DIABETES MALADY PREDICTION USING DATA MINING ALGORITHMS

BY

Syeda Saida Haider

ID: 171-35-174

This Report Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Software Engineering.

Supervised By

Nayeem Hasan

Senior Lecturer

Department of Software Engineering

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JUNE, 2021

APPROVAL

This thesis titled "Diabetes Malady Prediction Using Data Mining Algorithms", submitted by Syeda Saida Haider to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Software Engineering (SWE). The presentation has been held on June, 2021and approved as to its style and contents.

DECLARATION

I hereby declare that this thesis work has been completed by me under the supervision of Nayeem Hasan, Senior Lecturer, Department of Software Engineering, Daffodil International University. I also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for award of any degree.

Supervised by:

Nayeem Hasan
Senior Lecturer
Department of SWE
Daffodil International University

Submitted by:

Saidahaider

Syeda Saida Haider
ID: 171-35-174
Department of SWE
Daffodil International University

ACKNOWLEDGEMENT

Firstly, I want to disclose gratefulness to the Almighty Allah for all of the blessings that have enabled me to effectively finish my final year thesis work.

I am thankful to my supervisor Nayeem Hasan, Senior Lecturer, Department of SWE Daffodil International University, Dhaka. His encouragement assisted me in completing this research work effectively.

I would like to convey gratitude to all of my teachers and classmates at Daffodil International University for all of their supports.

Finally, I want to express my gratitude to my parents for their enormous support and inspiration.

ABSTRACT

Diabetes is considered as one of the world's most prevalent incurable diseases.422 million individuals around the world are affected by incurable diabetes malady is reported by the World Health Organization. Therapy can be amplified by anticipating diabetes malady at an early echelon. The techniques conversant to data mining are used extensively to anticipate diabetes at a preliminary phase. Individual’s chances of having a diabetes malady can be anticipated by some momentous attributes that are playing a crucial preamble to enumerate diabetes malady at an early echelon. Symptoms data of the forthwith diabetes malady invaded people or no diabetes invaded people have been used to predict the diabetes disease at an early period in this diabetes malady anticipating work. On the diabetes malady anticipation, dataset multiple data mining algorithms such as Random Forest (RF), Decision Tree, Extra Trees, XGBoost, and Bagging have been implemented. Data mining algorithms accuracy has been assimilated in this diabetes malady prediction work. Among the mentioned data mining algorithms the Random Forest provided the best accuracy while using the Percentage split technique. Ultimately, it can be said that to aid the diabetes malady anticipation operation the Random Forest is more suitable for this research work than the others algorithms which are used in this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

1. RF= Random Forest
2. ANN = Artificial Neural Network
3. KNN= K-Nearest Neighbors
4. LR = Logistic Regression
5. SVM = Support Vector Machines
6. NB = Naïve Bayes
7. XGBoost = Extreme Gradient Boosting
8. ROC = Receiver Operator Characteristic
9. AUC = Area Under the Curve

CHAPTER 1
INTRODUCTION

1.1 Introduction

Diabetes is a metabolic distemper that impels blood sugar levels to rise at an extreme level. The human body's organs can be impaired by the lofty dimension of blood sugar that diabetes malady impels. Peoples with diabetes are unable to use the insulin efficiently that their bodies generate, or else their physiques are unable to anticipate a sufficient amount of insulin (Watson, 2020). Various form of diabetes is affecting people's, they are Type One diabetes, Diabetes Secondary, and Type Two diabetes. The type of diabetes which is dependent on insulin is known as Type One Diabetes. This type of diabetes originates when the pancreas is invaded by the body along with the antibodies. Children are most commonly affected by type 1 diabetes. On the other side, type 2 diabetes is noninsulin-dependent. When an individual is affected by type 2 diabetes their body becomes insulin resistant. Adults are mainly attacked by type 2 diabetes. Another form of diabetes is Diabetes secondary. It occurs because of the outcome of another disease such as cystic fibrosis, chronic pancreatitis, and endocrine abnormalities (Klöppel et al., 1985).

According to the data of the World Health Organization (WHO), the number of people affected by diabetes incremented from 108 million in 1980 to 422 million in 2014 (World Health Organization,2021). 463 million patients who are affected by diabetes disease is predicted to survive worldwide and by 2045 this number will be increased to 700 million people globally (Elflein, 2021). Diabetes has become more prevalent in the Asian countries and over sixty percent of the world's diabetes patients are added by this continent. The people of the Asian continent have a rigid ethnic to diabetes and genetically they have a propensity for diabetes diseases .In consequence, diabetes is developed by them from an early age (Ramachandran et al., 2012).

Diabetes malady has been one of the causes of departure around the globe in passing years. In the year 2019, 1.5 million people had died globally because of chronic diabetes diseases (Elflein, 2021). According to the information of Diabetes Australia, diabetes disease can be carried by people up to 7 years before a medical diagnosis is made. By this period people will have started to be affected by a variety of afflictions, including strokes, kidney failure, heart attacks, eye injury, blindness ,foot ulcers ,and organ mutilation. In the majority of instances, these afflictions can be confined through identifying diabetes promptly and beginning the treatment as quickly as possible and by this each year $1415 per individual could save up (Diabetes Australia, 2018). To diagnose diabetes, the consultants suggest for conduct several procedures like OGTT (In a complete form Oral Glucose Tolerance Test), and the Hba1c (In a complete form Glycosylated hemoglobin).The expenditure of the OGTT and Hba1c tests are not much affordable (Ramachandran, 2014).

