

**WEKA VS RAPID MINER: A PERFORMANCE ANALYSIS OF DATA MINING
CLASSIFICATION TECHNIQUES ON HEALTH DATA**

BY

Tonima Islam
ID: 191-15-2686

AND

Ilma Akter Sharna
ID: 191-15-2681

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Amatul Bushra
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Nadira Anjum Nipa
Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

FEBRUARY 2023

APPROVAL

This Project titled “**Weka Vs Rapid miner: A Performance Analysis of Data Mining Classification Techniques on health data**”, submitted by Tonima Islam, ID: 191-15-2686, and Ilma Akter Sharna, ID: 191-15-2681 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 1st February 2023.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Tania Khatun

Tania Khatun (TK)
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Lamia Rukhsara 1.2.23

Ms. Lamia Rukhsara (LR)
Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Mohammad Shorif Uddin

Dr. Mohammad Shorif Uddin
Professor

Department of Computer Science and Engineering
Jahangirnagar University

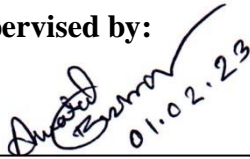
External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Amatul Bushra, Assistant Professor, Department of CSE** Daffodil International University.

We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Amatul Bushra

Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Nadira Anjum Nipa

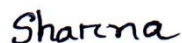
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Tonima Islam

ID: 191-15-2686
Department of CSE
Daffodil International University



Ilma Akter Sharna

ID: 191-15-2681
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First of all, we want to render our gratitude to the Almighty Allah for the enormous blessing that makes us able to complete the final thesis successfully.

We are really grateful and express our earnest indebtedness to **Amatul Bushra, Assistant Professor**, Department of CSE Daffodil International University, Dhaka, Bangladesh. The Profound Knowledge & intense interest of our supervisor in the field of “*Machine Learning & Data Mining*” made our way to carry out this thesis very smoothly. His remarkable patience and dedication, scholarly guidance, continual encouragement, vigorous motivation, direct and fair supervision, constructive criticism, valuable advice, and great endurance during reading many inferior drafts and correcting the work to make it unique paved the way of work very smooth and ended with a great result.

We would like to express our gratitude wholeheartedly to **Prof. Dr. Touhid Bhuiyan**, Professor, and Head, Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank the fellow Daffodil International University student, who participated in this discussion during the completion of this work.

We would like to express our immense thanks to the Different food application for visible user original reviews as a result we collected raw data to make our work possible.

We would also like to thank the people who provide the data done by us to collect the market real information.

Finally, we must acknowledge with due respect the constant support and passion of our parents and family members.

ABSTRACT

The term "data mining" is helpful since it simplifies the process of looking through and analyzing a large amount of data to find information that is important and confidential. Academics have recently shown a rising interest in the management of healthcare statistics using data mining techniques. The authors of this study used data mining techniques to attempt to categorize three distinct datasets related to breast cancer, diabetes, and renal disease using weka and fast miner. A variety of classification methods, including Decision Tree, K-Nearest Neighbors, Naive Bayes, Random Forest, and Support Vector, are used in the performance evaluation. Every categorization technique is implemented utilizing well-known data mining and tools for knowledge discovery like Rapid miner and weka. Weka tools outperform Rapid Miner tools in terms of accuracy across all datasets. Data mining is an appropriate word that makes it easier to explore and analyze huge amounts of data in search of private and useful information. Data mining approaches have recently piqued the interest of researchers who want to handle healthcare statistics. In this study, weka and Rapid miner were used to attempt classification using data mining techniques on three datasets (breast cancer, diabetes, and kidney). The performance of various classification techniques, including Decision Tree, K-Nearest Neighbors, Naive Bayes, Random Forest, and Support Vectors, is compared. Weka and Rapid miner, two popular data mining and knowledge discovery tools, are used for every categorization approach. When compared to Rapid Miner tools, Weka tools have been demonstrated to work with superior accuracy overall.

TABLE OF CONTENTS

CONTENTS	PAGE
Approval Page	ii
Declaration	iii
Acknowledgment	iv
Abstract	v
List of Figure	viii
List of Table	ix

CHAPTER

CHAPTER 1: INTRODUCTION	PAGE NO.
	1-5
1.1 Introduction	1-2
1.2 Motivation	2
1.3 Rationale of Study	3
1.4 Research Questions	3
1.5 Expected Output	4
1.6 Project Management and Finance	4-5
1.7 Report Layout	5
CHAPTER 2: BACKGROUND	6-9
2.1 Preliminaries	6
2.2 Related Works	6-8
2.3 Comparative Analysis and Summary	8
2.4 Scope of the Problem	8-9
2.5 Challenges	9

CHAPTER 3: RESEARCH METHODOLOGY	10-12
3.1 Research Subject and Instrumentation	10
3.2 Data Collection Procedure/Dataset Utilize	10
3.3 Statistical Analysis	10
3.4 Proposed Methodology/Applied Mechanism	11
3.5 Implementation Requirements	12
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	13-19
4.1 Experimental Setup	13-14
4.2 Experimental Results & Analysis	15-19
4.3 Discussion	19
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY	20-21
5.1 Impact on Society	20
5.2 Impact on the Environment	20-21
5.3 Ethical Aspects	21
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH	22-23
6.1 Summary of the Study	22
6.2 Conclusions	22
6.3 Implication for Further Study	23

REFERENCES	24-25
APPENDIX	26
PLAGIARISM REPORT	27

LIST OF FIGURES

FIGURES	PAGE NO.
Figure 3.3: Methodology of Research	12
Figure 4.A1: K-fold cross-validation	13
Figure 4.A.2: Confusion Matrix	14

LIST OF TABLE

TABLE	PAGE NO
Table1.5: The percentage of weka and rapid miner tools	4
Table 3.2. Dataset Description	11
Table 4.2: Best accuracy of the result	19

CHAPTER 1

INTRODUCTION

1.1 Introduction

A fascinating and crucial component of public health is the significance of maintaining excellent physical health in humans. Unhealthy lifestyle choices increase our risk of contracting kidney disease, diabetes, breast, and other organ malignancies, as well as other illnesses like kidney and liver disease. Millions of individuals are currently suffering as a result of these widespread illnesses, each of which brings with it a vast range of issues.

First, when a person's breast cells proliferate unchecked, breast cancer occurs. Although men can develop breast cancer, women typically do. Second, diabetes arises when the body is unable to utilize the hormone insulin normally, whose main purpose is to regulate the amount of sugar in the blood. Thirdly, kidney disease harms your kidneys and lessens their capacity to keep you healthy by removing waste products from your blood. Data mining techniques are capable of early illness detection. Early detection of this type of disease is impossible because there are so many patients with it. This is because there aren't enough doctors to go around. The mortality rate can be lowered by early detection. Numerous tools can be used to implement data mining methods. Researchers can swiftly adopt vital data mining technologies with the use of these tools. The greatest tools for putting data mining and machine learning algorithms into practice are Weka and Rapid miner. To obtain the classification performance in this study, use the Decision Tree, K-Nearest Neighbors, Naive Bayes, Random Forest, and Support Vector algorithms in Weka and Rapid Miner.

