outcome frequency

3.3. Dataset Features

CSV files which are often used to store tabular data in spreadsheets and databases, have been created from the dataset's original Microsoft Excel database.is a popular Open Data format widely used in a variety of domains for its simplicity and effectiveness in storing and disseminating data [63]. Tables of data (numbers and text) are stored in a CSV file, which is plain text. Each record in the file is a single line. Fields are separated by commas in each entry. Using a comma to separate the source fields gives this file format its name.

Files saved in binary format take up less disk space and are faster to read from the hard drive. The rising usage of deep learning is another major trend in machine learning, hence binary file formats are important.

3.4. Tools

Extracting data for prediction tasks was done using the Python programming language. The simplicity and adaptability of this language make it a popular choice among programmers. It is more widely used by the data science community because it provides a more standard programming language than some rivals [64] , machine learning [65], the Internet t of Things (IoT) [66].

An open-source, multiplatform widget toolkit called Tkinter is used to develop the integrated GUI, making it easier for the user to navigate and interact with the software. Python is used to implement the Tkinter interface [67]. AP.2, AP.3

Chapter 4

4. DESIGN OF THE MODEL

the first stage of the project, the existing standard data set was processed and produced. Inappropriate characteristics that may have a detrimental influence on prediction performance or overcomplicate computations are deleted from the data set at this phase. The data is subjected to basic filters. Then add up the remaining relevant characteristics. The major classification process begins in the second step, with algorithms being employed to categorize the current data set. Following the rating, various key performance indicators are calculated for each approach. The best classification approach was then presented after examination, analysis, and comparison. Figure 4.1 depicts a step-by-step summary of the suggested strategy.

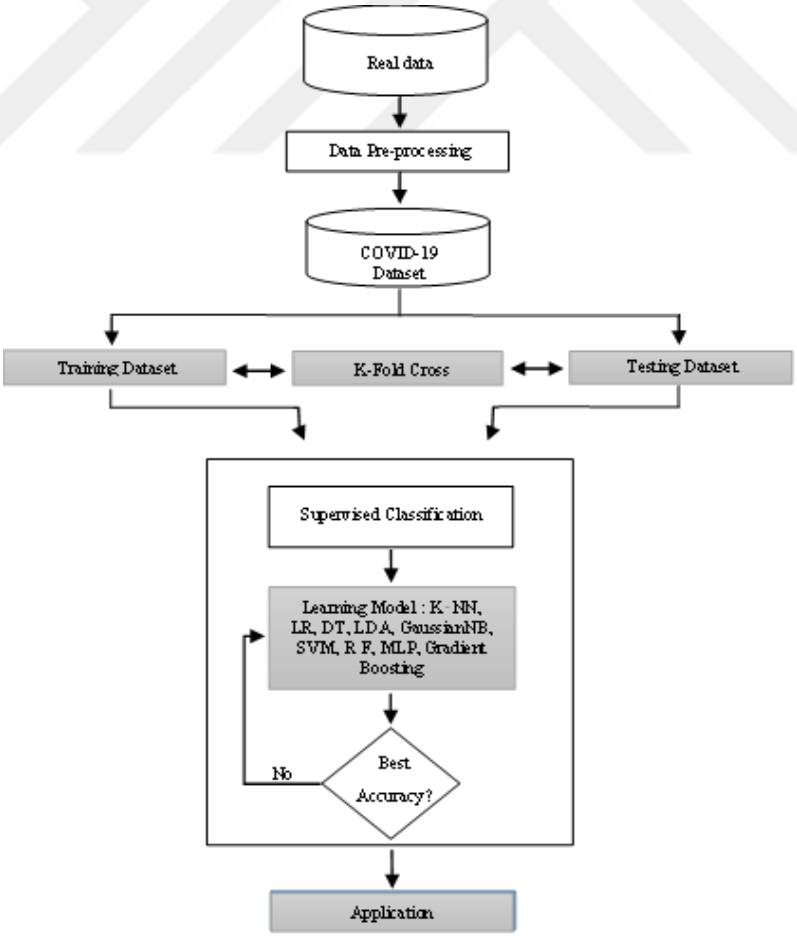


Figure 4.1: Sequential overview of proposed approach

4.1. Data Cleaning

Today's data is expanding by the day, and its sources are numerous, exposing it to a variety of issues that degrade its quality, such as a big number of missing data, data inconsistency, and a large amount of noise. And, of course, when the data quality is poor, the outcomes will suffer as a result. The ways of data cleansing will be highlighted. We demonstrate how important it is to clean the data before using it in the model. We make every effort to ensure that the data is as "perfect" as possible. Cleaning, verifying, and preparing data is a key aspect of developing a model, and a simple model with high-quality data is preferable to an accurate model with low-quality data.

4.2. Data Standardization

A dataset can contain a wide range of variables, and this is common in practice. One of the most important issues is that there is a wide range in the variables. If the original scale is used, the variables with a broad range may be given greater weight. Feature rescaling must be applied to independent variables or features of data during data pre-processing in order to address this issue.

It's common for the phrases normalization and standardization to be used interchangeably, yet they're actually rather different concepts. Feature Scaling is used to ensure that all characteristics are about equal in importance, making it easier for most ML algorithms to process.

4.3. Feature Selection Methods

4.3.1. Univariate Selection

For example, statistical tests may be performed to identify the most relevant attributes to the output variable. A variety of statistical tests may be used in conjunction with the SelectKBest class provided by the scikit-learn package to choose a certain number of features.

27 characteristics from the Covid-19 Range Prediction Dataset were selected using the chi2-statistical test for non-negative features. Table 4.1 shows the results of the analysis.

