A CAPSTONE REPORT

Analysis of different machine learning models for Liver disease prediction

by

Md. Al-Imran Hossain
17182103318
Saimun Islam
17182103352
Md. Saiful Islam
17182103353
Nayeem Hossain Khan
17182103354
Sanjida Islam Payel
17182103361

Project Report Submitted in Partial Fulfillment of

The Requirements for The Degree

of

Bachelor of Science in Computer Science and Engineering

Under the supervision of

T.M. Amir - Ul – Haque Bhuiyan

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BANGLADESH UNIVERSITY OF BUSINESS AND TECHNOLOGY

Summer 2022

Date: June 11, 2022

DECLARATION

We do hereby declare that the project entitled "**Analysis of different machine learning models for Liver disease prediction**" submitted for the degree of Bachelor of Science Engineering in Computer Science and Engineering in the faculty of Computer Science and Engineering of Bangladesh University of Business and Technology (BUBT), is our original work and that it contains no material which has been accepted for the award to the candidates of any other degree or diploma, except where due reference is made in the next of the project to the best of our knowledge, it contains no materials previously published or written by any other person except where due reference is made in this project.

Md. Al-Imran Hossain
ID: 17182103318
Intake: 38th
Section: 04

Saimun Islam
ID: 17182103352
Intake: 38th
Section: 04

Md. Saiful Islam
ID: 17182103353
Intake: 38th
Section: 04

Nayeem Hossain Khan
ID: 17182103354
Intake: 38th
Section: 04

Sanjida Islam Payel
ID: 17182103361
Intake: 38th
Section: 04

APPROVAL

This project "**Analysis of different machine learning models for Liver disease prediction**" report submitted by Md. Al-Imran Hossain, Saimun Islam, Md. Saiful Islam, Nayeem Hossain Khan and Sanjida Islam Payel students of Department of Computer Science and Engineering, Bangladesh University of Business and Technology(BUBT), under the supervision of **T.M. Amir-Ul- Haque Bhuiyan**, Assistant Professor, Department of Computer Science and Engineering has been accepted as satisfactory for the partial requirements for the degree of Bachelor of Science Engineering in Computer Science and Engineering.

T.M. Amir-Ul- Haque Bhuiyan
Assistant Professor
Department of CSE
Bangladesh University of Business and Technology

Md. Saifur Rahman
Assistant Professor & Chairman
Department of CSE
Bangladesh University of Business and Technology

ACKNOWLEDGEMENTS

We would like to express our deep and sincere gratitude to our research supervisor, T.M. Amir-Ul- Haque Bhuiyan, for giving us the opportunity to conduct research and providing invaluable guidance throughout this work. His dynamism, vision, sincerity and motivation have deeply inspired us. He has taught us the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under his guidance.

We are greatly indebted to our honorable teachers of the Department of Computer Science and Engineering at the Bangladesh University of Business and Technology who taught us during the course of our study. Without any doubt, their teaching and guidance have completely transformed us to the persons that we are today.

We are extremely thankful to our parents for their unconditional love, endless prayers and caring, and immense sacrifices for educating and preparing us for our future. We would like to say thanks to our friends for their kind support and care.

Finally, we would like to thank all the people who have supported us to complete the project work directly or indirectly.

Md. Al-Imran Hossain, Saimun Islam, Md. Saiful Islam,
Nayeem Hossain Khan and Sanjida Islam Payel.
Bangladesh University of Business and Technology.

DEDICATION

Dedicated to

– *Our Parents*

ABSTRACT

Recently, data mining has become a useful tool in the healthcare sector for predicting disease. The process of mining data involves retrieving information from vast databases, warehouses, and other sources. The task of predicting diseases from massive medical datasets poses a significant challenge to researchers. Solving those issues many researchers try various data mining techniques but most of them are not perfect on the basis of their accuracy. To solve these issues here we proposed a more accurate technique to predict liver disease. In this research, we used Liver Patient Dataset (LPD) and five supervised machine learning (Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, and K Nearest Neighbors) algorithms. In our findings, Random Forest Classifier gives the highest accuracy of 99.80%. The rest of the machine learning algorithms also give very good results.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Introduction

Scientists are facing more difficult tasks in the healthcare sector - predicting what diseases will emerge from the massive amounts of medical data. The healthcare industry nowadays is increasingly reliant on data mining. The classification, clustering, and association rule mining techniques are applied to medical data in order to find patterns that can be used to predict disease. The classification techniques used in data mining are popular for diagnosing and predicting diseases [1]. We apply Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, and K Nearest Neighbors classifier algorithms for predicting liver diseases in this study. There are a variety of liver illnesses that necessitate a practitioner's clinical attention [2]. The major goal of this study is to use the aforesaid classification algorithms to predict liver disorders such as Chronic Hepatitis, Cirrhosis, Liver Cancer, Bile Duct, and Acute Hepatitis from the Liver Patient Dataset (LPD) dataset. Besides being the second-largest internal organ in the body, the liver also plays an important role in biochemistry and in various other critical processes, such as red blood cell breakdown [3].

1.2 Problem Statement

The diagnosis of a disease is the most critical and vital task in medicine and this mostly depends on a doctor's intuition based on experiences in the past. Unfortunately, the difficulties in recognizing correct symptoms result in a misdiagnosis. To avoid such medical misdiagnosis, this study utilizes large datasets collected by healthcare industries to automate the diagnosis of diseases. The need to develop a tool that could aid doctors and prevent them from unwarranted errors and unwanted biases in diagnosis is established in this research. Therefore, an automated medical diagnosing system to tackle the problem of correct early detection of liver disease is considered as one of the outputs of this study.

1.3 Research Objectives

The objective of this project is to propose a rule-based classification model with machine learning and data mining techniques for the prediction Liver diseases.

1.3.1 Primary objective

- To gain practical experience by modeling an Artificial Intelligent and Data Mining technique based real world problem.
- To understand how Machine Learning model and Data Mining technique work on Liver disease classification.

1.3.2 Secondary objective

- To develop a Machine Learning model that deals with the real-world data.
- Predict Liver disease in early stage.