The perplexity of the diabetes diseases is conventional and diabetes treatment is a cost strain for the unprivileged people of the community (Ramachandran et al., 2012).

“The process which is applied to tracing anomalies, patterns, and correlations among the data set to predict the consequence entitled data mining” (SAS, n.d.). In the health sector, data mining accomplishes a crucial role to predict diseases efficiently in recent time. Data mining corroborates a significant induction for the diabetes diseases research.Various data mining techniques assist diabetes-related research work to enhance the medical care excellence for diabetes-affected people. Diverse data mining algorithms like KNN, Artificial Neural Network, RF (Random Forest), Naïve Bayes, Decision Tree are used by the researchers for the prediction of diabetes (Amalarethinam & Vignesh, 2015).

In this research work, Random Forest, Decision Tree, Extra Trees, XGBoost, and Bagging are implemented to anticipate diabetes diseases.

1.2 Motivation

Diabetes is a chronic disorder. It has been arriving considerably in the general demography. People over 65 years are particularly affected by this disease. Day by day numerous peoples are becoming affected by this global phenomenon (LeRoith et al., 2019). Diabetes-affected people have to suffer tremendously both financially and physically. A Diabetic patient’s blood contains a high glucose level. The appearance of this elevated blood glucose harms the body. As a consequence of the diabetes malady, complications like kidney damage, eye damage, heart attack, nerve damage, and foot problems can occur in the physique (Better Health Channel, n.d.).The diabetes malady is attached with manifold death from diverse diseases like degenerative disorder, infectious diseases, and cancers (New England Journal of Medicine, 2011).It is one of the life-threatening diseases .It is the prominent cause of death around the globe, it pushed people in the path of death to take lives.

In the year 2019, 1.5 million people have died globally because of chronic diabetes diseases (Elflein, 2021).It is a costly lifelong illness. Diabetes treatment seems like a financial burden for impoverished people because of its high expenditure. Diabetes patients have to spend an ample amount of money to sustain the treatment of diabetes disease. While the diabetes malady is diagnosed at the ages of 40, 50, 60, and 65 years the discounted lifetime treatment expenditure for the diabetes patients was $124600,$91200,$453800,and$35900,respectively (Zhuo et al., 2014).The existent system which is used to predict diabetes malady can be costly for the huge amount of people .A system is extremely needed for the people of the subordinate earning country where the medical equipment, experienced medical practitioners are limited, and the diabetes diagnosis is not available .A cost worthy system is desperately required to mitigate this problem. Multiplex research committed to getting out of this problem and still too many works are required to cooperate to moderate this problem.

1.3 Diabetes Malady

Diabetes is one kind of chronic disease in which the human body is unable to use the insulin effectively or produce the insulin. People of all ages are at risk of developing diabetes. In recent times, a huge amount of people are becoming affected by diabetes.

There are several kinds of diabetes occurring around the world. The details about diabetes type are given below:

- **Type One diabetes:** This form of diabetes mostly affects kiddies and adolescents. That's why it can be deemed as juvenile diabetes disease. Due to this Type 1 diabetes, the human physique is unable to manufacture insulin.
- **Type Two diabetes:** Around the globe this conjugation of diabetes is of utmost flourishing. Especially adults are the sufferers of this kind of diabetes. As a result of Type 2 diabetes, the human body has failed to utilize insulin effectively.
- **Diabetes Secondary:** This kind of diabetes is caused in the human body as fallout of further disorders like endocrine abnormalities, cystic fibrosis, and so on.

Numerous symptoms arrive when a person is affected by diabetes such as:

- Urination is much more frequent than usual.

- Drastic thirstiness arrives.
- People started to lose weight abruptly.
- The healing process becomes slow.
- Hard to sight plainly.
- Hair loss problems begin.
- Tiredness appears in the body.
- Hungriness increases.
- The muscle becomes numb.
- Abundance fat accumulates in the body.
- Itching problem attains in the physique.
- The mood becomes cranky.

The aforementioned symptoms are related to the diabetes malady. When people are affected by diabetes malady, they can perceive the appearance of the above-mentioned symptoms in their bodies.

1.4 Report Layout of the Research Work

This research report comprises the below contents:

- ❑ Chapter 1 provides the introduction regarding the research topic with the motivation behind this work. This chapter also gives a brief idea about the diabetes malady.
- ❑ Chapter 2 analyzes the related work on diabetes disease.
- ❑ Chapter 3 contains the details about the data collection and the different attributes of the dataset.
- ❑ Chapter 4 provides the idea about different data mining algorithms which are used in this research work.
- ❑ Chapter five shows the performance of the various models, a comparative exploration of the model's performance of this research with the existing works.
- ❑ Chapter 6 contains a summary of this diabetes prediction research work.

CHAPTER 2
LITERATURE SURVEY

2.1 Introduction

Abundant works have already been done on Diabetes malady prediction. Different data mining algorithms have been utilized to predict diabetes and to obtain the most excellent performance.

2.2 Related Work

Mahboob Alam et al. have implemented multiple algorithms to predict diabetes. The algorithms are ANN, K-Means Clustering algorithms, and the RF (Random Forest). The ANN comes up with the best accuracy of 75.7% (Mahboob Alam et al., 2019).