Table 4.1: Best features Range Prediction Dataset

No	Specs	Score
3	Cardiovascular Disease	9.568174e+01
26	Chronic obstructive pulmonary disease	3.088844e+01
22	Convulsions	6.775972e+00
5	Diabetes	5.510209e+00
1	Age	3.902637e+00
2	Acute flaccid paralysis	3.795240e+00
6	High Blood Pressure	3.463329e+00
7	Cancer	2.607880e+00
21	Runny Nose	2.276593e+00
17	Chest Swelling	2.245517e+00
8	Liver Disease	1.973479e+00
12	Bronchial respiration	1.724507e+00
19	Headache	1.580435e+00
4	Shortness Breath	1.430662e+00
25	Diarrhea	1.113335e+00
16	Nausea Vomiting	5.859210e-01
23	Frenzy Confusion	4.929181e-01
11	Smoking	4.049338e-01
20	Cyanosis	3.945871e-01
15	Burning pharynx	3.130355e-01
24	Slenderness	1.960187e-01
13	Fever	1.693114e-01
0	sex	6.618694e-02
18	Cough	5.618972e-02
14	Wheezing	1.005345e-03
9	AIDS	3.510525e-04
10	Kidney Disease	2.115232e-07
3	Cardiovascular Disease	9.568174e+01
26	Chronic obstructive pulmonary disease	3.088844e+01

4.3.2. Feature Importance

The feature priority attribute of the model may be used to determine the relative value of each feature in your dataset. The higher the score, the more essential or relevant a feature is to your output variable in terms of feature significance. For this dataset, we'll be utilizing an extra-tree-based classifier called Extra Tree Classifier to extract the most important characteristics as shown Figure 4.2.

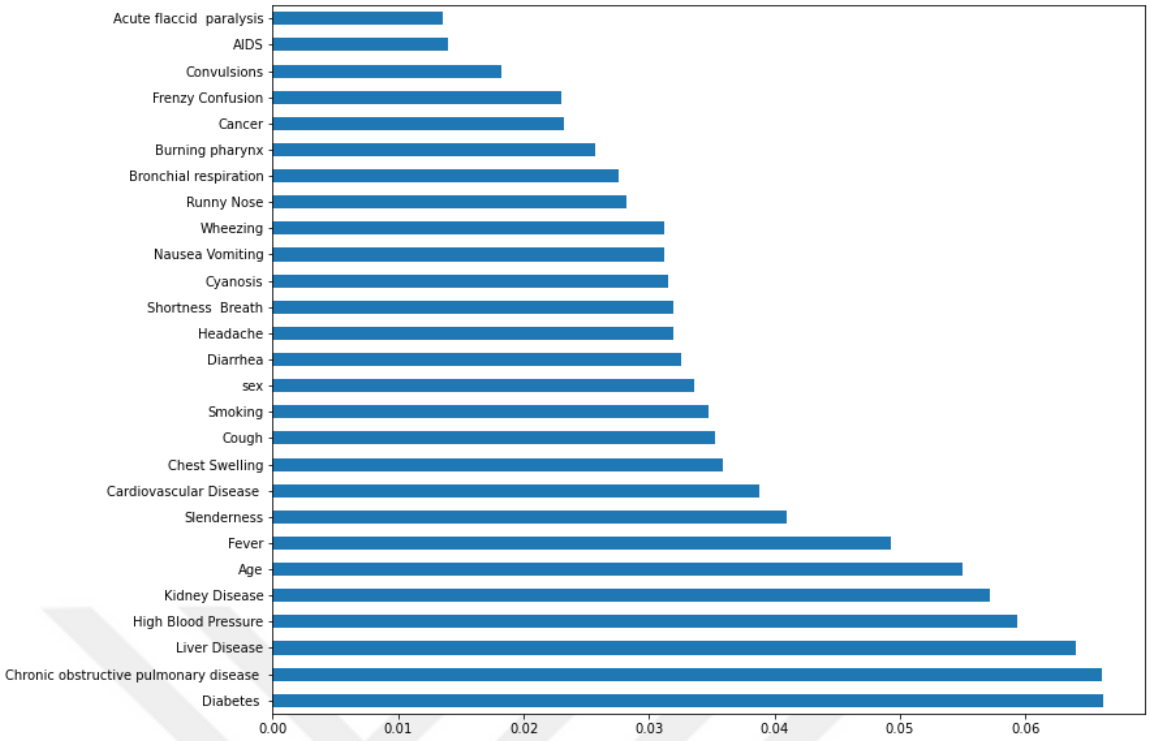


Figure 4.2: Ranking according to feature importance

4.3.3. Analysis and evaluation

Python programming language was employed in this research. It's a strong data mining language with a plethora of packages to help you construct data mining methods. On the data set of patients infected with the Covid-19, experimental results from the installation of classification algorithms (KNN, LR, DT, LDA, Gaussian NB, SVM, RF, MLP, Gradient Boosting) are evaluated and compared in this section. After comparing the findings, it was discovered that the Random Forest algorithm had the greatest classification accuracy, with 100 percent in training and 86.30 percent in testing data. The accuracy attained by nine classifiers is compared in Table 4.2 and Figure 4.3. Other key performance metrics such as sensitivity, specificity, F measure, accuracy, and ROC indices are evaluated to assess and compare the classification efficiency of the nine selected algorithms, despite the fact that accuracy is the most used measure of classification performance.

Table 4.2: classifiers model accuracy results

No.	Classifier	Training accuracy	Test accuracy
1	KNN	77.78 %	64.38 %
2	Gaussian NB	67.35 %	55.70%
3	SVC	91.14 %	76.25 %
4	Decision Tree	100.0 %	79.90%
5	LR	69.68%	62.55 %
6	Random Forest	100.0 %	86.30 %
7	Gradient Boosting	95.47%	81.27%
8	LDA	68.89 %	61.18 %
9	MLP	97.0 %	85.30 %