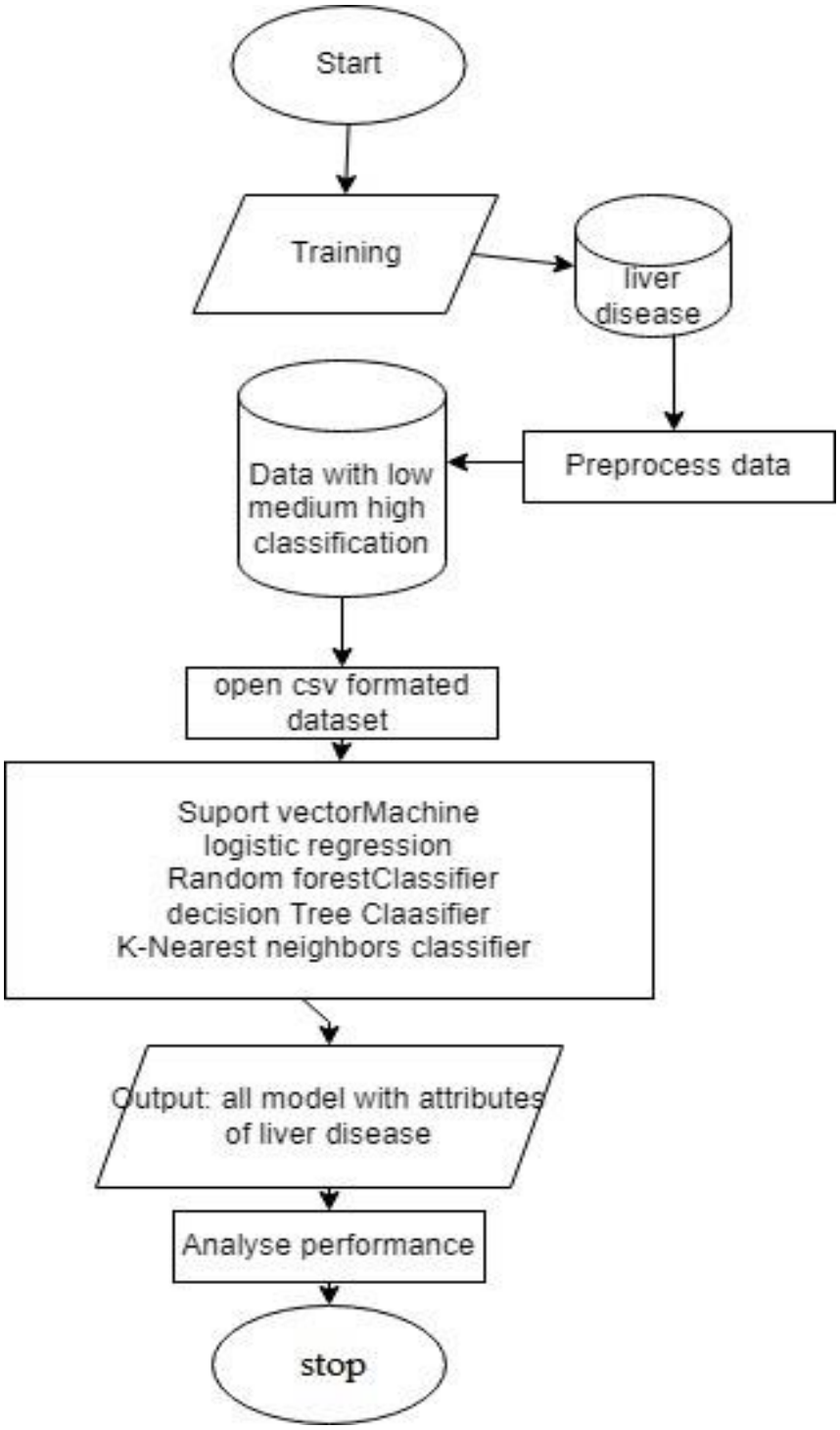
1.4 Motivation

The liver is a crucial and big organ in the human body, impacts the digestion system. Due to Liver diseases (LDs), so many deaths are occurred in worldwide that nearly 2 million deaths per year. The main LD complications are cirrhosis that 11th position in universal deaths, and others hepatocellular carcinoma and viral hepatitis that 16th leading position for global deaths. Fortunately, 3.5% of deaths are occurred due to LD. The capability of an ML approach for controlling LD can be identified through their factors, cofactors as well as complications respectively. So, in this datamining research project we want to build a project that can predict liver disease in early stage.

1.5 Flow of Research

The research work was performed in multiple steps. After finalizing the research topic, we first studied the basic theory of data mining and reports over liver disease that is needed to carry our research work. After the practice, we investigated for the best dataset for the project. We investigated the lacking of the proposed systems and conduct the higher

accuracy through ours project. After finalizing the design, we implemented the overall method. To test the proposed model, we collected a popular Dataset and ran tests and evaluations on our implemented project. Finally, we completed our project.



1.6 Research Contribution

The overall contribution of the research work includes:

- We investigate that most of the organizational information.
- Liver Disease is a system that can diagnose Disease or Non-Disease with the help of machine learning based on specific datasets.
- We can determine the accuracy of the diagnosis based on different models.

1.7 Summary

Based on the data set of Liver Disease using Python machine learning we can diagnose the disease using different models. We used a total of five models. We have been able to verify different accuracies using different models. We can determine the accuracy of the diagnosis based on different models. I got the highest accuracy Random Forest Classifier. I got the lowest accuracy Logistic Regression.

Chapter 2

RELATED WORKS

2.1 Introduction

The main goal in this project is to propose a rule-based classification model with machine learning and data mining techniques for the prediction Liver diseases. There has been so much prior research on the diagnosis of liver disease conducted by different researchers and the accuracy of different machine learning algorithms. Some of the best research work was reviewed in this section.

2.2 Literature Review

Dramodharan et.al [2] offered an early warning system for three common liver diseases such as Liver cancer, Cirrhosis, and Hepatitis based on distinct symptoms. In order to predict disease, the researchers used Naive Bayes and FT Tree algorithms. This article compares the classification accuracy measures of two algorithms based on their respective results. The researchers concluded from their experiments that Naive Bayes provided the greatest accuracy in predicting diseases.

Rosalina et al [4] predicted a poor hepatitis outcome. They used wrapper methods to remove noise features before the classification process. As a first step, SVM carried out feature extraction. The selection of features was implemented to minimize noise or irrelevant data. During the experimental study, the researchers observed a high accuracy rate in the cost of clinical lab tests with minimum execution time. Based on Least Squares Support Vector Machine (LS-SVM) and Modified Particle Swarm Optimization. Omar S. Soliman et al [5] proposed a hybrid classification system for HCV diagnosis. Principle Component Analysis is used to extract feature vectors. LS-SVM's sensitivity to parameter change is the reason the Modified-PSO Algorithm was applied to find optimal LS-SVM parameters in a smaller number of iterations. UCI's machine learning database repository

of HCV benchmark data was used to implement and evaluate the proposed system. PCA and LS-SVM were used in another classification system. Compared to other systems, the proposed system achieved the highest classification accuracy.

Using a soft computing technique, Karthik et.al [3] were able to diagnose liver disease intelligently. Three phases of classification and type detection have been implemented. A neural network (ANN) classification algorithm was used in the first phase of the study to classify liver disease. The second phase involved generating the classification rules by using the Learn by Example (LEM) algorithm. Fuzzy rules were used in the third phase to determine the type of liver disease. Using more input attributes, Chaitrali S. Dangare et. Al [6] have analyzed prediction systems for heart disease. On the heart disease database, Decision Trees, Naive Bayes, and Neural Networks are analyzed as data mining classification methods. Based on the accuracy of their results, these techniques are compared. According to the authors' analysis, Neural Networks has the best prediction of heart disease out of these three classification models.

2.3 Project significance

Liver disease prediction has its own significance to the medical domain. Due to Liver diseases (LDs), so many deaths are occurred in worldwide that nearly 2 million deaths per year. The main LD complications are cirrhosis that 11th position in universal deaths, and others hepatocellular carcinoma and viral hepatitis that 16th leading position for global deaths. Fortunately, 3.5% of deaths are occurred due to LD. The capability of an ML approach for controlling LD can be identified through their factors, cofactors as well as complications respectively. So, in this data mining research project we want to build a project that can predict liver disease in early stage.