Islam et al. have applied Ten-fold Cross-Validation and Percentage Split techniques to assess the effectiveness of the algorithms. They explored the dataset by using several algorithms like RF (Random Forest), J48 Decision Tree, Naïve Bayes Algorithm, and Logistic Regression Algorithm. RF achieved the best accuracy, 99% for the Percentage Split technique and 97.4% for the Cross–Validation technique (Islam et al., 2019).

Alpan and Ilgi collected 520 data for the experiment. They applied 7 diverse algorithms they are Support Vector Machines, Random Tree (RT), K-Nearest Neighbors (KNN), Bayes Network, Random Forest, Naïve Bayes, and the Decision Tree. Afterward applying the Cross-validation technique KNN has been found to have the best accuracy with 98.07% corrected instances (Alpan & Ilgi, 2020).

Chaves and Marques tested diverse algorithms to achieve the accuracy of the algorithms. They applied Random Forest, Neural Network, K-Nearest Neighbors, AdaBoost, Naïve Bayes, and SVM (Support Vector Machines). Their study has a signal that Neural Networks gained the highest accuracy (98.08%) while using the Cross-validation technique (Chaves & Marques, 2021).

To anticipate the evaluation of the algorithm’s performance, Iyer et al. used two techniques like Cross-Validation and Percentage Split. To find out the exactness score Naïve Bayes and J48 Decision Tree algorithms are implemented by them. Their studies results indicate that Naïve Bayes achieved the highest accuracy of 79.5652 % when using Percentage Split (Iyer et al., 2015).

Sisodia and Sisodia applied algorithms such as Decision Tree, SVM, and Naive Bayes. Naive Bayes achieved the maximal exactness of 76.3% (Sisodia & Sisodia, 2018).

Silva et al. implemented SMO Support Vector Machine, Decision Tree, and Naïve Bayes to anticipate the prediction of diabetes disease. To assess the algorithms both the Percentage split and K-fold Cross-Validation had been used by them. In their work, the Decision Tree can detect the highest amount of the correct instances (84.6667%) (Silva et al.,2016).

Peker et al. used the Orange data mining toolbox to commit a data exploration. They actualized algorithms like Decision Tree, ANN, Random Forest, KNN, and SVM. ANN provides the highest accuracy in their work, with a correctness of 93.85% (Peker et al., 2018).

Malik et al. implemented 10 different data mining algorithms to predict the diabetes malady with the efficient model. Authors used Bagging, Naïve Bayes, Multi-Layer Perceptron, Decision tree, AdaBoost, SVM, KNN, Random Forest, Bayes Net, and Logistic Regression (LR). Among all the algorithms KNN obtained 98.62% highest accuracy (Malik et al., 2020).

In the year 2019, Mujumdar & Vaidehi analyzed the accuracy of several algorithms including Gaussian Naïve Bayes, Support Vector Classifier, Logistic Regression (LR), Bagging, Decision Tree Classifier, Ada Boost, Random Forest, Perceptron, Extra Tree Classifier, Linear Discriminant Analysis, KNN, and Gradient Boost Classifier. In their research work, LR provided the maximal accuracy of 96% (Mujumdar & Vaidehi, 2019).

To anticipate the prediction of diabetes malady, Mitushi Soni and Dr. Sunita Varma used multiple algorithms like SVM, KNN, Random Forest, Gradient Boosting, Decision Tree, and Logistic Regression (LR). In this author's study, Random Forest achieved the foremost accuracy of 77% (Soni & Varma, 2020).

Pradhan et al. used various algorithms like ANN, Decision tree, KNN, Naïve Bayes (NB), Support Vector Machine, and K-means algorithms to obtaining the accuracy of the algorithms and to make a comparative analysis among the accuracy of the algorithms. Authors get the highest accuracy from ANN which is 85.09% (Pradhan et al., 2020).

The Five-Fold Cross-Validation technique is used by Zou et al. to assess the execution of the algorithms. They used the Artificial Neural Network (ANN), j48 Decision Tree, and Random Forest. In their study, ANN achieved maximal accuracy of 80.84% (Zou et al., 2018).

Shuaibu et al. used Support Vector Machine (SVM), and XGBoost algorithms. The XGBoost provides the highest accuracy in their research work, with a correctness of 77% (Shuaibu et al., 2021).

To evaluate the competency of the XGBoost algorithm, Bhulakshmi and Gandhi utilized the Ten-fold Cross-Validation technique.In their study, XGBoost provided the accuracy of 81% ((Bhulakshmi & Gandhi, 2020).

CHAPTER 3

DATASET DETAILS

Symptoms data of the forthwith diabetes malady invaded people or no diabetes invaded people have been used to predict the diabetes disease at an early period in this diabetes malady anticipating work. The dataset consists of 520 instances with 17 attributes. This data gathering, which has been gathered from the Sylhet Diabetes Hospital (Islam et al., 2019) ,is publicly accessible in Data World.