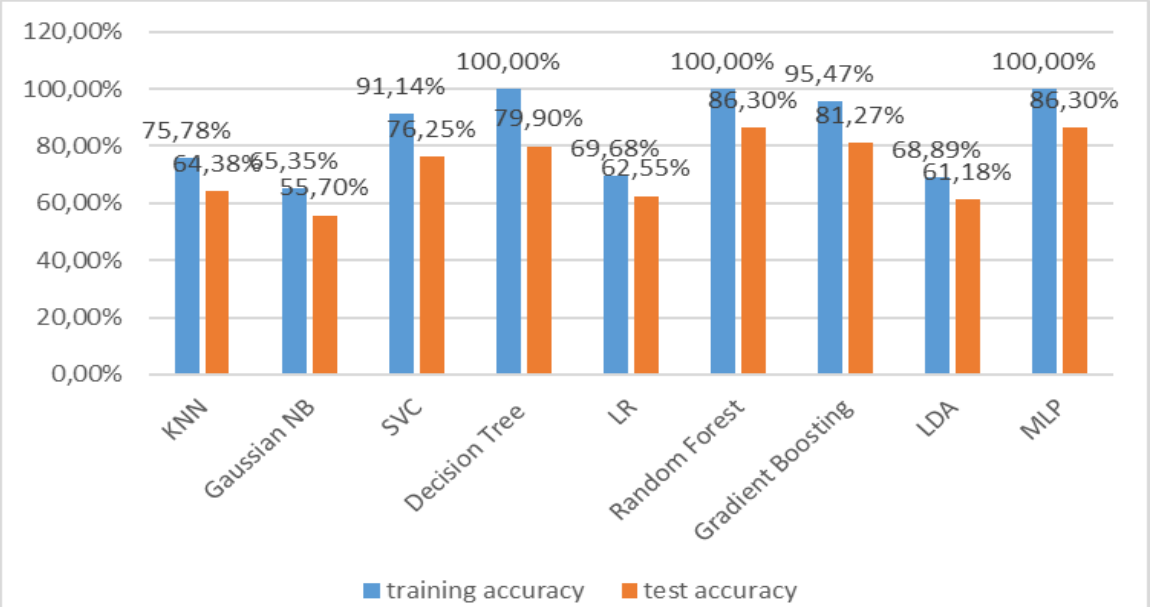


Figure 4.3: Comparison of accuracy achieved by nine classifiers

4.4 Interface

It is necessary to use graphical interfaces in order to make the application easy to use by the beneficiaries of the program in hospitals and health centers

As we mentioned earlier, the Tikinter library was used to design graphical interfaces to facilitate the entry of new data to predict the situation. A welcome interface AP.2 was added to introduce the program and a second window (the input window) AP.3 was added that contains the input fields in addition to four Buttons (print set) that prints the inputs in the form of a one-dimensional array in order to do the process of verifying the inputs and (Predicting) button to get the results of the negative or positive prediction and the (reset the filed) button to empty the fields to return to do another prediction process.

Chapter 5

5. DISCUSSION AND RESULT

5.1. Result

The first model was designed was for predicting the covid-19 infected using the KNN algorithm.

During the experiment, 70% of the data was used for training and 30% for testing, resulting in a training accuracy of 77.78% and testing accuracy of 64.38%, with precision of 0.77 for Uninfected and 0.76 for infected, recall of 0.90 for Uninfected and 0.54 for infected, F1 Score of 0.83 for Uninfected and 0.63 for infected, and Support of 138 for Uninfected and 81 for infected Table 5.1 show values the KNN classification report.

Table 5.1: Classification Report for KNN Algorithm

	precision	**recall**	**f1-score**	**support**
Uninfected	0.77	0.90	0.83	138
infected	0.76	0.54	0.63	81
accuracy			0.75	219
macro avg	0.76	0.72	0.73	219
weighted avg	0.77	0.77	0.76	219

Also, the KNN produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (124) of them were true positive i.e. they were predicted to be correctly infected and (14) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (37) of them were False Positive, meaning they were predicted to be wrongly non-infection while (44) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19. Figure 5.1 View the Confusion Matrix to KNN model.

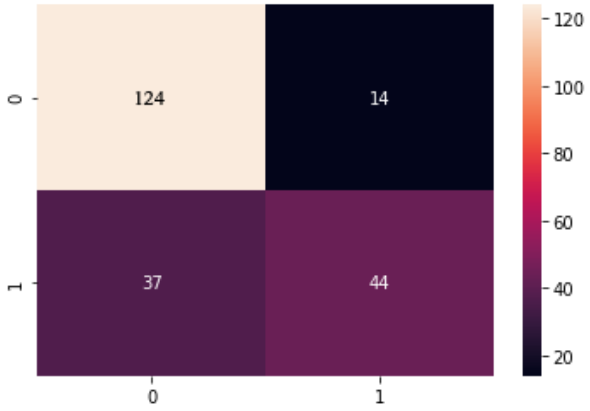


Figure 5.1: Confusion Matrix for KNN

The second model was designed was for predicting the covid-19 infected using the Gaussian NB algorithm.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to training accuracy of 67.35 % and testing accuracy of 55.70%, with precision of 0.76 for Uninfected and 0.54 for infected, Recall of 0.70 for Uninfected and 0.62 for infected, F1 Score of 0.72 for Uninfected and 0.58 for infected and Support of 138 for Uninfected and 81 for infected, Table 5.2 show values the Gaussian NB classification report.

Table 5.2: Classification Report for Gaussian NB Algorithm

	precision	**recall**	**f1-score**	**support**
Uninfected	0.76	0.70	0.72	138
infected	0.54	0.62	0.58	81
accuracy			0.65	219
macro avg	0.65	0.66	0.65	219
weighted avg	0.68	0.67	0.67	219

Also, the Gaussian NB produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (96) of them were true positive i.e. they were predicted to be correctly infected and (42) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (31) of them were False Positive, meaning they were predicted to be wrongly non-infection while (50) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19, Figure 5.2 View the Confusion Matrix to Gaussian NB model.

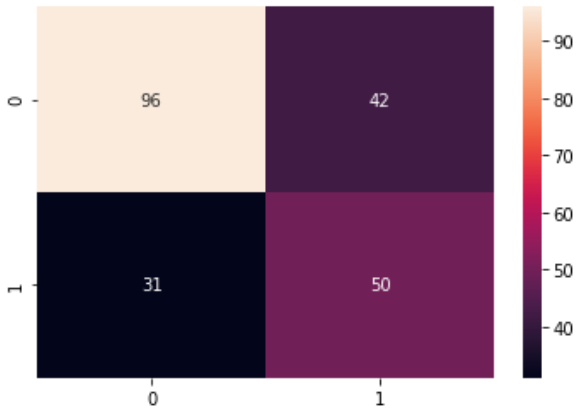


Figure 5.2: Confusion Matrix for Gaussian NB

The third model was designed was for predicting the covid-19 infected using the for Decision Tree algorithm.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to training accuracy of 100.0 % and testing accuracy of 79.90%, with precision of 0.100 for Uninfected and 0.100 for infected, Recall of 0.100 for Uninfected and 0.100 for infected, F1 Score of 0.100 for Uninfected and 0.100 for infected and Support of 138 for Uninfected and 81 for infected, Table 5.3 show values the Decision Tree classification report.