2.4 Summary

This chapter reviewed the latest techniques **for Liver disease prediction system**, including the drawbacks. The thesis's target is to eliminate the imperfections as much as possible and introduce a new combined approach to the system.

Chapter 3

METHOD AND MATERIALS

3.1 Requirement analysis

To complete this project we need some software and hardware device those are given bellow,

- Google colab.
- NumPy.
- Pandas.
- Seaborn.
- Matplotlib.
- Sklearn.
- Xgboost.
- Dataset (Indian liver patient)

3.2 Dataset

Dataset is very important to train a machine learning model. Only good and clean dataset can give better result in machine learning project. IN this project we use Kaggle Liver Disease Patient Dataset [7]. This dataset consists of total 30000 people data about liver disease. Among 30000, 21917 patients have liver disease the rest of 8774 people don't have any liver disease. The total number of columns are 11. The first 10 column(Age of the patient, Gender of the patient, TB Total Bilirubin, DB Direct Bilirubin, Alkphos Alkaline Phosphatase, Sgot Alamine Aminotransferase, Sgot Aspartate Aminotransferase, TP Total Proteins, ALB Albumin and A/G Ratio Albumin and Globulin Ratio) of this dataset are main features of this dataset. The last column (Dataset) is the label or prediction column of the dataset. Number of male and female patient of this dataset is 15000 and

7000 respectively(fig-3.1).There are sum null value also present in this dataset those are Age of the patient(2),TB Total Bilirubin(648), DB Direct Bilirubin(561), Alkphos Alkaline Phosphatase(796), Sgpt Alamine Aminotransferase(538), Sgot Aspartate Aminotransferase(462), TP Total Proteins(463), ALB Albumin(494) and A/G Ratio Albumin and Globulin Ratio(559).

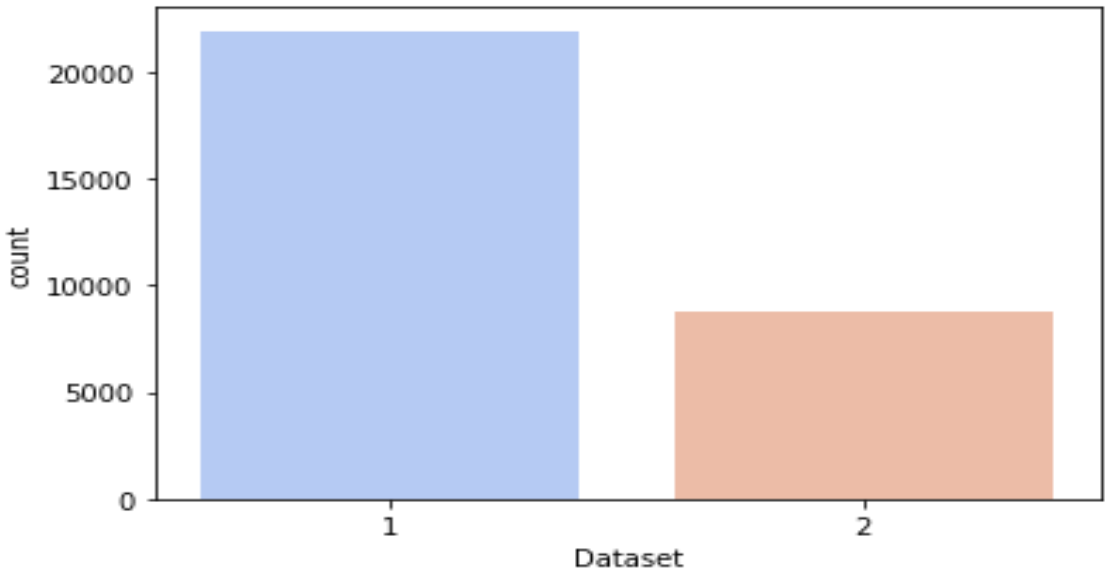


Figure 3.1: Total number of liver disease and non-liver disease patient values is given here. Column 1 represent the number of liver disease patient and column 2 represent the number of non-liver disease patient.

3.3 Machine Learning Models

To classify Liver disease we use five supervised machine learning; those are given bellow:

- Logistic Regression
- XGBoost
- Naive Bayes
- Decision Tree
- Random Forest

3.3.1 Logistic Regression

As part of supervised learning, logistic regression is used to predict the likelihood of a target variable. In logistic regression, the target variable would be a dichotomous variable in the sense that there would be only two classes [10]. Basically, its data are binary either it can be a 1 (which indicates true/yes) or a 0 (which indicates false/no). Equation of logistic regression is,

$$y = log(p/1 - p) \tag{3.1}$$

3.3.2 Naive Bayes

Generally, the working procedure of Naive Bayes is designed based on Bayes' Theorem. Naive Bayes is not a single classification algorithm, it's a collection of multiple classification algorithms [11]. In Naive Bayes, multiple classification algorithms shared a common principle to train the model effectively. Here we gave the main working equation of Naive Bayes,

$$P(A/B) = P(B/A)P(A)/P(B) \tag{3.2}$$

3.3.3 XGBoost

In order to refer the engineering goal and push the limits of boosted tree algorithms with limited computing resources XGBoost is used. Gradient boosted decision trees can be,

$$j_m(\phi_m) = \Sigma_{i=1}^{n} L(y_i, \hat{f}^{(m-1)}(x_i) + (\phi_m)(x_i)) \tag{3.3}$$

implemented by using XGBoost that can be faster and efficient than decision trees or random forests [12]. Equation of XGBoost is given here,

3.3.4 Decision Tree

A decision tree is a chart that illustrates the potential outcomes of a series of choices. Using decision trees individuals or organizations can compare the costs, risks, and benefits of possible actions. In a decision tree, each node represents the possible outcome [13]. In Figure-3.2 we gave a simple decision tree example, which can clear the concept of the decision tree.

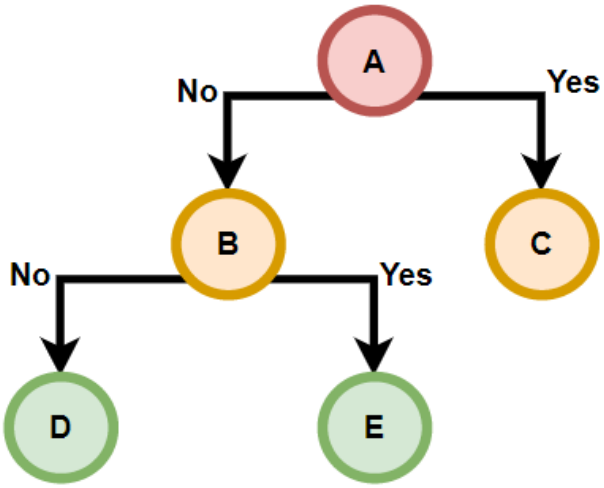


Figure 3.2: A simple example of the decision tree is given here. Node-A is the start node or parent node and B and C are the leaf or child node of A, Similarly D and E nodes are the leaf or child node of B.