Weka:

You can examine a dataset and choose several properties for the x- and y-axes in Weka's Visualize panel, ideally numeric ones. The representation of instances is as a point, with various classes represented by various colors. You can draw a rectangle and narrow the dataset's attention to the points inside.

Rapid Miner:

Rapid Miner can ensure that your data science projects are a resounding success. For organizations looking to quicken the pace of transformation, Rapid Miner is the

enterprise-ready data science platform that amplifies the combined impact of your people, skills, and data.

We conducted a performance examination of various data mining classification approaches on healthcare data for this paper. This research assisted in identifying the most accurate data mining categorization approach for the given dataset. We have trained the Kaggle public healthcare dataset for this. Decision tree, K-Nearest Neighbors, Naive Bayes, Random Forest, and Support Vector Algorithm are the categorization methods under scrutiny. Based on their accuracy, these strategies are evaluated for performance. Future researchers will benefit from this study by learning the best data mining categorization and obtaining effective findings.

1.2 Motivation

In this study, there are just five classifier algorithms used to detect the best accuracy. But a large number of classified algorithms are available in machine learning. In a further study, the researcher can use more classifier algorithms. On the other hand, many machine learning tools are available, but just two tools are used and compared in this study. At last, in further study, the researcher can use more classifier algorithms and tools to correspond with them.

Our goal is to use this data to make a comparison where using ML we can automatically compare from accurate data. Because we know for the people also lot of people don't know they suffer from diseases. They might help them without the user being awkward. From all the data sets we choose Kaggle to collect our data from. Because it has a wider range of users and people share their feeling via small passages mostly to the public. It is also easy to collect huge amounts of data related to our work from there. Even though we know that human disease is a major problem for most people. We kept our data collection according to our ease of data sorting. So that it can work with basic classifier algorithms.

1.3 Rationale of the Study

It is a really common practice among recent researchers to use ML in different fields in order performance analysis using weka and rapid miner. Also, it is being used in the 5 algorithms and 3 diseases for our research. Our work follows this path where we are using the tools and analyzing the results from them using Weka and rapid miner tools. There have been many works done in this field previously. The same type of themes was practiced but in different ways. Different data gave different outcomes to different works. Our work focuses on the use of recent diseases from the media human bodies and findings in our data set using Kaggle.

The major goal of the performance is to assess new diseases' current performance, enhance it, and lower the disease's future potential and value.

Powerful data mining tool (weka and Rapid miner) that enables model deployment, model operations, and data mining. Our comprehensive data science platform provides all data preparation and performance-based results analysis.

1.4 Research Questions

We even start with our research work some questions are essential to our work. These questions will create an outline for our work. The following questions are the ones we encountered.

- Why do we need to use the tools for data?
- What information do I need?
- What kind of methods to use for data collection?
- Why we will use these tools?
- Which kind of algorithm we will use?
- What will be our criteria for data collection?
- How can I use new tools and algorithms in our projects?

We will discuss a few quantitative results Of studies that have been done before moving on to the results. In order towards the performance of our accuracy and as a result, we started to notice more diseases can be detected. We tried to make sure that performance analysis finds the best accuracy using different tools

This research will offer insight into how to identify disease symptoms as effectively. Different diseases are known to have harmful effects on human bodies We have also studied different papers in search of great techniques which can accurately best the accuracy of the results based.

We have two main tools of data for our project. One source is an online available dataset that contains data from Kaggle. It has 3 types of datasets diabetes, breast cancer and the last one is kidney disease. We have collected data from Kaggle using tools. We manually labeled the data that we collected from the Kaggle. Depending on different algorithms we also labeled our data depending on our dataset.

1.5 Expected Output

Here are the outcomes of our project:

Algorithm	Weka			Rapid Miner		
	Breast Cancer	Diabetics	Kidney Disease	Breast Cancer	Diabetics	Kidney Disease
Decision tree	93.14%	73.82%	99.50%	91.15%	71.43%	99.17%
K-nearest neighbors	96.13%	70.18%	99.25%	75.55%	69.48%	60.89%
Naïve Bayes	92.61%	76.30%	97.75%	91.15%	73.48%	95.83%
Random Forest	96.48%	75.52%	1.00%	94.69%	73.38%	99.17%
Support Vector Machine	97.89%	77.34%	98.75%	94.74%	76.52%	95.00%

Table1.5: The percentage of weka & rapid miner tools

Our original output from the quick miner is shown on the right, and our output from the weka classification algorithm is shown on the left. So, as we can see from this table, all of the objects in the findings were located with the greatest precision. Weka tools will be used in this article to determine the best accuracy.

1.6 Project Management and Finance

For the machine to be able to foresee and forecast the future close price of any cryptocurrency, it is essential to train it to learn from the available dataset. These datasets will be utilized to create models using a variety of techniques, which will finish the prediction/forecasting assignment. Rapid miner software frequently uses graphs, plots, charts, and tables for data analysis, making it simple to see the results and contrast different attributes and models. However, a machine must be taught to learn from the available dataset, based on which models will be constructed using various methodologies, in order to accurately predict the future stock price. The output will be

defined by the frequency and probability distribution function of the training dataset after the models have been trained using a number of techniques and a training dataset.

1.7 Report Layout

In Chapter 1, we provide an overview of our research study and discuss the fundamentals of our argument. We talk about how these tools function, how it relates to the users, and how it significantly affects the way we gather data for our thesis. This chapter also discusses our motivation. We have also discussed the steps we took to start this task and our efforts. We have also briefly discussed probable results.

Our work's background is examined in Chapter 2. In this chapter, we discussed the idea behind our work and the inspiration for it. Additionally, we have attempted to give a fundamental summary of the global social network analysis that we collected from the internet. We have also talked about the obstacles we still have to overcome.

The major goal of the experimental analysis section of Chapter 3 is to identify and describe our research plan. The key topic we have covered is this. Because the type of data we use and how we use it have a major impact on the quality of our work. since the sort of data we select will affect both our outcome and result. We have talked about and examined the two datasets we used. We discussed and looked at the dataset.

We have also discussed the studies we conducted while doing our experiment in Chapter 4 Experimental Results and Discussion. We spoke about the strategies we employed. Both in our study and the methods we employed. We have discussed the theories, the results of the experiments, and their descriptions.

The future extent of this research project is briefly described as the study's scope in Chapter 5. This chapter wraps up the entire research article with a helpful conclusion that succinctly summarizes the study's key results.

Chapter 6 contains an overview of all of our study, a conclusion on what we've done so far, recommendations, potential future applications, and what we can learn from it.

CHAPTER 2

BACKGROUND STUDY

2.1 Preliminaries

The main objective for undertaking this comparative study was to gain more knowledge about data mining and data mining tools in general. After taking the Data Mining and Data Warehousing Course, I felt that it would be beneficial to have some sort of resource where future students would be able to learn about data mining tools without having to spend a large amount of time deciding which tool to use and where to find resources to learn them. This is mainly because students are required to work on a data mining project but have sufficiently less amount of time to learn various data mining tools. Another objective of this comparative study is also to provide relevant and quality resources.