Table 5.3: Classification Report for Decision Tree Algorithm

	precision	**recall**	**f1-score**	**support**
Uninfected	1.00	1.00	1.00	138
infected	1.00	1.00	1.00	81
accuracy			1.00	219
macro avg	1.00	1.00	1.00	219
weighted avg	1.00	1.00	1.00	219

Also, the Decision Tree produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (138) of them were true positive i.e. they were predicted to be correctly infected and (0) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (0) of them were False Positive, meaning they were predicted to be wrongly non-infection while (81) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19, Figure 5.3 View the Confusion Matrix to Decision Tree model.

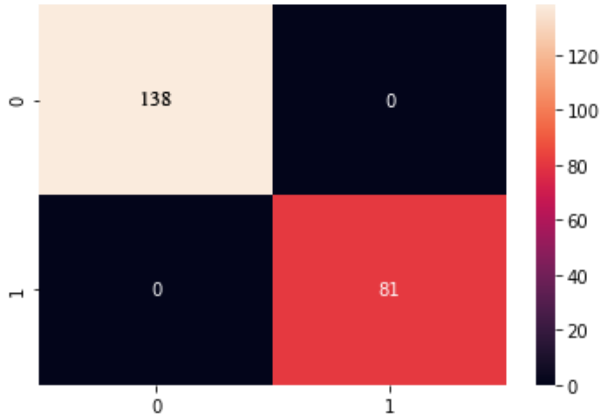


Figure 5.3: Confusion Matrix for Decision Tree

Forth model was designed was for predicting the covid-19 infected using the for SVM algorithm.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to training accuracy of 91.14 % and testing accuracy of 76.25 %, with precision of 0.83 for Uninfected and 0.98 for infected, Recall of 0.99 for Uninfected and 0.64 for infected, F1 Score of 0.90 for Uninfected and 0.78 for infected and Support of 138 for Uninfected and 81 for infected, Table 5.4 show values the SVM classification report.

Table 5.4: Classification Report for SVM Algorithm

	precision	**recall**	**f1-score**	**support**
Uninfected	0.83	0.99	0.90	138
infected	0.98	0.64	0.78	81
accuracy			0.91	219
macro avg	0.90	0.82	0.84	219
weighted avg	0.88	0.86	0.86	219

Also, the SVM produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (137) of them were true positive i.e. they were predicted to be correctly infected and (1) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (0) of them were False Positive, meaning they were predicted to be wrongly non-infection while (81) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19, Figure 5.4 View the Confusion Matrix to SVM model.

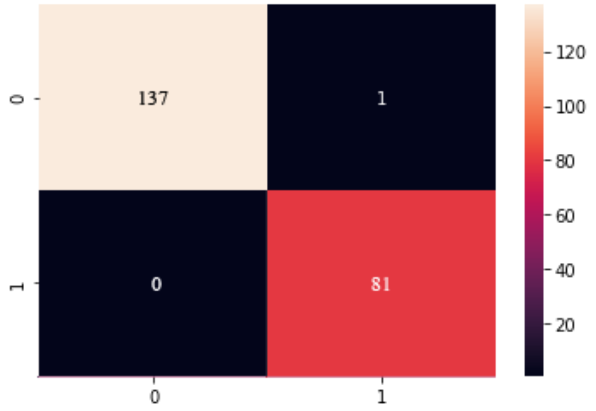


Figure 5.4: Confusion Matrix for SVM

The fifth model was designed was for predicting the covid-19 infected using the for Logistic Regression algorithm.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to training accuracy of 69.68 % and testing accuracy of 62.55 %, with precision of 0.71 for Uninfected and 0.64 for infected, Recall of 0.87 for Uninfected and 0.40 for infected, F1 Score of 0.78 for Uninfected and 0.49 for infected and Support of 138 for Uninfected and 81 for infected, Table 5.5 show values the Logistic Regression classification report.

Table 5.5: Classification Report for Logistic Regression Algorithm

	precision	**recall**	**f1-score**	**support**
Uninfected	0.71	0.87	0.78	138
infected	0.64	0.40	0.49	81
accuracy			0.69	219
macro avg	0.68	0.63	0.64	219
weighted avg	0.68	0.69	0.67	219

Also, the Logistic Regression produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (120) of them were true positive i.e. they were predicted to be correctly infected and (18) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (49) of them were False Positive, meaning they were predicted to be wrongly non-infection while (32) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19, Figure 5.5 View the Confusion Matrix to Logistic Regression model.

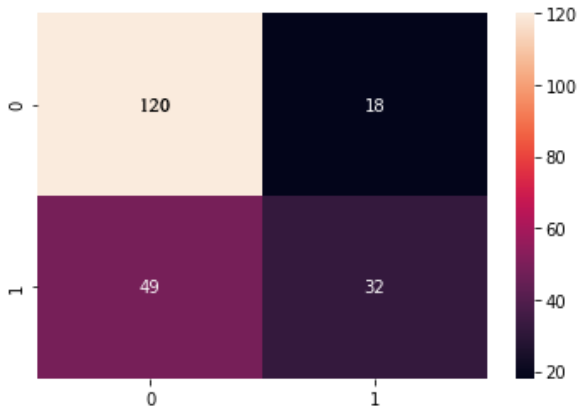


Figure 5.5: Confusion Matrix for Logistic Regression

The sixth model was designed was for predicting the covid-19 infected using the for Random Forest algorithm.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to training accuracy of 100.0 % and testing accuracy of 86.30 %, with precision of 0.100 for Uninfected and 0.100 for infected, Recall of 0.100 for Uninfected and 0.100 for infected, F1 Score of 0.100 for Uninfected and 0.100 for infected and Support of 138 for Uninfected and 81 for infected, Table 5.6 show values the Random Forest classification report.

Table 5.6: Classification Report for Random Forest Algorithm

	precision	**recall**	**f1-score**	**support**
Uninfected	1.00	1.00	1.00	138
infected	1.00	1.00	1.00	81
accuracy			1.00	219
macro avg	1.00	1.00	1.00	219
weighted avg	1.00	1.00	1.00	219

Also, the Random Forest produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (138) of them were true positive i.e. they were predicted to be correctly infected and (0) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (0) of them were False Positive, meaning they were predicted to be wrongly non-infection while (81) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19, Figure 5.6 View the Confusion Matrix to Random Forest model.