3.1 Input attributes of the dataset

Table 3.1: Description of 17 input attributes

Sr. No.	Name of the attributes	Description	Values
1	Age	Age in years	16 to 90

2	Gender	Male or Female	1.Male(Yes/NO)
			2.Female(Yes/NO)

3	Polyuria	Urination frequently than usual	Yes and No
4	Polydipsia	A feeling of extreme thirstiness	Yes and No
5	Sudden weight loss	Weight loss occurs unexpectedly	Yes and No
6	Weakness	The tiredness of body	Yes and No
7	Polyphagia	The hefty intuition of hunger	Yes and No
8	Genital thrush	Usual yeast infection	Yes and No
9	Visual blurring	Tough to see evidently	Yes and No
10	Itching	A sensation that is unrestrained and vexatious	Yes and No

11	Irritability	Feeling of anxiety	Yes and No
12	Delayed healing	The curing process becomes sluggish	Yes and No
13	Partial paresis	Muscles become weak	Yes and No
14	Muscle stiffness	Muscles perceive uncomfortable also hard to move	Yes and No
15	Alopecia	Extreme hair fall problem	Yes and No
16	Obesity	Abundant fat accumulation	Yes and No
17	Class	Represents the results (Positive or Negative)	Positive and Negative

3.2 Independent attributes of the dataset

- Age
- Gender
- Polyuria
- Polydipsia
- Sudden weight loss
- Weakness
- Polyphagia
- Genital thrush
- Visual blurring
- Itching
- Irritability
- Delayed healing
- Alopecia
- Partial paresis
- Obesity

3.3 Dependent attribute of the dataset

- Class

3.4 Graphical Representations of Data

This dataset contains the data of 520 individuals. The individuals are different from each other in age, gender, and different health characteristics. In this part, diverse attributes of the dataset have been shown.

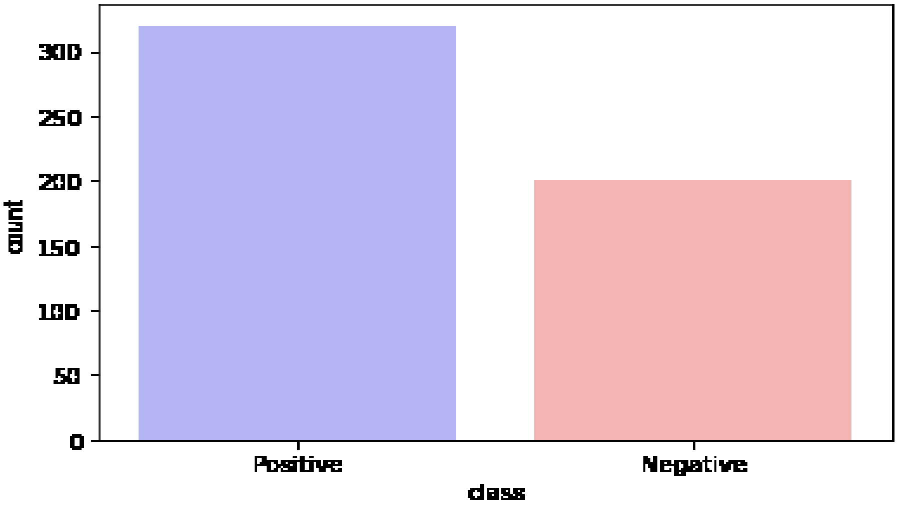


Fig 3.1: Have diabetes and doesn't have diabetes disease

The figure above shows the number of the person with diabetes and without diabetes through a bar graph.

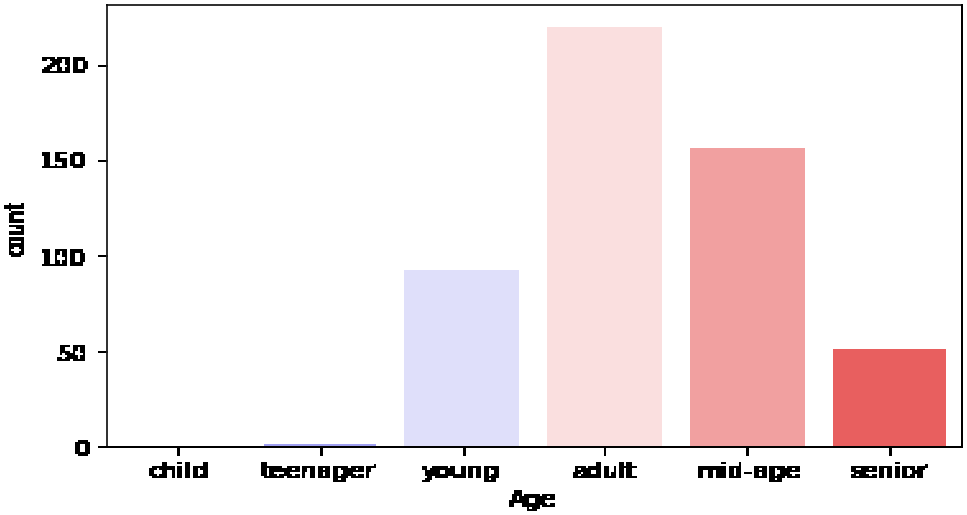


Fig 3.2: Age distribution

The above figure shows the distribution of the age which is contained in the dataset. In this figure, we can see that age is classified into six classes. Those are child, teenager, young, adult, mid-age, and senior. According to this figure, teenager's number is very few in the dataset and the maximum people in this dataset are into the adult group.