One of the objectives of this comparative study is to provide a web interface for learning data mining tools and making them available to students for their help. In this the project, we are going to describe the implementation of some of the data mining tools like Weka, and Rapid Miner. Also, we will see sample implementations of classification and clustering algorithms in Rapid Miner and Weka.

2.2 Related Works

[1] To find the best accuracy in healthcare, their approach combines Weka & Rapid Miner with K-NN, Naive Bayes, and Decision Tree algorithms. For analysis, the accuracy % and error rate are employed as metrics of the health data. [2] They compared six popular data mining programs, including Orange, Weka, Rapid Miner, Knime, Matlab, and Scikit-Learn, by classifying heart disease using six machine learning methods, including Logistic Regression, Support Vector Machine, K Nearest Neighbors, Artificial Neural Network, Naive Bayes, and Random Forest. To assess the accuracy of each tool's performance across the methodologies, three performance metrics were used. [3] They must classify the five classification methods (Naive Bayes on WEKA) using two illness datasets (breast cancer and diabetes) in this study. This is a useful categorization tool utilized in this paper for SMO, REP Tree, J48, and MLP. On the breast cancer and diabetes dataset, the performances of these five algorithms have been examined. [4] Knowing which data tool produces better results for the chosen algorithm and which tool is better for the chosen dataset size is crucial. The three datasets in this study are (Iris Dataset, Car Evaluation Dataset, and Tic-Tac-Toe end game Dataset). The aforementioned data mining tools apply these algorithms to the three chosen datasets (WEKA, Rapid Miner, Orange) The accuracy of each method (Decision tree, Naive

Bayes, and Random forest) in each data mining tool will be taken into account by using these algorithms. [5] In the current study, we present comprehensive data information. mining tactics with a larger focus on categorization as a key supervised learning method. In addition, WEKA software is discussed as a suggested tool for performing classification analysis on different kinds of available data. To make it simpler for a range of users to utilize the software, a complete technique is provided. WEKA's main features include 15 attribute/subset evaluators, 49 tools for preparing data, 76 algorithms for classification and regression, 8 methods for clustering, 3 algorithms for determining association rules, and 10 search algorithms for feature selection. By removing crucial information from the data, WEKA aids in the selection of an appropriate algorithm for building an accurate prediction model from the data. [6] The fundamentals of data pre-processing, classification, clustering and the WEKA tool are presented in this study. Oura is a tool for data mining. In this work, we lay out the procedures for using the WEKA tool with these technologies. It offers the option to categorize the data using different methods. [7] To choose the best model and tool, this paper analyzes agricultural soil health using machine learning techniques. Additionally, bibliometric analysis is used to find relevant sources and authors' keywords to determine the proposed work's area of focus. Different ML algorithms are used to build models on the SK-Learn, KNIME, WEKA, and Rapid Miner tools. On these tools, soil data is analyzed using Naive Bayes, Random Forest (RF), Decision Tree (DT), Ensemble learning (EL), and k-Nearest Neighbor (KNN). The Decision Tree model performs better than other algorithms, according to the results, followed by better accuracy provided by the RF algorithm, which is a collection of different decision tree algorithms, the SK-Learn tool, the WEKA tool, and then the KNIME tool. [8] In this work, models were built using the same classification techniques in several data mining tools using the bank marketing data set from the UCI Machine Learning Data Set. The performances of the classification models were evaluated using accuracy, precision, and f-measure criteria. The data set was split into training and test data sets with separation ratios of 60–40%, 75–25%, and 80–20%. R, Knime, RapidMiner, and WEKA are data mining applications that are utilized for these procedures. Additionally, the 5 decision trees, Naive Bayes, and k-nearest neighbor (k-NN) classification algorithms are frequently utilized in these platforms. [9] To extract information that can be used by a medical specialist to make an early disease diagnosis, healthcare data sets must be mined. In this paper, the features of popular simulators including Weka, RapidMiner, Spyder, Orange, R tool, and KNIME are presented. Heart disease risk is predicted utilizing a performance analysis of data mining classification approaches applied to two simulation tools over health care data sets. [10] A thorough analysis of educational data mining (EDM) and learning analytics (LA) in education was undertaken as a result of the realization that data mining analytics could have an impact on students' learning processes and outcomes. To analyze student performance using EDM, we did a thorough evaluation of the relevant literature for this essay. [14] To

identify the best classifier for a given application, the authors here attempted to use the WEKA tool to compare the effectiveness of multiple classifiers on a dataset. Several tools offer classifiers, and there are many performance analysis metrics to evaluate a classifier's effectiveness. WEKA has been used to implement Bayes Net, Naive Bayes, and binomial in the current investigation. [16] The classification of data mining techniques was covered in this work. In this study, we employ the WEKA interface and the two classification algorithms NAIVE BAYES AND MLP. Datasets related to hypothyroidism and breast cancer were used to examine how well these two algorithms performed. We selected these two datasets from the UCI Machine Learning Repository. [18] The goal of this project is to develop more effective analytical methods for the early diagnosis of cancer. Additionally, precision is crucial in making predictions that will raise the standard of care and consequently, the survival rate. The datasets for this work were taken from the University of Wisconsin Hospitals' UCI Machine Learning Repository. The K Nearest Neighbor (KNN) classifier is used for the diagnostic and classification process with various values of the K variable, establishing the procedure known as KNN Clustering. [20] For our research work, the BUPA liver dataset is acquired from the UCI machine learning library. Accuracy, positive predictive value, negative predictive value, sensitivity, specificity, and F1 score are used to evaluate the performance of the suggested scheme. RFs and ANNs, used in conjunction with the F1 score of 75.86% and 82.76% in the testing phase, gave the scheme accuracy results of 80% and 85.29%, respectively.

2.3 Comparative Analysis and Summary

To complete this project, we had to master the use of Weka and Rapid Miner. These tools for data mining and data visualization each have unique features. They are distinctive and well-liked in their own right due to their distinctions.

The following list of comparable characteristics offered by each of these data mining tools is what I came up with after researching the features of each tool. Users can judge each tool's usability. This explains which user interface is, in

2.4 Scope of the problem

After defining comparative research, this article outlines some of the centers of its central problems, including These are:

- The problem resulting from the complexity of these data.
- The problem of using methods in the study of human diseases increases or decreases.
- The problem of verification and prediction-making in complexity.

- Case selection, unit, level, and scale of analysis.
- Construct equivalence.
- Variable or case orientation.
- Causality.

2.5 Challenges

Data analysis to comparative analysis is great information, which is the most widely used to decrease different kinds of diseases.