3.3.5 Random Forest

Random decision forests or random forests are constructed by constructing as many decision trees as possible. Then, these decision trees are used to learn classification and regression functions. Using random forest, the class selected by most trees is the output for classification tasks. The average or mean predictions of individual trees are returned for the regression tasks [14]. The random decision forest corrects for a decision tree's tendency to overfit its training set [15]. In general, random forests are more accurate than decision

trees, but a gradient-boosted tree may be more accurate. Random forests use mean squared error (MSE) to solve regression problems.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}\left(y_d - y_i\right)^2 \tag{3.4}$$

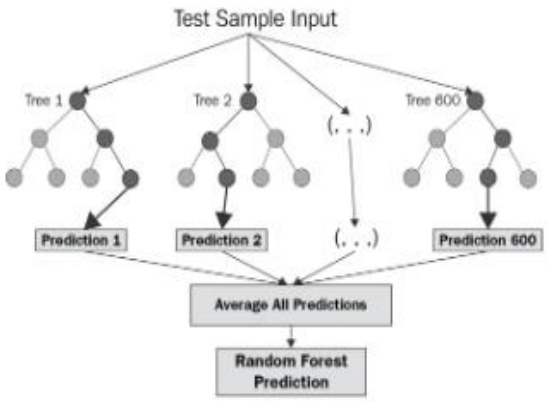


Figure 3.3: A simple example of the random forests is given here. A random forest is a collection of two or more decision trees.

Chapter 4

METHODOLOGY

A methodology is the combination of logically related methods and step by step techniques for successful planning, control and delivery of the project. In figure-4.1 we see a sort summary of our all step of this project. We divided our whole methodology in six steps. Those steps are described below,

- Dataset loading.
- Applying Various Data Mining Technique.
- Scaling Dataset.
- Dataset splitting.
- Machine Learning Model Training.
- Prediction.

4.1 Dataset loading

In this step we read the dataset. To read dataset we use pandas module of python. In this project we use Kaggle Liver Disease Patient Dataset [7]. This dataset consists of total 30000 people data about liver disease. Among 30000, 21917 patients have liver disease the rest of 8774 people don't have any liver disease. The total number of column is 11. The first 10 column of this dataset are main features of this dataset. The last column is the label or prediction column of the dataset.

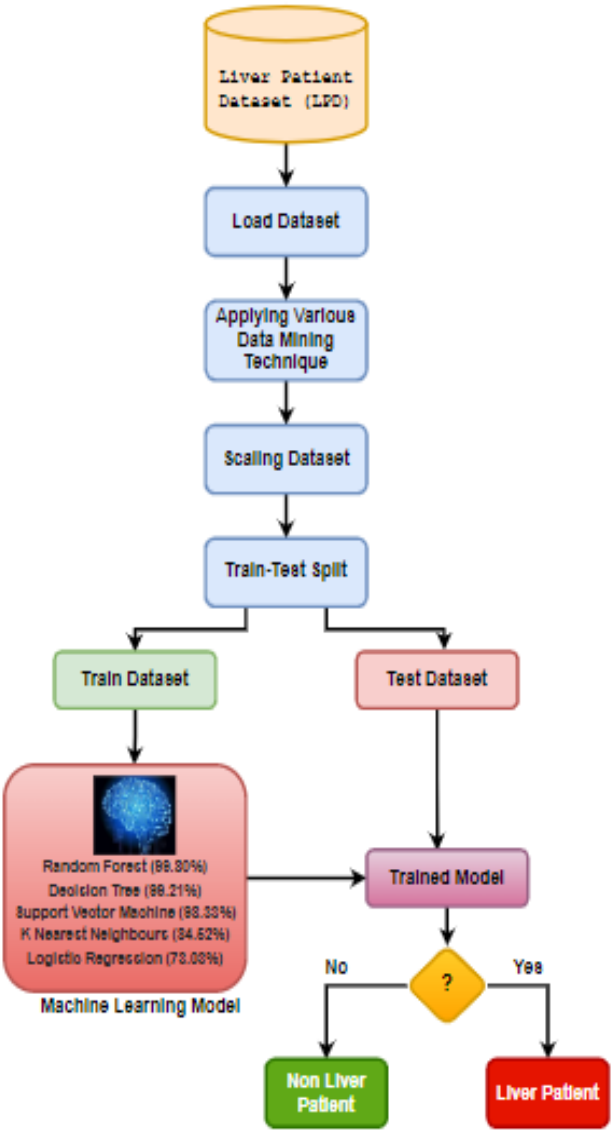


Figure 4.1: System architecture of our Analysis of different machine learning models for Liver disease prediction.

4.2 Applying Various Data Mining Technique

In this step we apply various data mining technique to process the dataset. Generally, in this step we detect the correlation between the features of the dataset and find the outliers of this dataset. Figure-4.2 give the picture of correlation between the features of this

dataset. The dataset is cleaned and there are no outliers present in this dataset. In this step we also find the null value of the dataset and fill them into their average value.