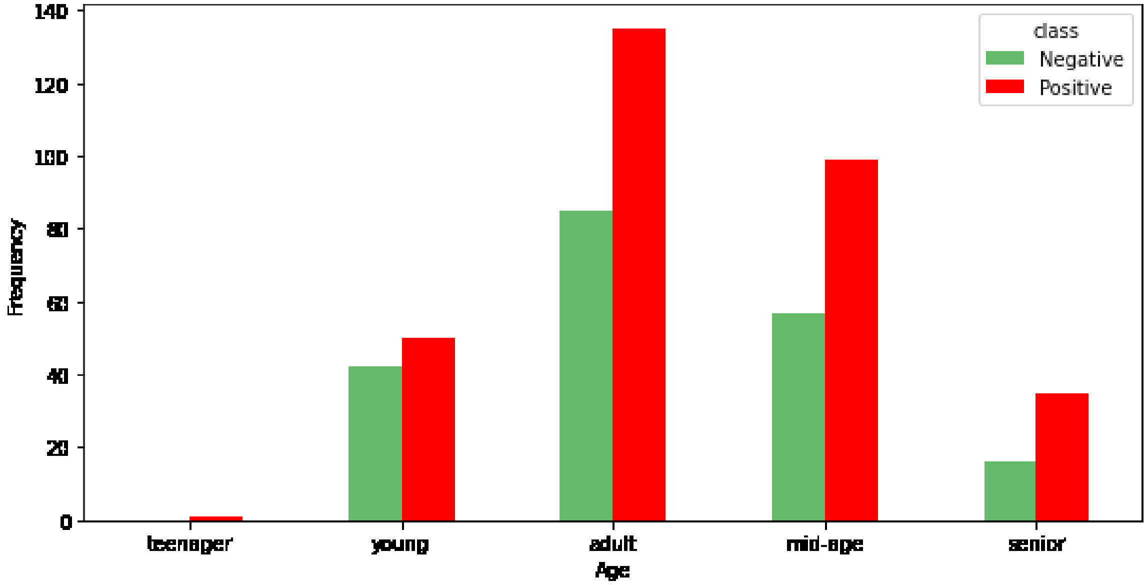


Fig 3.3: Diabetes disease association with age

The above figure shows the number of the person with diabetes and without diabetes through a bar graph, this based on their age group.

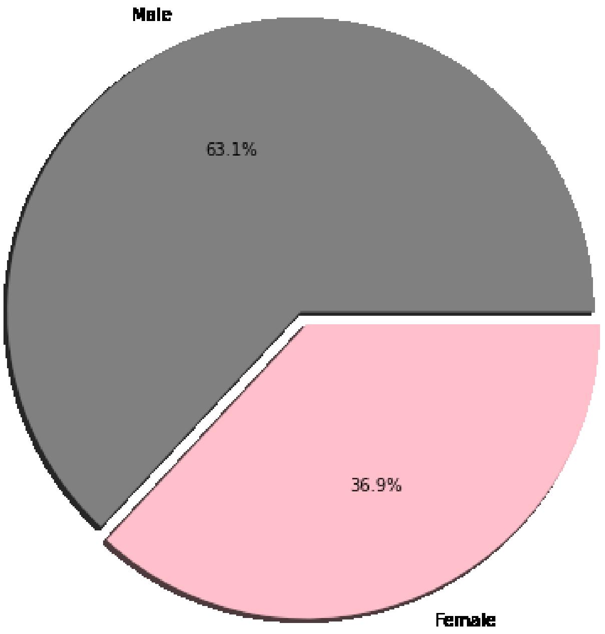


Fig 3.4: Sex distribution

The above figure shows the distribution of the sex which is contained in the dataset. We can see that there are 63.1% males and 36.9% females are in this dataset.

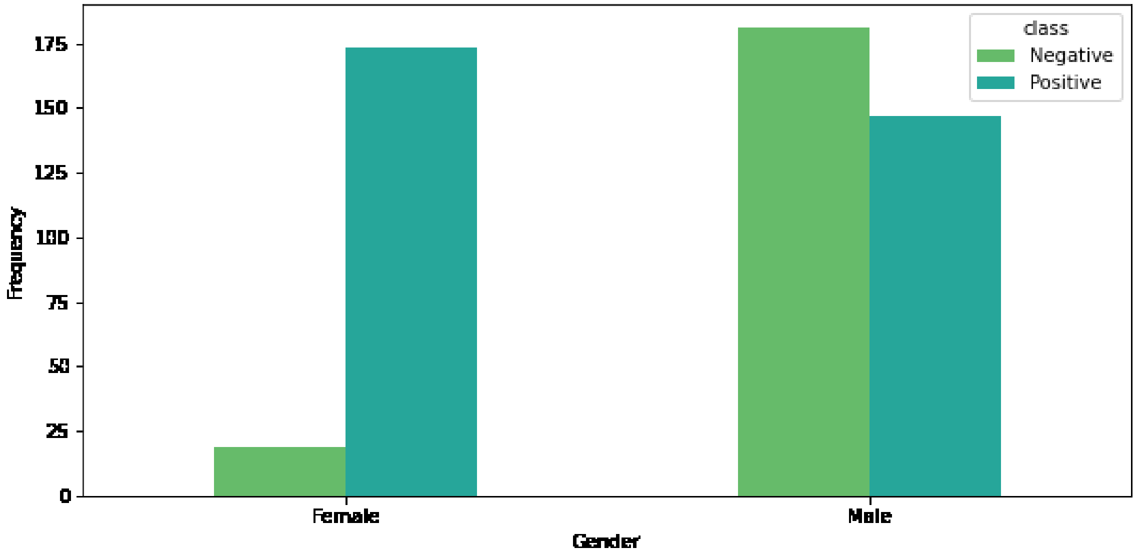


Fig 3.5: Diabetes disease association with gender

The above figure shows the number of the person with diabetes and without diabetes through a bar graph, this based on their sex.