During our work on this project, we faced challenges like

- Proper data collection
- Selection of algorithms
- Looking for data collection sources
- Data validation

In this research, we attempt to provide a framework that can compare the results using these tools. Since it is challenging to collect enough annotated training data and test data using machine learning techniques has not been extensive.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrumentation

The best results employing machine learning are the focus of our investigation. To create our model for this job, we require an actual dataset. Weka and Rapid Miner were employed to carry out our project. A few deep-learning libraries are also used. Five classification algorithms were employed as the framework in this study.

3.2 Dataset Utilized

Data collection is one of the most important tasks in every research. The datasets used for this experiment are taken from the online data repository Kaggle. In total, three datasets have been used namely breast cancer, diabetics, and kidney disease dataset. Table 2 illustrates the details of the datasets selected for this research work:

Dataset	Instance	Attributes	Types Of Classification	Types of Class Variables
Breast Cancer	500	32	Binary	Benign & Malignant
Diabetics	768	9	Binary	Tested-positive & Tested-negative
Kidney Disease	700	20	Binary	Yes or No

Table 3.2. Dataset Description

3.3 Statistical Analysis

The statistical analysis we ran statistical tests to verify the consistency between the parameters we selected and the length of the activities, assessing the effectiveness of the project management approach. We used the tests of independence and correlation to accomplish this, which were respectively discussed Test for Independence The test of independence between two qualitative variables is performed using information from a contingency table.

3.4 Proposed Methodology

To achieve our goals, we compared several classification algorithms in data mining using the programs weka and rapid miner [18]. This section summarizes the approach we used to achieve those goals. In the beginning, it requires specific training datasets about some of the serious diseases that endanger human life. Next, choose either weka or Rapid miner, two appropriate open-source programs for this task. Thirdly, assess the models produced by chosen approaches and look at how well each methodology works with each tool.

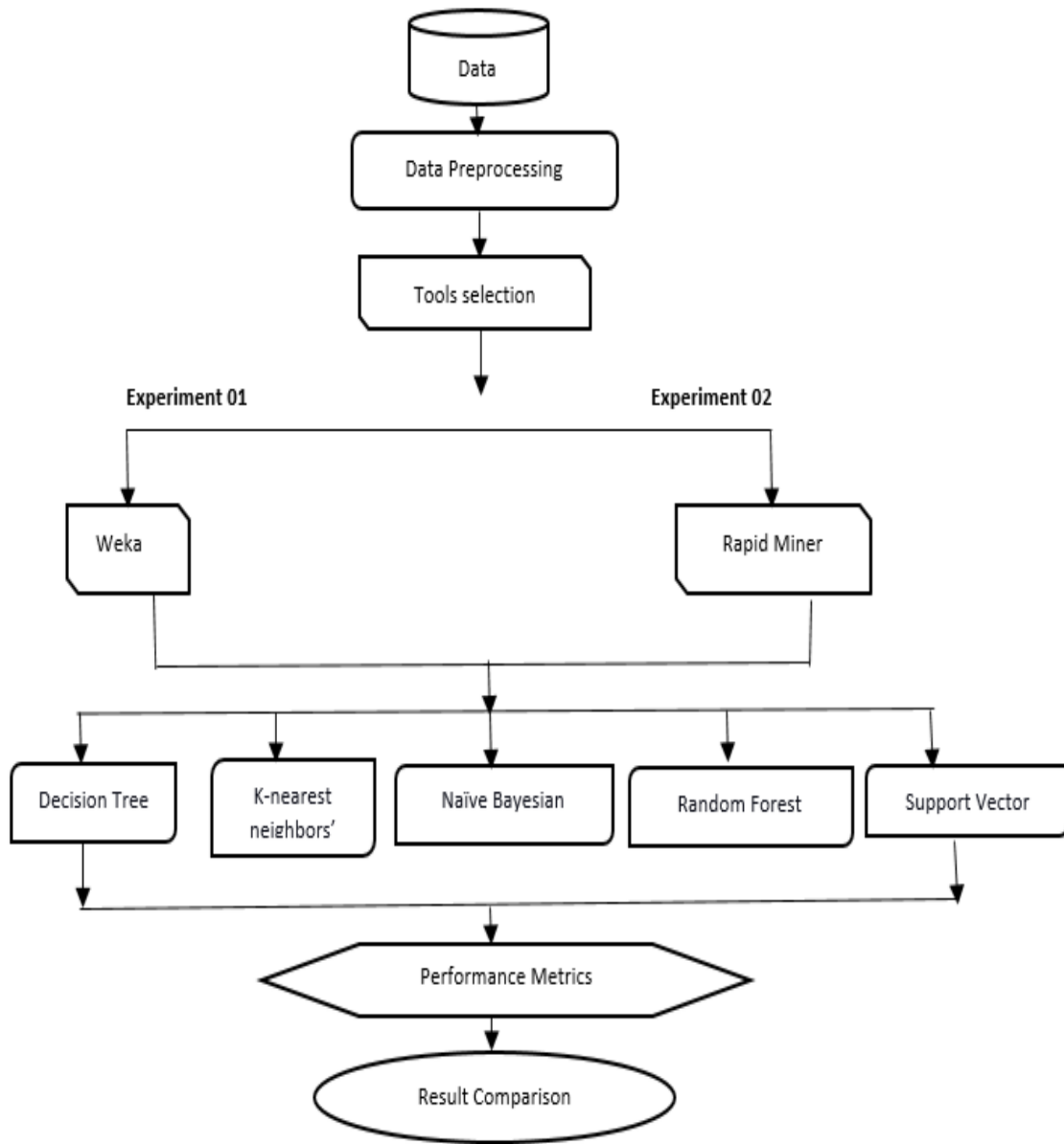


Figure 3.3: Methodology of Research

3.5 Implementation Requirements

The main technologies used to design the web interface for the comparative study includes a description of the Rapid Miner, Weka implementation of classification algorithms in Rapid miner, and Weka with their description and visual steps. The comparative study also involves a sample quiz based on the knowledge of these data mining tools.

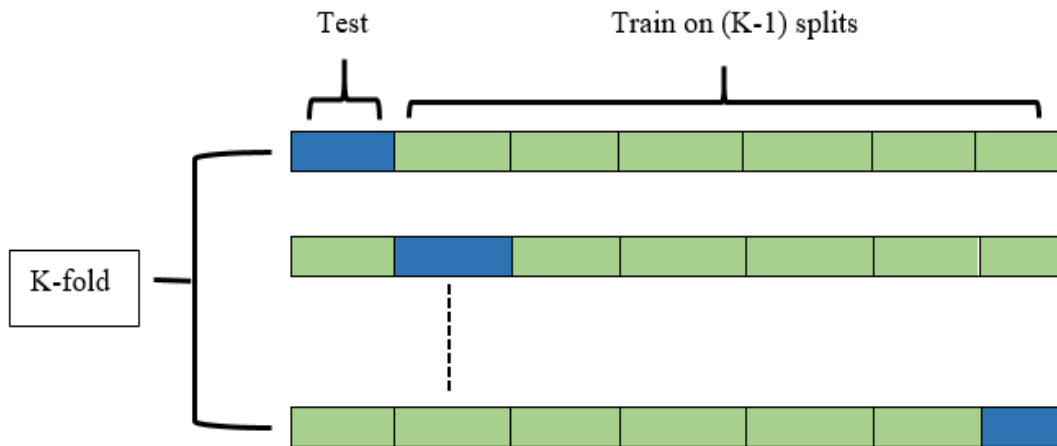
As the main focus of this comparison study is on data mining tools, We have initially provided information about Rapid Miner & Weka.