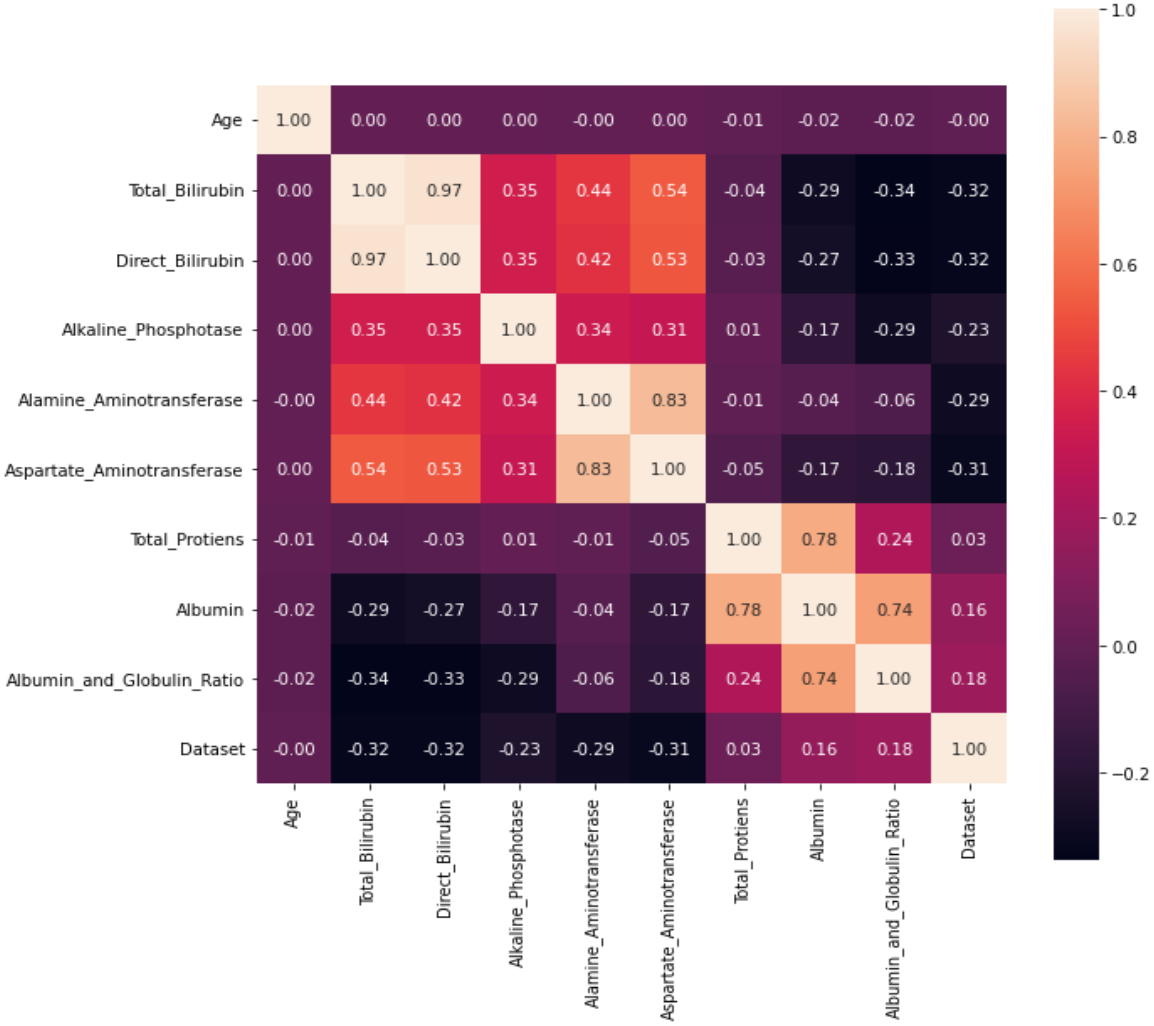


Figure 4.2: This figure depicts the correlation between the features of the dataset.

4.3 Scaling Dataset

Scaling Dataset consist of outlier detection and dataset value minimize a defined range. There is no outlier present in this dataset. So, we don't need to remove outlier in this dataset. But there are sum value is not in range like gender have to object type data "Male" and "Female" so we change "Male" as 0 and "Female" as 1. We also use StandardScaler method of sklearn library to minimize the integer and float value into a range.

4.4 Dataset splitting

In this step we split the whole dataset into two different datasets. We use train test split method of sklearn library to split the dataset. The first one of the splatted datasets is training dataset and the second one is testing dataset. Training dataset is used for train the machine learning model and the testing dataset is used for test the model accuracy [13].

4.5 Machine Learning Model Training

In this project we use five machine learning model to predict liver disease. Those five models are,

- Logistic Regression.
- Support Vector Machine (SVM).
- Decision Tree.
- Random Forest.
- K Nearest Neighbors.

All the machine learning model is trained by Kaggle Liver Disease Patient Dataset [17]. Most of them is give very good result. Among them Random Forest give the highest accuracy of 99.80%. The rest of the model accuracy is Decision Tree (99.21%) Support Vector Machine (98.33%) K Nearest Neighbors (84.52%) Logistic Regression (73.03%).

4.6 Prediction

Prediction steps consist of machine learning model testing and predicting the label of testing dataset. Here we use the testing dataset that was splitting in dataset splitting step [18]. We already told that Random Forest give the highest accuracy of 99.80%. So, we use random forest to test the testing dataset and build the main project with this model.

Chapter 5

RESULT

5.1 EVALUATION MATRIX

5.1.1 Confusion Matrix

A confusion matrix is a table that indicates how well an algorithm works on test data where the true values are known. As you can see, there are only two classes in,
the confusion matrix: “true/yes” and “false/no”. We represented true or false as “1” or “0” in our confusion matrix. As part of this calculation, true positives will be referred to as TP’s, true negatives as TN, false positives as FP, and false negatives as FN. Figure-5.1 shows all the models confusion matrix.

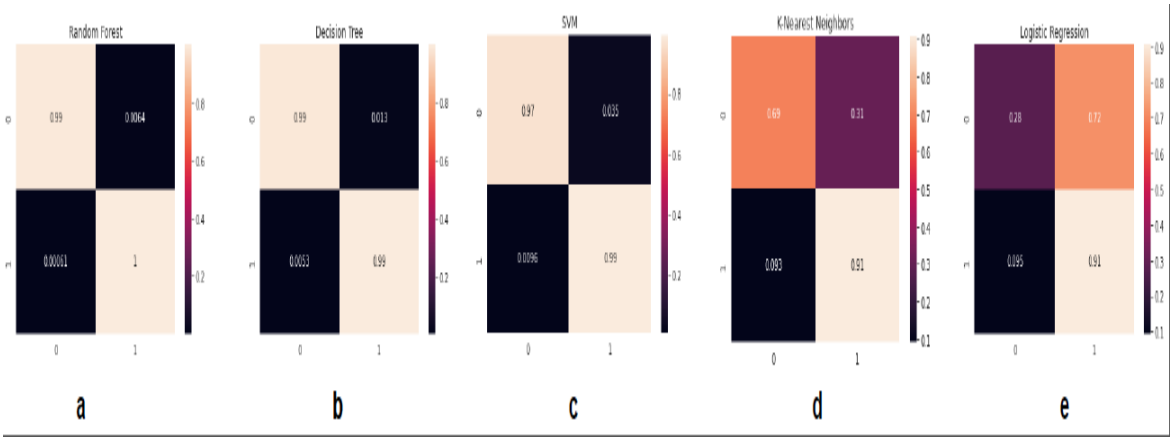


Figure 5.1: Here we saw all the model confusion matrix.

Accuracy: The successfully predicted divided by the total number of predictions is how accurate a model is. This is also known as accuracy. Equation of accuracy is 19],

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (5.1)$$

Precision: Precision means the ratio of total Positive samples correctly classified to the total number of samples classified as Positive. Equation of precision is [19].

$$Precision = TP/(TP + FP) \quad (5.2)$$

Recall: The total number of Positive samples accurately categorized as Positive divided by the total number of Positive samples is known as recall. Equation of recall is [19],

$$Rcall = TP/(TP + FN) \quad (5.3)$$

5.1.2 Receiver operating characteristic (ROC) Curve

A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds[16]. This curve plots two parameters:

- True Positive Rate.
- False Positive Rate.

True Positive Rate (TPR): True Positive Rate (TPR) is a synonym for recall and is therefore defined as follows:

$$TPR = TP/(TP + FN) \quad (5.4)$$

False Positive Rate (FPR): False Positive Rate (FPR) is defined as follows:

$$FPR = FP/(FP + TN) \quad (5.5)$$

A ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. Figure-5.2 shows all the models confusion matrix.

5.2 RESULT ANALYSIS

All the machine learning model is trained by Kaggle Liver Disease Patient Dataset [21]. Most of them is give very good result. Among them Random Forest give the highest accuracy of 99.80%. In Table-5.1 and Figurte-5.3 we give all the model accuracy, precision, recall, F1_Score, Mean_Absolute_Error, and Root_Mean_Square_Error.