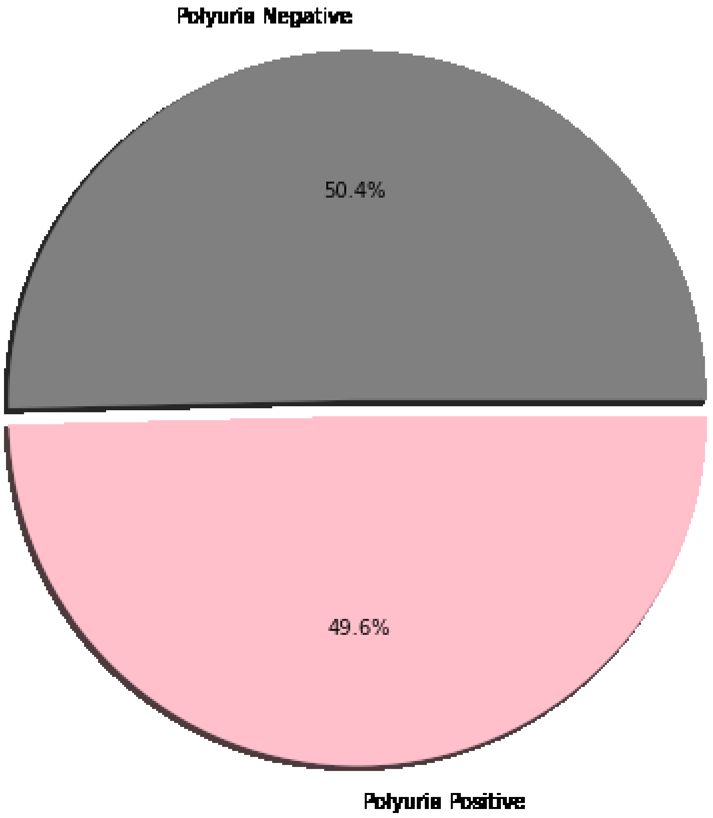


Fig 3.6: Polyuria distribution

The above figure shows the distribution of the polyuria which is contained in the dataset. We can see that in this dataset 49.6% of people have the symptoms of polyuria and 50.4% of people do not have the symptoms of polyuria.