CHAPTER 4

EXPERIMENTAL RESULT ANALYSIS

4.1 Experimental Setup

A training set and a testing set comprise the dataset for this investigation. The training set contains 80% of the data, while the testing set (validation dataset) only contains 20%. The training dataset is split into k equal portions using the K -fold cross-validation procedure, which was used in this study. One split is set aside for the validation dataset for each iteration, while the remaining $k-1$ splits are kept as training data. A diagram of



K -fold cross-verification can be found below:

Figure 4.A1. K -fold cross-validation

That sketch is followed by an overview of the various metrics used to evaluate the effectiveness of the chosen classifiers. In general, the criteria for comparing algorithms can be done by assessing their speed, accuracy, scalability, interpretability, and robustness [23]. Three aspects are taken into account in this study: scalability, which is the capacity of a classifier to create a model effectively when the classifier is applied to a big set of data, and speed, which refers to the amount of time required to generate and develop the model. The capacity of a classifier to accurately anticipate the class label is used to define algorithm accuracy. The Confusion matrix [Provost and Kohai, 1998] is another technique that shows the expected and actual categorization. Accuracy, recall,

and F-Measure are just a few of the criteria built using a confusion matrix. Formulas utilizing a 2 x 2 confusion matrix are shown in Figure 4.A2.

True positive (TP)	False Negative (NG)
False Positive (FP)	True positive

Figure 4.A2: Confusion Matrix

Evaluation Matrix:

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

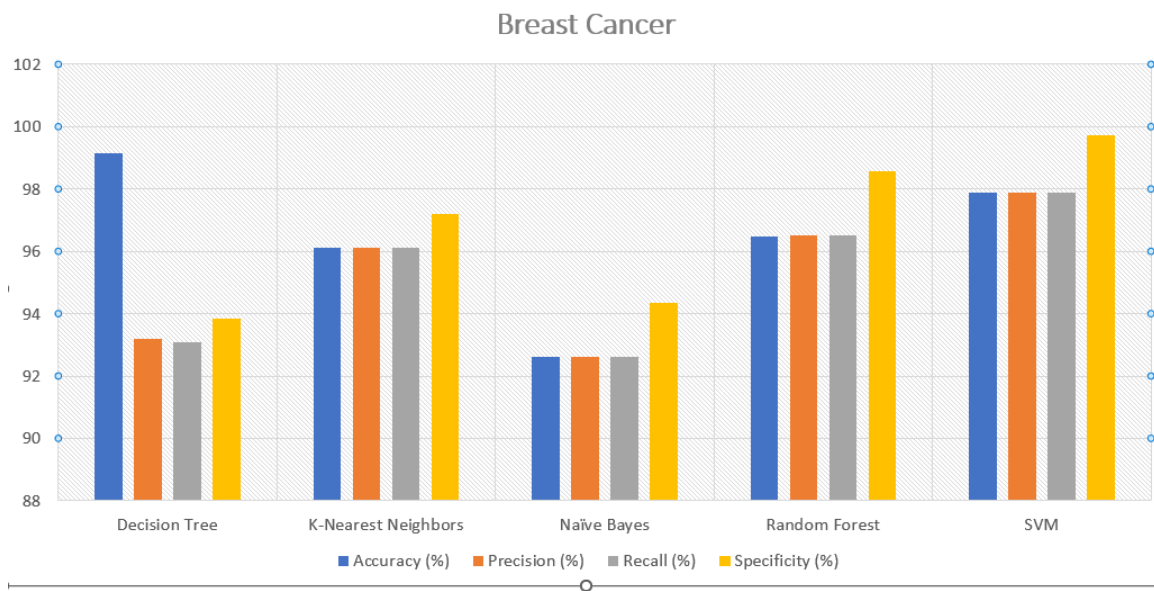
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

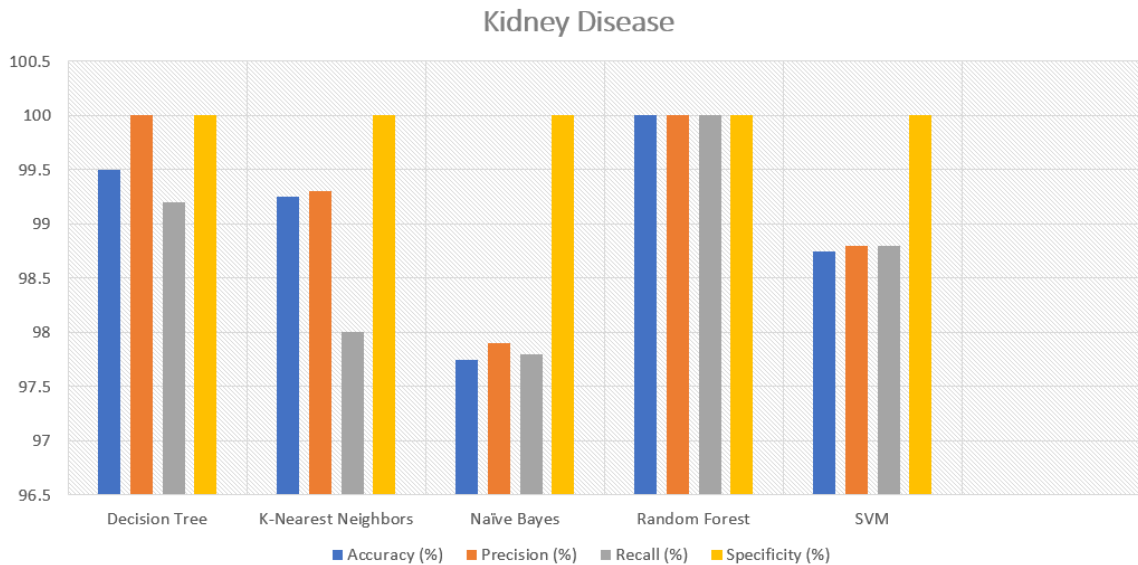
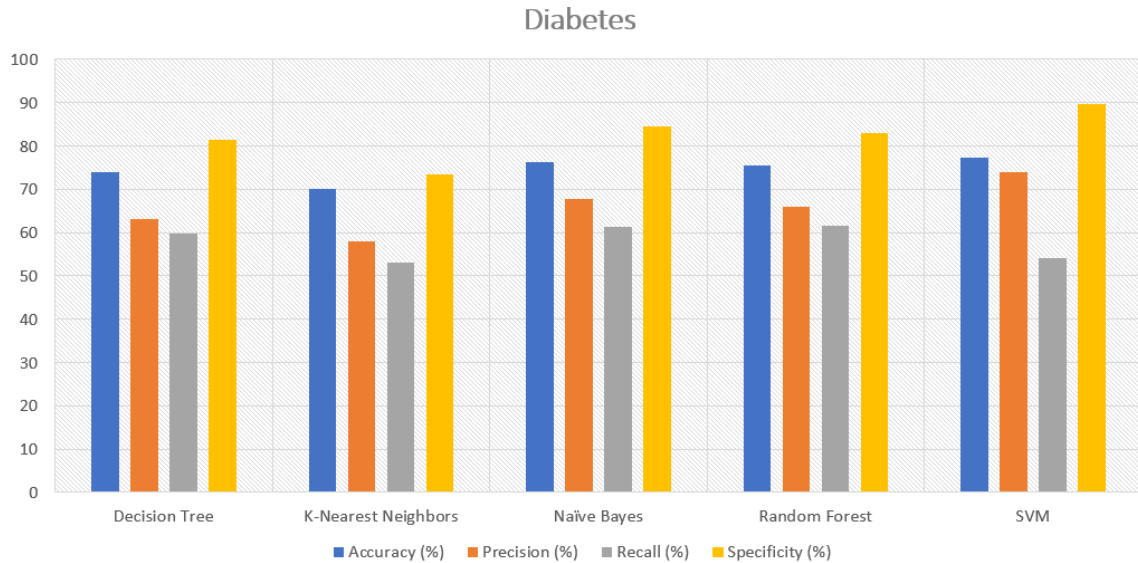
4.2 Experimental Results & Analysis