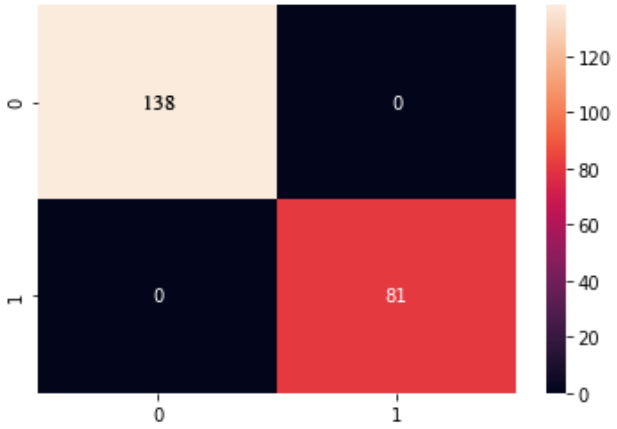


Figure 5.6: Confusion Matrix for Random Forest

The seventh model was designed was for predicting the covid-19 infected using the for Gradient Boosting algorithm.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to training accuracy of 95.47% and testing accuracy of 81.27%, with precision of 0.98 for Uninfected and 0.100 for infected, Recall of 0.100 for Uninfected and 0.96 for infected, F1 Score of 0.99 for Uninfected and 0.98 for infected and Support of 138 for Uninfected and 81 for infected, Table 5.7 show values the Gradient Boosting classification report.

Table 5.7: Classification Report for Gradient Boosting Algorithm

	precision	**recall**	**f1-score**	**support**
Uninfected	0.98	1.00	0.99	138
infected	1.00	0.96	0.98	81
accuracy			0.95	219
macro avg	0.99	0.98	0.99	219
weighted avg	0.99	0.99	0.99	219

Also, the Gradient Boosting produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (138) of them were true positive i.e. they were predicted to be correctly infected and (0) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (3) of them were False Positive, meaning they were predicted to be wrongly non-infection while (78) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19, Figure 5.7 View the Confusion Matrix to Gradient Boosting model.

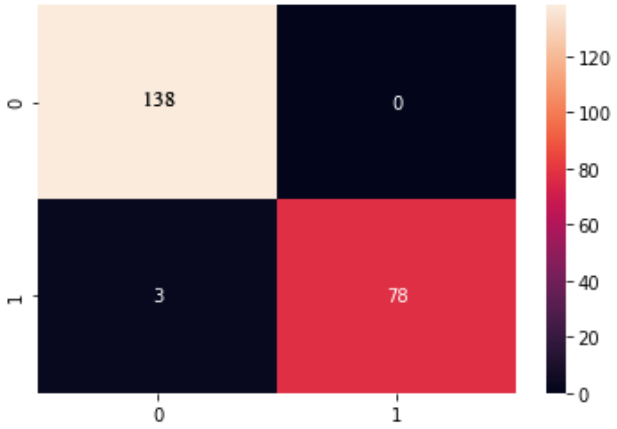


Figure 5.7: Confusion Matrix for Gradient Boosting

The Eighth model was designed was for predicting the covid-19 infected using the for Linear Discriminant Analysis algorithm.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to training accuracy of 68.89 % and testing accuracy of 61.18 %, with precision of 0.71 for Uninfected and 0.65 for infected, Recall of 0.87 for Uninfected and 0.41 for infected, F1 Score of 0.78 for Uninfected and 0.50 for infected and Support of 138 for Uninfected and 81 for infected, Table 5.8 show values the Linear Discriminant Analysis classification report.

Table 5.8: Classification Report for Linear Discriminant Analysis Algorithm

	precision	**recall**	**f1-score**	support
Uninfected	0.71	0.87	0.78	138
infected	0.65	0.41	0.50	81
accuracy			0.68	219
macro avg	0.68	0.64	0.64	219
weighted avg	0.69	0.70	0.68	219

Also, the Linear Discriminant Analysis produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (120) of them were true positive i.e. they were predicted to be correctly infected and (18) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (48) of them were False Positive, meaning they were predicted to be wrongly non-infection while (33) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19, Figure 5.8 View the Confusion Matrix to Linear Discriminant Analysis model.

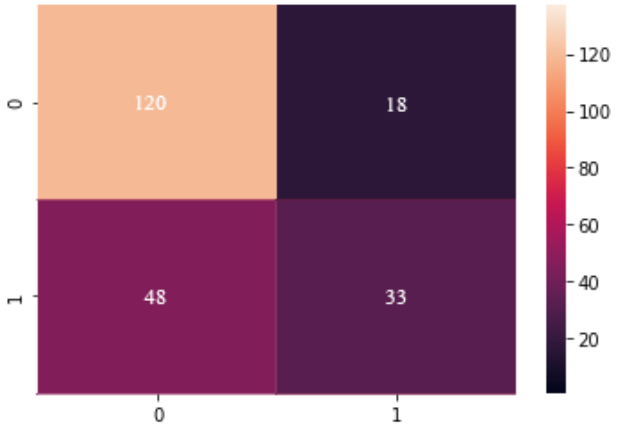


Figure 5.8: Confusion Matrix for Linear Discriminant Analysis

The ninth model was designed was for predicting the covid-19 infected using the for MLP algorithm.

During the experiment, 70% of the data was used for training while the remaining 30% was used for testing, which amounted to training accuracy of 97.0 % and testing accuracy of 85.30 %, with precision of 0.99 for Uninfected and 0.100 for infected, Recall of 0.100 for Uninfected and 0.98 for infected, F1 Score of 0.99 for Uninfected and 0.99 for infected and Support of 138 for Uninfected and 81 for infected, Table 5.9 show values the MLP classification report.

Table 5.9: Classification Report for MLP Algorithm

	precision	**recall**	**f1-score**	**support**
Uninfected	0.99	1.00	0.99	138
infected	1.00	0.98	0.99	81
accuracy			0.97	219
macro avg	0.99	0.99	0.99	219
weighted avg	0.99	0.99	0.99	219

Also, the MLP Algorithm produced a fair confusion matrix which showed that from the covid-19 dataset used for testing (137) of them were true positive i.e. they were predicted to be correctly infected and (1) of them were False Negative, meaning they were wrongly predicted to be non-infection while they are actually infected. It also indicates of the non-infection dataset used for testing (1) of them were False Positive, meaning they were predicted to be wrongly non-infection while (80) of the dataset were True Negative, meaning they were correctly predicted to be non-infection with covid-19, Figure 5.9 View the Confusion Matrix to MLP model.