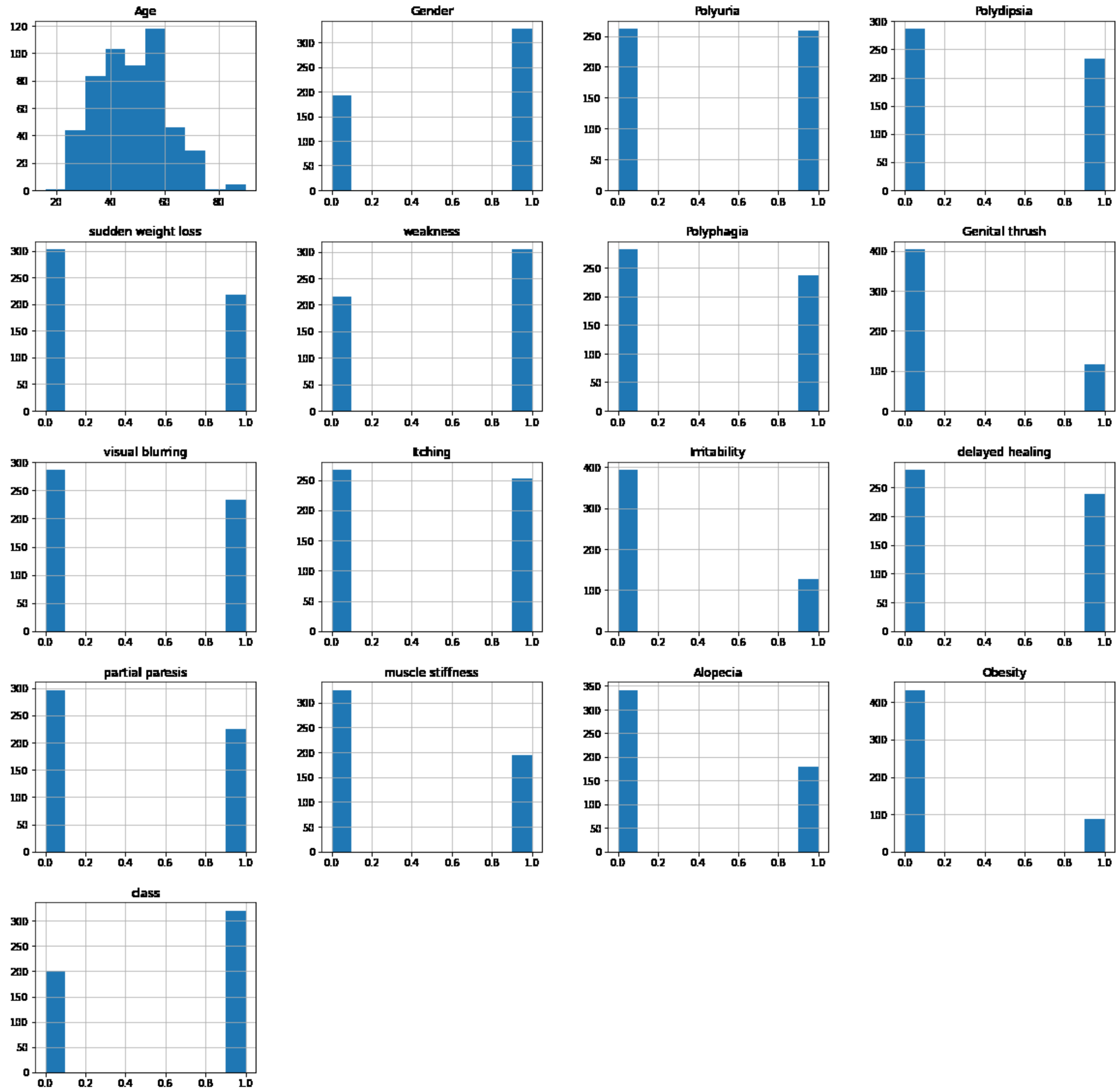


Fig 3.7: Histogram of all attributes

The above figure shows the distribution of all the attributes which are contained in the dataset.