Finally, We have implemented five classification algorithms in Rapid Miner and Weka to have a fair idea of the difference in steps needed to perform these two tools. We have implemented the K-NN algorithm, Naïve Bayes, SVM, Random Forest, and Decision tree classifications algorithm as an example by explaining each step systematically in such a way that students understand the working of both tools. Experiment 1&2 describes the implementation of classification algorithms and the best accuracy of these results.

Experiment 1:

Result:





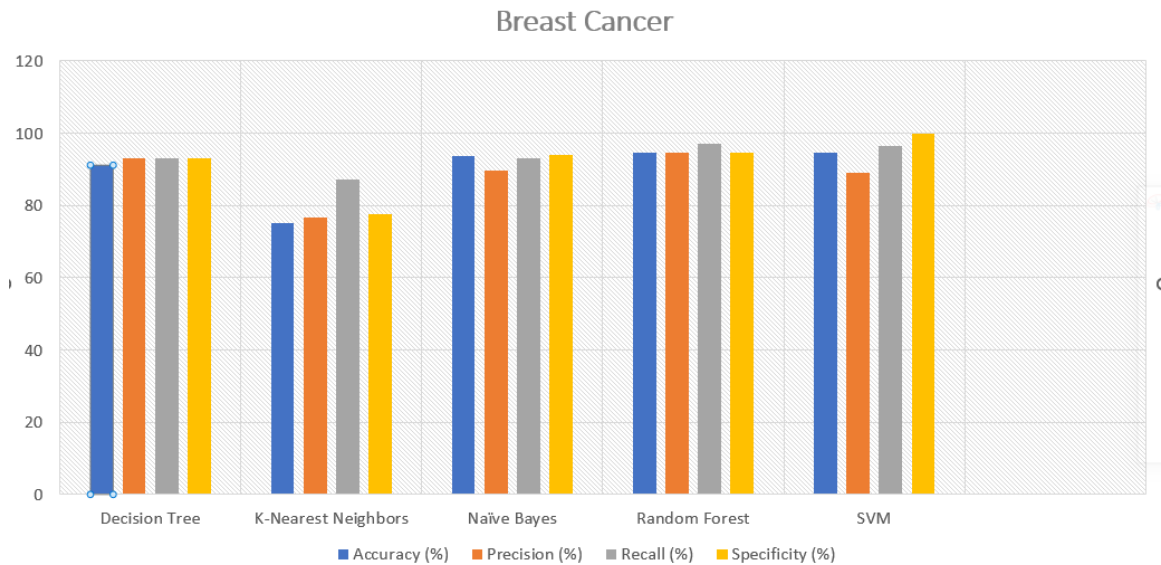
The accuracy acquired in training the breast-cancer dataset by the five classifiers. It shows that Sequential Minimal Optimization has the higher accuracy with 97.89%, followed by Random Forest based classification with 96.48%, then Instance-based K-nearest Neighbors and Decision trees with approximately the same accuracy respectively 96.13% and 93.14%. Naïve Bayesian Classification with 92.61%.

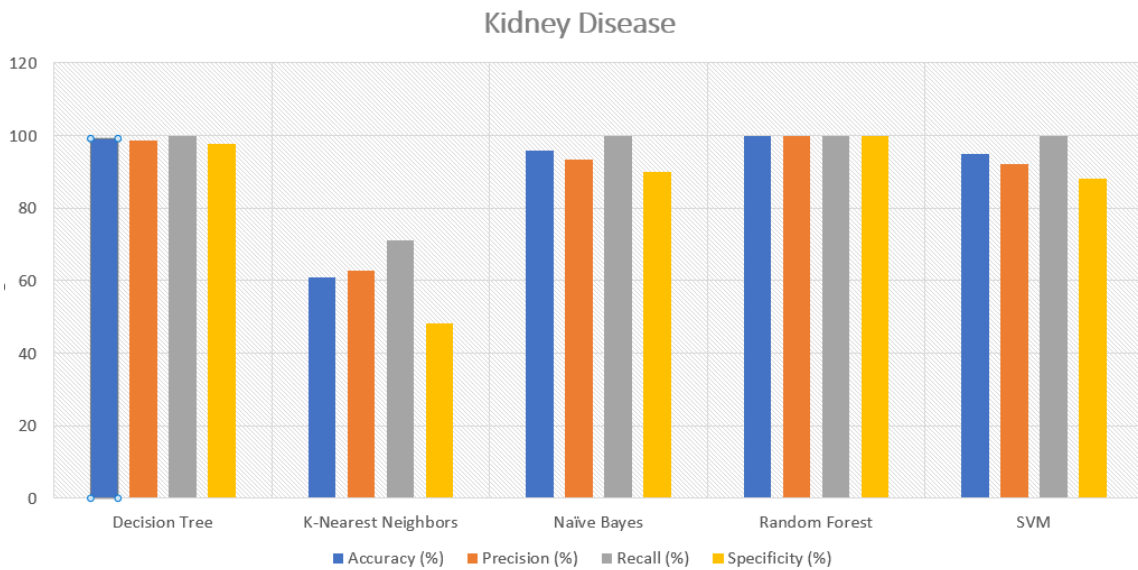
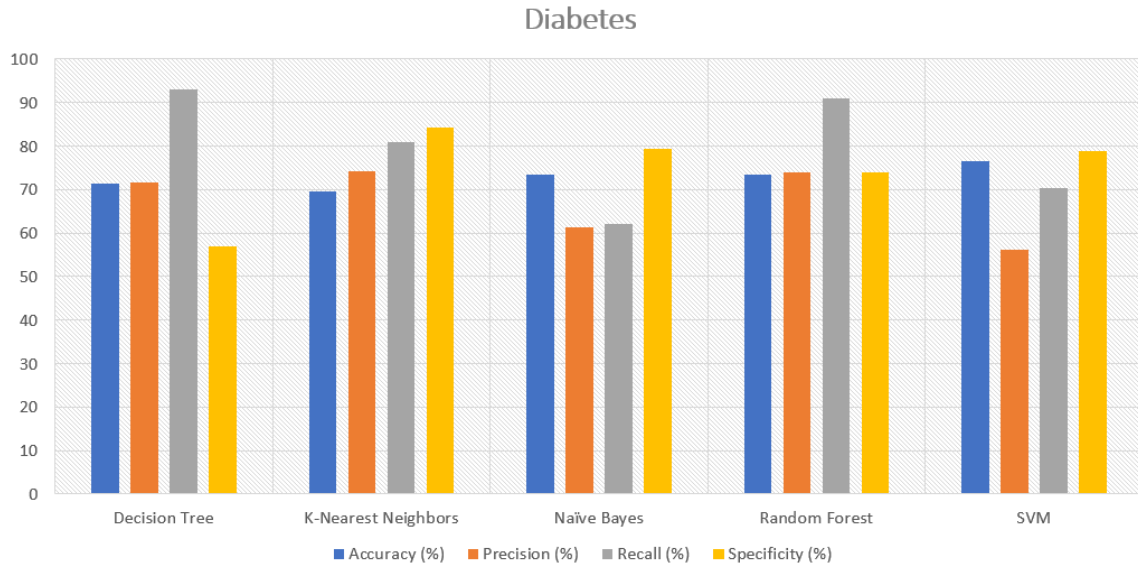
On the other hand, the training data is related to diabetes, and the results are different; Sequential Minimal Optimization has the best accuracy with 77.34%, followed by Naïve Bayesian Classification with 76.30%, then Random Forest based classification with