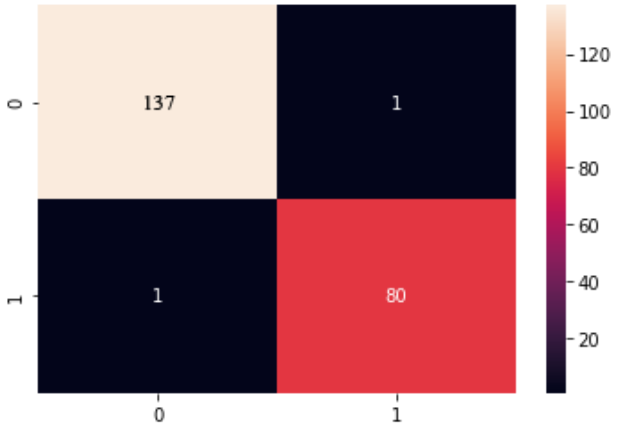


Figure 5.9: Confusion Matrix for MLP Algorithm

5.2. Conclusion

During this study, a comprehensive literature evaluation was undertaken to find the most appropriate COVID-19 prediction method. There was no conclusive proof that a single algorithm was the best method for predicting the future. To this end, the clinical data was used to train a variety of algorithms, including K-Nearest Neighbour (K-NN), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (Gaussian NB), and Support Vector Machine (SVM), as well as Random Forest, MLP Classifier, and Gradient Boosting. Each algorithm is trained on a different number of patients in order to evaluate its accuracy. The trained algorithms were evaluated using an accuracy performance metric. The Random Forest classifier (RF) was shown to be more accurate than other algorithms in a comparison study. Patients' COVID-19 risk was also evaluated using the taught algorithms.

5.3. Discussion

The results indicated that the nine algorithms were used when using algorithms and decision trees, and they had the highest accuracy in testing and training. The first algorithm is best for approximating values between testing and training. The study showed a relationship between the performance of the model and the advantage of COPD. Figure 4.2.

The use of this study and a number of previous studies of data mining techniques and machine learning algorithms demonstrates that building these models is indispensable in quickly and accurately supporting decisions for medical personnel during injury diagnosis,

which contributes to saving many lives, and also contributes to knowledge of the most important features that contribute to injury. The generalization of the results of this study is limited to increasing the number of samples and expanding it by adding laboratory tests to the case to obtain a comprehensive view of the behavior of the virus.

5.4. Recommendations for future work

Machine learning has a lot of potential in the healthcare industry. It is suggested that future study focus on calibrated and ensemble approaches that can more quickly and effectively address quirky situations than the present algorithms. An AI-based app may also be constructed utilizing sensors and characteristics to identify and diagnose illnesses.

REFERENCES

[1] Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., ... & Alshahrani, M. S. (2021). Applications of big data analytics to control Covid-19 pandemic. Sensors, 21(7), 2282.

[2] Jamal. Ali Khalil Dahshan, "The role of artificial intelligence in facing the Corona pandemic in the face of coexistence with it." The educational magazine of the Faculty of Education in Sohag 76.76 (2020): 1261-1286.

[3] Hafeez, A., Ahmad, S., Siddqui, S. A., Ahmad, M., & Mishra, S. (2020). A review of COVID-19 (Coronavirus Disease-2019) diagnosis, treatments and prevention. Ejmo, 4(2), 116-125.

[4] Agbehadji, I. E., Awuzie, B. O., Ngowi, A. B., & Millham, R. C. (2020). Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. International journal of environmental research and public health, 17(15), 5330.

[5] Patterson, G. E., McIntyre, K. M., Clough, H. E., & Rushton, J. (2021). Societal impacts of pandemics: Comparing COVID-19 with history to focus our response. Frontiers in public health, 206.

[6] Batty, M. (2013). Dialogues in Human Geography.

[7] Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health information science and systems, 2, 3. https://doi.org/10.1186/2047-2501-2-3

[8] Ismail, L., Materwala, H., Karduck, A. P., & Adem, A. (2020). Requirements of health data management systems for biomedical care and research: scoping review. Journal of medical Internet research, 22(7), e17508.

[9] Yagci, Mustafa & Güler, İnan. (2011). Muayenehane Otomasyonu Tasarim Aşamalari Ve Gerçekleştirilmesi. Selçuk-Teknik Dergisi. 10. 241-258.

[10] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018, September). Prediction of diabetes using machine learning algorithms in healthcare. In 2018 24th international conference on automation and computing (ICAC) (pp. 1-6). IEEE.

[11] Laatifi, M., Douzi, S., Bouklouz, A., Ezzine, H., Jaafari, J., Zaid, Y., ... & Naciri, M. (2022). Machine learning approaches in Covid-19 severity risk prediction in Morocco. Journal of big Data, 9(1), 1-21.

[12] Pourhomayoun, M., & Shakibi, M. Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. medRxiv 2020. Google Scholar.

[13] Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., ... & Huang, Y. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials & Continua, 63(1), 537-551.

[14] Yan, L., Zhang, H., & Goncalves, J. An interpretable mortality prediction model for COVID-19 patients. Nat. Mach. Intell. 2, 283–288 (2020).

[15] Li, W. T., Ma, J., Shende, N., Castaneda, G., Chakladar, J., Tsai, J. C., ... & Ongkeko, W. M. (2020). Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC medical informatics and decision making, 20(1), 1-13.

[16] Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. Journal of medical systems, 44(8), 1-12.

[17] Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., ... & Li, S. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. MedRxiv.

[18] Tschoellitsch, T., Dünser, M., Böck, C., Schwarzbauer, K., & Meier, J. (2021). Machine learning prediction of SARS-CoV-2 polymerase chain reaction results with routine blood tests. Laboratory medicine, 52(2), 146-149.

[19] Ur, A., Rahman, S., Naseer, I., Majeed, R., Musleh, D., Gollapalli, M. A. S., ... & Khan, M. A. (2021). Supervised machine learning-based prediction of covid-19. Computers, Materials and Continua, 21-34.

[20] Ardabili, S., Mosavi, A., Band, S. S., & Varkonyi-Koczy, A. R. (2020, November). Coronavirus disease (COVID-19) global prediction using hybrid artificial intelligence method of ANN trained with Grey Wolf optimizer. In 2020 IEEE 3rd International Conference and Workshop in Óbuda on Electrical and Power Engineering (CANDO-EPE) (pp. 000251-000254). IEEE.