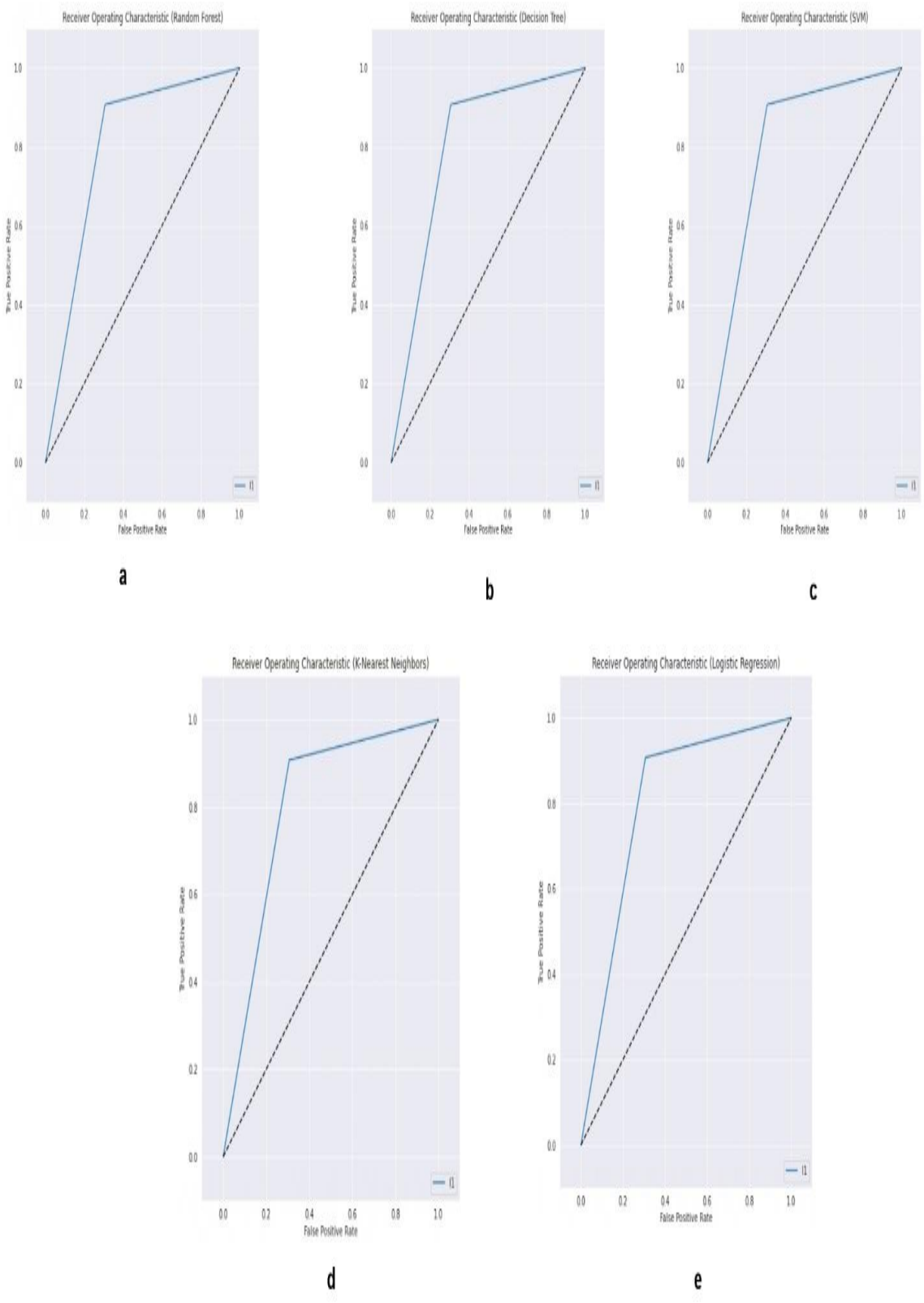


Figure 5.2: Here we saw all the model ROC curve.

Table 5.1: All the model accuracy, precision, recall, F1 Score, Mean Absolute Error, and Root Mean Square Error is given here.

ML Model	Accuracy	Precision	Recall	F1 Score	MAE	RMSE
Random Forest	99.77%	99.74%	99.94%	99.84%	0.228063	4.775590
Decision Tree	99.24%	99.47%	99.47%	99.47%	0.760209	8.718994
Support Vector Machine	98.32%	98.60%	99.04%	98.82%	1.683319	12.974278
K Nearest Neighbours	84.57%	87.98%	90.73%	89.33%	15.432233	39.283881
Logistic Regression	72.52%	75.66%	90.55%	82.44%	27.476108	52.417657

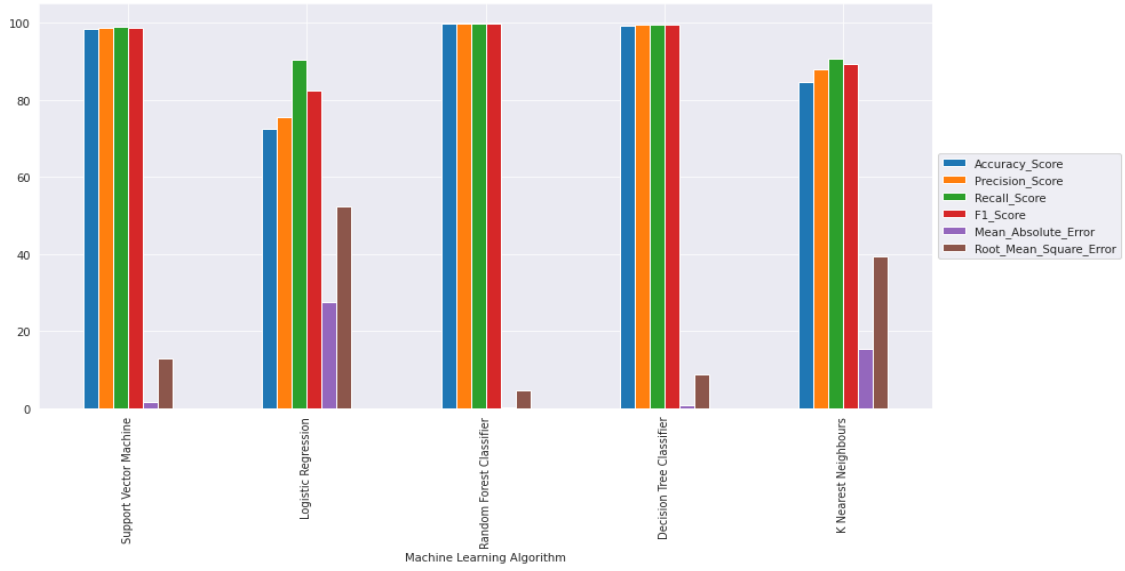


Figure 5.3: Here we saw all the model accuracy, precision, recall, F1 Score, Mean Absolute Error, and Root Mean Square Error.

5.3 Summary

After processing dataset, we fed it into above mentioned machine learning classifiers one by one. Firstly, we run the classifiers on raw dataset after processing, assessed results and run a comparative analysis among the classifiers. Then we ran classifiers again after selecting useful features to improve performance of our existing classifiers. We have done an elaborate experiment on all the classifiers mentioned above and found KNN as the best performing classifier. Performance Comparison is made among these classification algorithms before and after applying feature selection. From the analysis of each algorithm we can say that, at first, we get most accuracy in Support Vector Machine.