75.52%, next Decision trees with 73.82%, and Instance-based K-nearest Neighbors with 70.18%. And Random Forest based classification has the greatest accuracy at 1.00%, followed by Decision trees at 99.50%, then Instance-based K-nearest Neighbors at 99.25%, then Sequential Minimal Optimization at 98.75%, and Naïve Bayesian Classification at 97.75%.

Experiment 2:

Result:





The accuracy acquired in training the breast-cancer dataset by the five classifiers. It shows that Sequential Minimal Optimization has a higher accuracy with 94.74%, followed by Random Forest based classification with 94.69%, then Decision trees and Instance-based K-nearest Neighbors with approximately the same accuracy respectively 91.15% and 75.55%.

On the other hand, the training data is related to diabetes, and the results are different; Sequential Minimal Optimization has the best accuracy with 76.52%, followed by Naïve Bayesian Classification with 73.48%, then Random Forest with 73.38%, next Decision

trees with 71.43%, Instance-based K-nearest Neighbors 69.48% And Random Forest has the greatest accuracy with 1.00%, followed by Decision trees with 99.17%, then Naïve Bayesian Classification with 95.83%, then Sequential Minimal Optimization with 95.00%, and Instance-based K-nearest Neighbors with 60.89%.

Result Analysis:

A comparison of machine learning technologies employing WEKA and Rapid Miner with classifier methods is done for three diseases. We carried out an experimental investigation on the weka and quick miner tools using five classifier algorithms. a summary of the performance information for two tools based on machine learning. Out of the two instruments, Weka provides the best result.

In the result analysis above, the majority of the 99.25% Weka accuracy can be shown, showing that Weka performs better with the classifier. Weka also boasts the highest level of precision, allowing it to precisely identify successful results. Additionally, the 99% precision of rapid mining shows that Weka rated it as positive.

Algorithms	Rapid Miner	Weka
Decision Tree	99.17%	99.15%
K-NN	75.55%	99.25%
Naïve Bayes	95.83%	97.75%
Random Forest	99.17%	96.48%
Support Vector Machine	95.00%	98.75%

Table 4.2: Best accuracy of the result

4.3 Discussion

Three metrics are typically used to assess accuracy. The optimal Dataset level makes up the first measurement., the second measure is data pre-processing per dataset and when the data process is successfully done then We can visualize the structure in this graph undertaken. This project has taught me how to do comprehensive research on several topics and use that research to further expand our knowledge. We are more comfortable working on Weka. This comparative study will further encourage the students to learn several other data mining tools and use them for their data mining working on this project has taught us new concepts and enhanced our working development.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

Different diseases, distress, and behavioral disturbances affect us in different ways. Anxiety-related issues are particularly prevalent. These circumstances can have an impact on your disease symptoms in addition to how you feel emotionally.

Thankfully, there are effective therapies for this, such as medication and psychological counseling. It will be simpler to adhere to those disease treatment regimens if these issues are resolved. Any worries you have about your psychological health can be evaluated by a healthcare professional, such as your doctor or a cardiac rehabilitation nurse.

5.2 Impact on the Environment

Conditions including kidney disease, some types of breast cancer, and respiratory disorders can all be exacerbated by environmental pollutants. People with lower incomes are more likely to live in polluted areas and have access to contaminated water. Additionally, children and pregnant women are more at risk for pollution-related health problems.

The physical environment is negatively impacted by a variety of human activities including deforestation, pollution, overcrowding, and the combustion of fossil fuels. Changes like this have contributed to climate change, soil erosion, poor air quality, and undrinkable water.

5.3 Ethical Aspects

- The objectives of the research are promoted, and respondent and researcher trust is increased.
- To protect the welfare, rights, and dignity of study participants, it is crucial to adhere to ethical standards.
- Researchers can be held responsible and brought to account for their decisions.
- Social and moral values are promoted by ethics.

- Promotes the objectives of the study, such as understanding, accuracy, and avoiding errors.
- Ethical norms uphold the virtues necessary for collaborative work, such as faith, accountability, respect for one another, and impartiality.
- Public support for the research is also influenced by ethical norms in the field. When people have faith in a study project's reliability and validity, they are more likely to trust it.

5.4 Sustainability plan

The topic of sustainable development has been the subject of numerous studies, articles, analyses, and comparisons during the past few decades. The definition of sustainable development is not something we can readily define, and neither is the creation of a strategy. It is also essential to thoroughly research the various diseases' sustainability metrics and indices. The country's sustainable development strategies, which include and identify the indicators, as well as the system are a tremendous aid in these.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

This research has given us many insights into this topic. Some diseases are a sensitive topic at this time. This is the reason it is not widespread well aware news or information. That is the reason we used ML to recognize the hidden information in the data. These patterns are what ML uses to do predictions.

Already we have mentioned that while doing our study we used data from Kaggle and the reason behind this is we needed treatment where people are comfortable or frequently share their diseases Kaggle met expectations. This data helped our algorithms to train and learn sentence patterns and which was used for comparative analysis. There are a few problems that were addressed before in the beginning. We were able to achieve the goal which we were working for. Different algorithms gave us different kinds of outcomes. We discussed it in detail.