[21] Sun, L., Song, F., Shi, N., Liu, F., Li, S., Li, P., ... & Shi, Y. (2020). Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19. Journal of Clinical Virology, 128, 104431.

[22] Al Mamlook, R., Al-Mawee, W., Alden, A. Y. Q., Alsheakh, H., & Bzizi, H. (2021, February). Evaluation of Machine Learning Models to Forecast COVID-19 Relying on Laboratory Outcomes Characteristics in Children. In IOP Conference Series: Materials Science and Engineering (Vol. 1094, No. 1, p. 012072). IOP Publishing.

[23] Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., ... & Huang, Y. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. Computers, Materials & Continua, 63(1), 537-551.

[24] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD (Vol. 96, pp. 82-88).

[25] Chan, K. C. C., Wong, A. K. C., Piatesky–Shapiro, G., & Frawley, W. J. (1991). Knowledge Discovery in Databases.

[26] Akhila, G. S., Madhu, G. D., Madhu, M. H., & Pooja, M. H. (2014). Comparative study of classification algorithms using data mining. In Discovery Science (Vol. 9, No. 20, pp. 17-21).

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

[28] Osman, A. S. (2019). Data mining techniques.

[29] Berry, M. J., & Linoff, G. S. (2004). Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons.

[30] Sisodia, D., Singh, L., Sisodia, S., & Saxena, K. (2012). Clustering techniques: a brief survey of different clustering algorithms. International Journal of Latest Trends in Engineering and Technology (IJLTET), 1(3), 82-87.

[31] Rao, I. R. (2003, December). Data mining and clustering techniques. In DRTC Workshop on Semantic Web (Vol. 8).

[32] Govindarajan, M. (2013). Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm. International Journal of Advanced Computer Research, 3(4), 139.

[33] IBM TEAM. (2021, July 7). Sequential patterns and rules. IBM. Retrieved from https://www.ibm.com/docs/en/db2oc?topic=procedures-sequential-patterns

[34] Zhang, Shichao & Zhang, Chengqi & Yang, Qiang. (2003). Data Preparation for Data Mining. Applied Artificial Intelligence. 17. 375-381. 10.1080/713827180.

[35] Yse, D. L. (2020, May 10). How to prepare your data. Medium. Retrieved May 5, 2022, from https://towardsdatascience.com/the-basics-of-data-prep-7bb5f3af77ac

[36] Agrawal, M. M. (2021). Understanding different techniques of data cleaning and different operations involved. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(12), 3820-3826.

[37] Irawan, Y. (2019). Implementation of Data Mining for Determining Majors Using K-Means Algorithm in Students of SMA Negeri 1 Pangkalan Kerinci. Journal of Applied Engineering and Technological Science (JAETS), 1(1), 17-29.

[38] Minewiskan. (n.d.). Data Mining Concepts. Microsoft Docs. Retrieved June 14, 2022, from https://docs.microsoft.com/en-us/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions#DeployingandUpdatingModels

[39] SNIJDERS, Chris, Uwe MATZAT, Ulf-Dietrich REIPS, 2012. "Big Data": Big Gaps of Knowledge in the Field of Internet Science. In: International Journal of Internet Science. 7(1), pp. 1-5. eISSN 1662-5544

[40] De Oliveira, M. F., & Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. IEEE transactions on visualization and computer graphics, 9(3), 378-394.

[41] Data Vedas, & *. (2018, February 1). Data Exploration and Preparation Theory. Data Vedas. Retrieved May 7, 2022, from http://www.datavedas.com/univariate-bivariate-analysis/

[42] Paradis, E., O'Brien, B., Nimmon, L., Bandiera, G., & Martimianakis, M. A. (2016). Design: Selection of data collection methods. Journal of graduate medical education, 8(2), 263-264.

[43] Axinn, W. G., & Pearce, L. D. (2006). Mixed method data collection strategies. Cambridge University Press.

[44] Xing, W., & Bei, Y. (2019). Medical health big data classification based on KNN classification algorithm. IEEE Access, 8, 28808-28819.

[45] Li, K., Luo, X., & Jin, M. (2010). Semi-supervised Learning for SVM-KNN. J. Comput., 5(5), 671-678.

[46] Mythili, T., Mukherji, D., Padalia, N., & Naidu, A. (2013). A heart disease prediction model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications, 68(16).

[47] Park, H. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. Journal of Korean Academy of Nursing, 43(2), 154-164.

[48] Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research (IJSR), 5(4), 2094-2097.

[49] Park, C. H., & Park, H. (2008). A comparison of generalized linear discriminant analysis algorithms. Pattern Recognition, 41(3), 1083-1097.

[50] Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and Tensor Flow: Concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.".

[51] Saritas, M. M., & Yasar, A. (2019). Performance analysis of ANN and Naive Bayes classification algorithm for data classification. International Journal of Intelligent Systems and Applications in Engineering, 7(2), 88-91.

[52] Hooshmand, A. (2020). Naive Bayesian Machine Learning to Diagnose Breast Cancer.

[53] Taylor, K. L. Oracle Data Mining Concepts, 11g Release 2 (11.2) E16808-06.

[54] Gandhi, R. (2018, June 7). Support Vector Machine — Introduction to Machine Learning Algorithm.

[55] deep, A. (2021, November 10). Random Forest classifier using Scikit-learn. GeeksforGeeks. Retrieved December 1, 2021, from https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/.

[56] Donges, N. (n.d.). A complete guide to the random forest algorithm. Built In. Retrieved December 1, 2021, https://builtin.com/data-science/random-forest-algorithm

[57] Scoralick, J. P., Iwashima, G. C., Colugnati, F. A., Goliatt, L., & Capriles, P. V. (2020, December). A Extreme Gradient Boosting Classifier for Predicting Chronic Kidney Disease Stages. In International Conference on Intelligent Systems Design and Applications (pp. 901-910). Springer, Cham.

[58] H. M. (2021, November 25). What is gradient boosting, and how is it different from AdaBoost? Retrieved December 11, 2021, from https://www.mygreatlearning.com/blog/gradient-boosting/

[59] Susmaga, R. (2004). Confusion matrix visualization. In Intelligent information processing and web mining (pp. 107-116). Springer, Berlin, Heidelberg.