Chapter 6

STANDARDS, CHALLENGES AND CONSTRAINTS

6.1 Sustainability

Data mining is a rapidly growing industry in this technology trend world. Everyone requires the data to be managed in an appropriate manner and up in the right approach in order to obtain useful and accurate information [22].

Using data mining for sustainable data management will:

- Reduce the consequences of unmanaged data growth with sustainable practices.
- Lower data storage costs with unified data protection.
- Decrease the risk of data loss or theft with global data visibility.
- Ensure regulatory compliance with comprehensive enterprise data governance.
- Embrace a cleaner digital environment for a cleaner natural one.

6.2 Challenges

These days Data Mining and information disclosure are developing a critical innovation for researchers and businesses in numerous spaces. Data Mining was forming into a setup and confided in control, as yet forthcoming data mining challenges must be tackled. Some of the Data mining challenges are Security and Social Challenges, Noisy and Incomplete Data, Distributed Data, Complex Data, Performance, Improvement of Mining Algorithms, Incorporation of Background Knowledge, Data Visualization, Data Privacy and Security, User Interface, Mining dependent on Level of Abstraction, Integration of Background Knowledge, Methodology Challenges.

6.3 Project Constraints

A constraint is something that limits or controls what you can do. Their decision to abandon the trip was made because of financial constraints.

6.3.1 Design Constraints

Design constraints are those inflicted on the design solution and it is the limitations on a design. Implementing cryptographic solution needs hands-on experience about the algorithms. When we got the opportunity to work on it, it was very difficult to understand the terminologies, equally mundane to remember algorithm names and very tough to design a solution using cryptography. It needed significant amount of time to spend at search engines, still it did not match with the understanding acquired by resolving production issues, understanding existing designs and making others understand how does it ensure security.

6.3.2 Component Constraints

The component requirements of the proposed architecture include,

- Minimum Processor Requirement: Intel i3 (7th Gen, 3GHz)
- Minimum Memory Requirement: 4GB (DDR3, 1600 bus)
- Minimum Platform Requirement: Google Colab

6.3.3 Budget Constraints

The estimated budget is to be calculated by the current market price of the component requirements.

Chapter 7

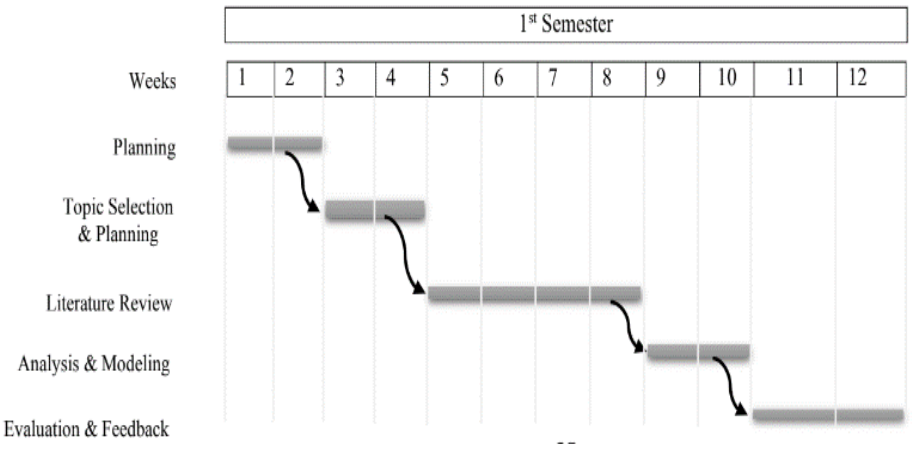
SCHEDULES, TASKS AND MILESTONES

7.1 Timeline

Our project work is separated into three sections due to the fact that we have three semesters to complete it. Our work was completed according to our supervisor's instructions. We submitted a proposal and examined associated thesis work during the first semester. In addition, we created a prototype for the planned systems by analyzing and planning with existing systems. In the second section, we partially implemented it in the second semester. Finally, we developed the overall design and reported the overall workflow in the third semester.

7.2 Gantt Chart

The following Gantt chart (Figure 7.1) represents the work execution process to complete this project work. The project work is completed within three semesters, where per semester is four-month which means 12 months in total.



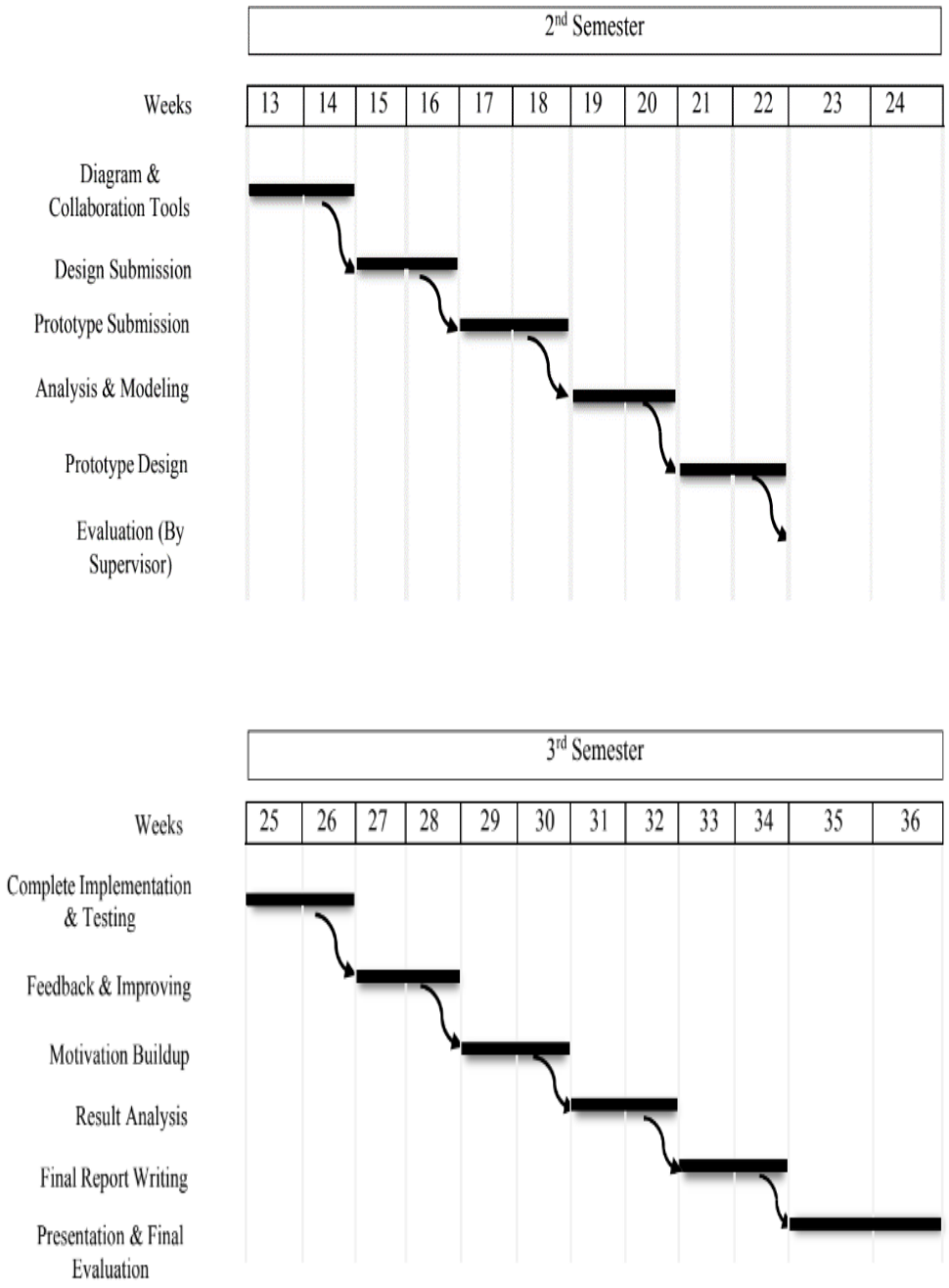


Figure 7.1. Gantt chart of the work execution process.