6.2 Conclusion

The goal of this research was to establish strategies to choose an appropriate classifier and collection of tools for use with health problem datasets involving breast cancer, diabetes, and kidney illness. Weka and Rapid-Miner are two sophisticated machine learning programs that employ five different classification methods in Data Mining: decision trees, instance-based k-nearest neighbors, Naive bayesian classification, random forest-based classification, and sequential minimal optimization. In this research, the findings of several classifiers show that Rapid miner provides the highest accuracy while using the same dataset as Weka. The best outcomes are achieved using the Rapid miner and the Weka software using the random forest algorithm.

6.3 Implication for Further Study

In this study, there are just five classifier algorithms used to detect the best accuracy. But a large number of classified algorithms are available in machine learning. In a further study, the researcher can use more classifier algorithms. On the other hand, many machine learning tools are available, but just two tools are used and compared in this study. At last, in further study, the researcher can use more classifier algorithms and tools to correspond with them.

REFERENCES

- [1] Y. Farhaoui, B. Aksasse, and S.S. Alaoui, 2018. Data classification techniques using data mining. *Int. J. Tomogr. Simul*, 31, pp. 34–44.
- [2] Abelha, Abelha, Pinto, Ferreira, Neto, and Machado, 2020. to forecast chronic kidney disease in its early stages via data mining. *Procedia Computer Science*, vol. 177, p. 562–567.
- [3] Zeb, K., Al-Rakhani, M., Derhab, A., and S.A.C. Bukhari, 2021. A thorough evaluation of diabetes detection and prediction using data mining. *IEEE Access*, Volume 9, Pages 4371–43735.
- [4] German, L.B., Margarita, R.V., Elisa Clementina, O.M., Jose, C.O., Marlon Alberto, P.M., Eugenia, A.R., Roberto Cesar, M.O., and Fabio Enrique, M.P., 2020, July. an approach for analyzing breast cancer recurrences based on data mining techniques. *International Swarm Intelligence Conference* (pp. 584-596). Cham Springer.
- [5] Dickinson, J.A. and Tonelli, M. 2020. Implications for low-, middle-, and high-income nations of early CKD identification. *American Society of Nephrology Journal*, 31(9), 1931–1940.
- [6] A. Sandbank, T. Lauritzen, K. Borch-Johnsen, W. Herman, R. K. Simmons, S. J. Griffin, M. J. Davies, K. Khunti, G. E. Rutten, and M. B. Brown, 2015. Intensive Treatment in People With Screen-Detected Diabetes in Primary Care: An Anglo-Danish-Dutch Study Our findings demonstrate that early diagnosis and treatment of type 2 diabetes minimize cardiovascular morbidity and death (ADDITION-Europe).
- [7] German, L.B., Paola Patricia, A.C., Eugenia, A.R., Elisa Clementina, O.M., Roberto Cesar, M.O.2020, July, Jose, C.O., Margarita, R.V., Marlon Alberto, P.M., Fabio Enrique, M.P. a methodology based on data mining methods for analyzing breast cancer recurrence. *International Swarm Intelligence Conference* (pp. 584-596). Cham Springer.
- [8] N. H. Ismail, F. Ahmad, and A. A. Aziz, 2013. using WEKA as a data mining tool to implement a naive Bayes classifier to analyze student academic performance. *Postgraduate Research Conference at UniSA*.
- [9] Rodriguez, J.C., Han, J., and M. Beheshti, 2008, December Rapidminer is used for diabetes data analysis and model discovery. 2008 was the second international conference on next-generation networking and communication (Vol. 3, pp. 96-99). IEEE.
- [10] P. Bajaj and P. Gupta, 2014. Review of data mining approaches used in the diagnosis of heart disease. *IJSR*, 3(5), 1593–1596. *International Journal of Science and Research*.
- [11] Robu, R., and S. Holban. using the genetic algorithm to classify data *Journal of Artificial International*. Wang, K., and Luo, J. (2016) .
- [12] Recognizing illness symptoms on faces that can be seen. 2016(1), pp. 1–8, in *EURASIP Journal on Bioinformatics and Systems Biology*.
- [13] F. Haghanikhameneh, P.H.S. Panahy, N. Khanahmadliravi, and S.A. Mousavi, 2012. a comparison of categorization methods using the Squid dataset and data mining algorithms. 9, pp. 59–66, *IJAI*.

- [14] Stuetzle, W., Murua, A., and Sieberts, S. (2001). Model-based categorization and clustering of documents draft being prepared.
- [15] 2013 study by Beshah, T., Ejigu, D., Abraham, A., Snasel, and P. Road User Behavior and the Role in Road Accident Data Mining: Implications for Increasing Road Safety. *International journal of simulation and tomography*.
- [16] Mu, X. Shen, and J. Kirby's 2017 paper. For text-based chatbot conversation, a support vector machine classifier based on an approximative entropy measure is used. *Int. J. Artif. Intell.*, 15(2), pp. 1–16.
- [17] Kalim, A., Alqahtani, H., Faruque, M.F., and Sarker, I.H., 2020. Diabetes mellitus prediction and analysis for eHealth services based on K-nearest neighbor learning. *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 26 (p. e4–e4).
- [18] (2017) Wu, H., Yang, S., Huang, Z., He, J., and Wang, X. data mining-based type 2 diabetes mellitus prediction model. 10, pp. 100–107 of *Informatics in Medicine Unlocked*.
- [19] September 2012, Liu, Y., Wang, Y., and Zhang, J. Random forest is a brand-new machine learning algorithm.
- [20] S. Sossi Alaoui, Y. Farhaoui, and B. Aksasse, April 2017. a comparison of the four well-known data mining classification techniques. *Advanced Information Technology, Services, and Systems Conference* (pp. 362-373). Cham Springer.
- [21] STEFANOWSKI, J., 2008 Classifier evaluation in data mining.
- [22] F. Provost and R. Kohavi, 1998. Introduction by the guest editors: Applied machine learning research.

APPENDIX

It is challenging to identify the issues and circumstances that arose as we were putting the finishing touches on our project, even when they are listed in the appendix. To achieve perfection and the best possible outcomes, we first choose the best algorithms from among all the others. Additionally, everyone must fully know how to use machine learning. The process of gathering and compiling such a huge dataset won't be as simple as we anticipated. Finally, we were successful in completing that.

Report_Final_docx.docx

ORIGINALITY REPORT

23%	18%	9%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	8%
2	Submitted to Daffodil International University Student Paper	5%
3	Submitted to University of Warwick Student Paper	1%
4	journals.riverpublishers.com Internet Source	1%
5	Submitted to Universiti Tenaga Nasional Student Paper	1%
6	Submitted to Holy Family University Student Paper	1%
7	Submitted to Florida International University Student Paper	1%
8	www.researchgate.net Internet Source	1%
9	iranarze.ir Internet Source	<1%