[60] Kharwal, A. (2021, July 7). Classification report in Machine Learning. Data Science | Machine Learning | Python | C++ | Coding | Programming | JavaScript. Retrieved May 7, 2022, from https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/

[61] Kang, H. (2013). The prevention and handling of the missing data. Korean journal of anesthesiology, 64(5), 402-406.

[62] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), 3-13.

[63] Christodoulakis, C., Munson, E. B., Gabel, M., Brown, A. D., & Miller, R. J. (2020). Pytheas: pattern-based table discovery in CSV files. Proceedings of the VLDB Endowment, 13(12), 2075-2089.

[64] Hunt, J. (2021). A beginners guide to Python 3 programming.

[65] Raschka, S., & Mirjalili, V. (2017). Python machine learning: Machine learning and deep learning with python. Scikit-Learn, and Tensor Flow. Second edition ed, 3.

[66] Arora, K. (2021, March 9). Understanding The Role of Python In IoT Development.

[67] Chaudhary, B. (2013). Tkinter GUI Application Development HOTSHOT. Packt Publishing.

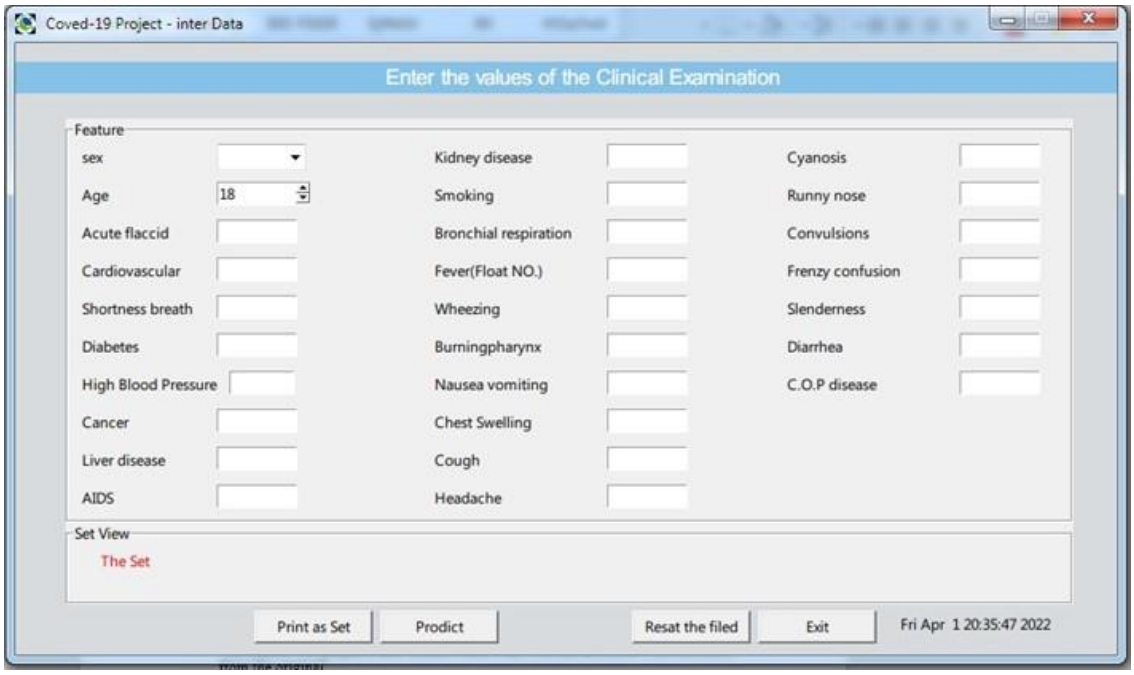
APPENDIX

AP 1: Form Diagnostics of Clinical Signs of Covid-19

An inquiry form for the case of suspected SARS - Corona (Covid_19)											
J	YEAR	Month	Day	Detection date		Epidemiological number					
	Mother's name					Patient name					
						Patient JOB					
				phone number		Age					
Sex		house	alley	Locality	Area	City	Address				
F	M										
YEAR		Month		Day		Date of notification					
Doctor's name						The health institution that was informed of the case					
Year		Month		day		date on which symptoms and signs began					
Year		Month		day		Date of hospitalization					
Recovered ☐ / Not Recovered ☐ /Death ☐ /Unknown ☐						fate of the patient					
Symptoms and indications											
N	Y	Nausea, vomiting	N	Y	Burning pharynx	N	Y	Wheezing	N	Y	Fever above 38 degrees
N	Y	a headache	N	Y	Bronchial respiration	N	Y	Cough	N	Y	Chest bang
N	Y	Frenzy , confusion	N	Y	Convulsions	N	Y	Runny nose	N	Y	cyanosis
N	Y	Other	N	Y	diarrhea	N	Y	slenderness	N	Y	Shortness of breath or difficulty breathing
Does he suffer from acute flaccid paralysis: yes ( ) no ( ) if the answer is (yes) / date of paralysis / site of paralysis ( )											
Smoking : smoker ( ) non-smoker ( )											
During the 14 days before symptoms start											
Has the patient traveled : Yes ( ) No ( ) / Country: City:											
Does the patient come into contact with animals: Yes ( ) No ( )											
Does the patient have contact with a suspected case of the emerging corona virus ( ): confirmed case ( ): respiratory patterns ( )											
Risk factors											
Pregnancy : Duration of Pregnancy : ☐ Cardiovascular Disease : ☐ High Blood Pressure : ☐ Diabetes : ☐ Cancer ☐											
Liver disease : ☐ After birth / up to 6 weeks : ☐ Impaired immunity, including AIDS : ☐ Kidney disease : ☐											
Chronic lung disease : ☐ Others : ☐											

AP 2: welcome interface of the designed model

AP 3: Application data enter interface



RESUME

Personal İnformation	
Name	Ruslan Salim Naseef
Place of Birth	
Date of Birth	
Nationality	❒ T.C. ☑ Diğer:

Education İnformation	
License	
University	Al-Mammon University College
Department	Computer Science
Graduation Year	2003

Degree	
University	Kırşehir Ahi Evran University
Institute Name	Institute Of Science
Department	Advanced Technology Department
Program	Advanced Technologies Master's with Thesis
Graduation Year	2022