Chapter 8

CONCLUSION AND FUTURE WORK

8.1 Conclusion

Classification is the major data mining technique which is primarily used in health-care sectors for medical diagnosis and predicting diseases. This research work used classification algorithms namely Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree, and K Nearest Neighbors classifier for liver disease prediction. Comparisons of these algorithms are done and it is based on the performance factors classification accuracy and execution time. From the experimental results, this work concludes, the Random Forest classifier is considered as a best algorithm because of its highest classification accuracy. On the other hand, while comparing the execution time, the Logistic Regression classifier needs minimum execution time.

8.2 Future work

Nothing is 100% correct in this world. So, our work is not done yet. We need to improve our model using various technique. Some of them is given bellow,

- Adding some ensemble method to increase accuracy of our model.
- Applying data synthesis method to increase dataset.
- Developing an application for android devices that works on the same database.
- Putting this model on network and updating it through internet.

References

[1] Bendi Venkata Ramana, M Surendra Prasad Babu, NB Venkateswarlu, et al. A criticalstudy of selected classification algorithms for liver disease diagnosis. International Journal of Database Management Systems, 3(2):101–114, 2011.

[2] S Dhamodharan. Liver disease prediction using bayesian classification. 2016.

[3] S Karthik, A Priyadarishini, J Anuradha, and BK Tripathy. Classification and rule extraction using rough set for diagnosis of liver disease and its types. Adv Appl Sci Res,2(3):334–345, 2011.

[4] AH Roslina and A Noraziah. Prediction of hepatitis prognosis using support vector machines and wrapper method. In 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, volume 5, pages 2209–2211. IEEE, 2010.

[5] Omar S Soliman and Eman Abo Elhamd. Classification of hepatitis c virus using modified particle swarm optimization and least squares support vector machine. International Journal of Scientific & Engineering Research, 5(3):122, 2014.

[6] Chaitrali S Dangare and Sulabha S Apte. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10):44–48, 2012.

[7] Abhishek Shrivastava. Liver Disease Patient Dataset 30K train data. https://www.kaggle.com/abhi8923shriv/liver-disease-patient-dataset, 2021. [Online; accessed 20-March-2022].

[8] Tin Kam Ho. Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE,1995.

[9] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. springer series in statistics. In: Springer, 2001.

[10] Ahmed MT, Imtiaz MN, and Mitu NS. (2020). Impact of weather on crops in few northern parts of Bangladesh: HCI and machine learning based approach, Aust. J. Eng. Innov. Technol., 2(1), 7-15. https://doi.org/10.34104/ajeit.020.07015 .

[11] Aneeshkumar A. S. and Venkateswaran C. J. (2012). Estimating the surveillance of liver disorder using classification algorithms. International Journal of Computer Applic-

ations, 57(6), Pp. 0975-8887. https://www.ijcaonline.org/archives/volume57/number6/9121-3281 .

[12] Dhamodharan, S. (2014), Liver disease prediction using bayesian classification. In 4th National Conference on Advanced com-puting, applications & Technologies, pp. 1-3.

[13] A, Vohra R, and Rani P. (2014). Liver Patient Classification Using Intelligent Techniques. International J. of Computer Science and Information Techno-logies, 5(4), 5110-5115.http://www.ijcsit.com/docs/Volume%205/vol5issue04/ijcsit2014050462.pdf

[14] Karthik, S., Priyadarishini, A., Anuradha, J. and Tripathy, B.K. (2011). Classification and rule extraction using rough set for diagnosis of liver disease and its types. Adv Appl Sci Res, 2(3), pp.334-345. https://www.researchgate.net/publication/318645 553

[15] Liu, K.H. and Huang, D.S., 2008. Cancer classification using rotation forest. Computers in biology and medicine, 38(5), pp.601-610.

https://doi.org/10.1016/j.compbiomed.2008.02.00 7

[16] Pahariyavohra J, Makhijani J. and Patsariya S. (2014). Liver patient classification using intelligence techniques. International journal of advanced research in computer science and software engineering. 4(2): 295-299.

[17] Rahman, A. S., Shamrat, F. J. M., Tasnim, Z., Roy, J. and Hossain, S. A. (2019). A Comparative Study on Liver Disease Predic- tion Using Supervised Machine Learning Algorithms. International J. of Scientific & Technology Research, 8(11), pp.419-422.

[18] Ramana BV, Babu MSP, Venkateswarlu NB. (2012). Liver Classification Using Modified Rotation Forest. International Journal of Engineering Research and Development, 1(6), 17-24.

[19] Ramana B. V., Babu M. S. P. Venkateswarlu N. B. (2011). A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. International Journal of Data-base Management Systems (IJDMS), 3(2), 101-114. https://pdfs.semanticscholar.org/c92d/38a7a76 c20a317de63fb9278bb10102c758b.pdf

[20] Rifkin R, Mukherjee S, Pablo, Tamayo, P, and Mesirov JP. (2003). An Analytical Method For Multi-Class Molecular Cancer Classification, SIAM Review 45(4): 706-723. https://doi.org/10.1137/S0036144502411986

[21] Rajeswari P., Reena G. S. (2010). Analysis Of Liver Disorder Using Data Mining Algorithm. Global Journal Of Computer Science And Technology, 10(14). Pp. 48-52. https://core.ac.uk/download/pdf/231162636.pdf

[22] Rifkin R, Mukherjee S, Pablo, Tamayo, P, and Mesirov JP. (2003). An Analytical Method For Multi-Class Molecular Cancer Classification, SIAM Review 45(4): 706-723. https://towardsdatascience.com/data-mining-for-sustainable-data-management-659e7c7ea41e

[23] Rosalina. A.H, Noraziah. A. (2010). Pre- diction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method, IEEE, Pp.2209-22. https://doi.org/10.1109/FSKD.2010.5569542

[24] Russell, S. and Norvig, P. (2002). Artificial intelligence: a modern approach. Prentice Hall, Englewood Cliffs, New Jersey 07632. https://www.cin.ufpe.br/~tfl2/artificial-intellige nce-modern-approach.9780131038059.2536 8.pdf

[25] Schiff, E. R., Sorrell, M. F., & Maddrey, W. C. (2007). Schiff's Diseases of the Liver (10 ed.). https://miami.pure.elsevier.com/en/publications /schiffs-diseases-of-the-liver.