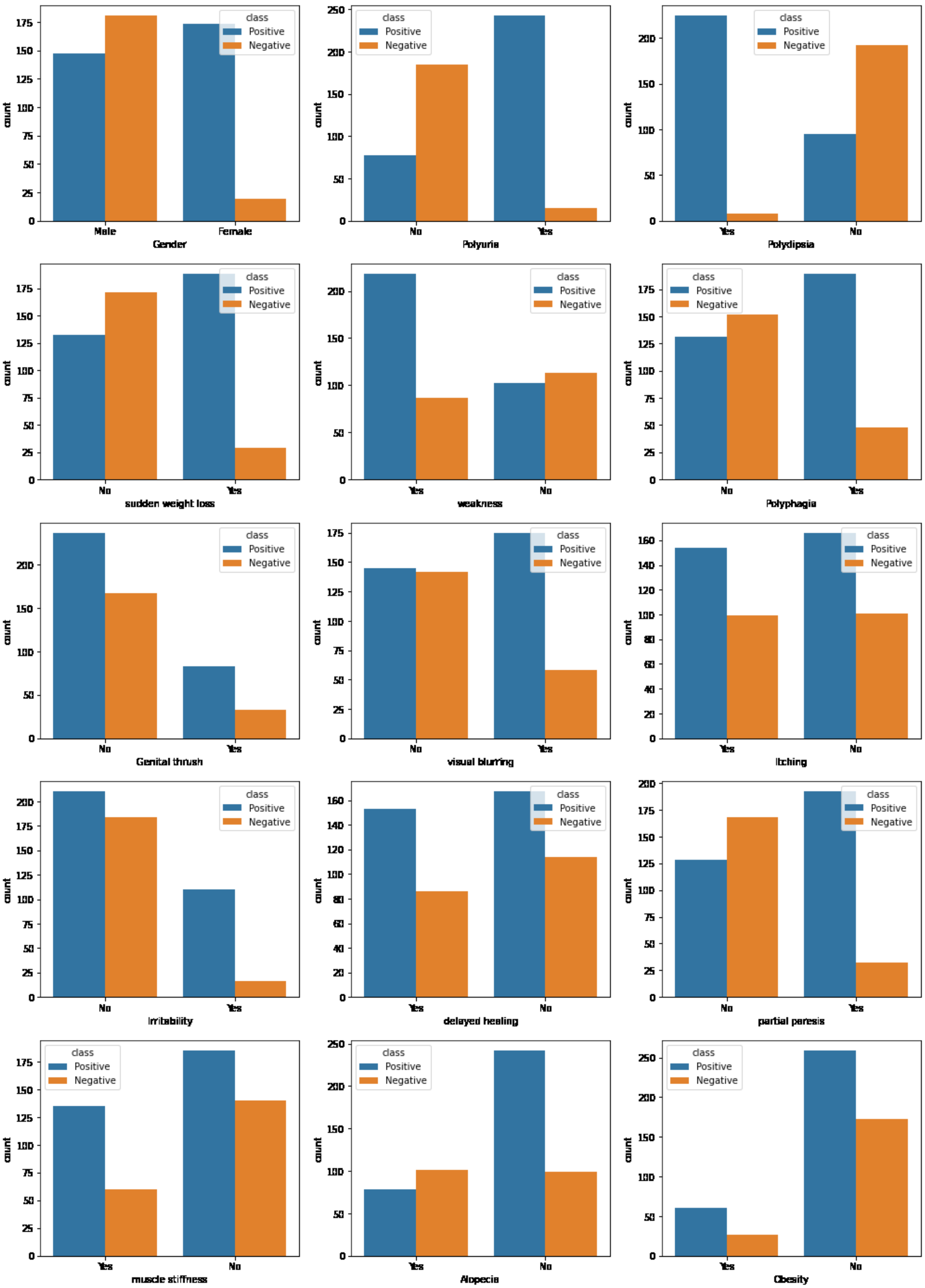
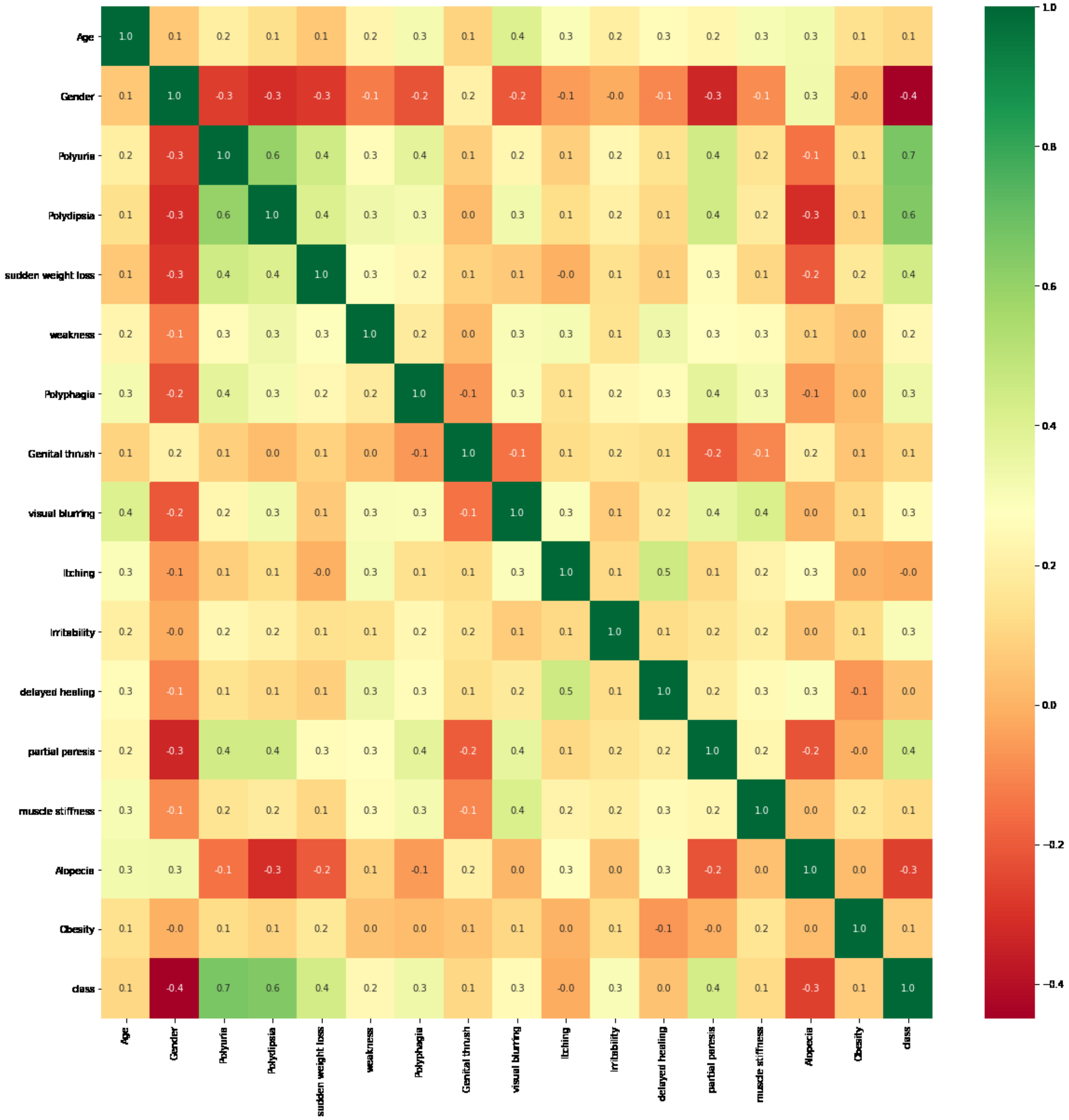


Fig 3.8: Diabetes association with all attributes

The above figure shows the number of the person with diabetes and without diabetes through the bar graphs. These graphs are made based on the all attributes of this dataset.



	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
Age	1.0	0.1	0.2	0.1	0.1	0.2	0.3	0.1	0.4	0.3	0.2	0.3	0.2	0.3	0.3	0.1	0.1
Gender	0.1	1.0	-0.3	-0.3	-0.3	-0.1	-0.2	0.2	-0.2	-0.1	-0.0	-0.1	-0.3	-0.1	0.3	-0.0	-0.4
Polyuria	0.2	-0.3	1.0	0.6	0.4	0.3	0.4	0.1	0.2	0.1	0.2	0.1	0.4	0.2	-0.1	0.1	0.7
Polydipsia	0.1	-0.3	0.6	1.0	0.4	0.3	0.3	0.0	0.3	0.1	0.2	0.1	0.4	0.2	-0.3	0.1	0.6
sudden weight loss	0.1	-0.3	0.4	0.4	1.0	0.3	0.2	0.1	0.1	-0.0	0.1	0.1	0.3	0.1	-0.2	0.2	0.4
weakness	0.2	-0.1	0.3	0.3	0.3	1.0	0.2	0.0	0.3	0.3	0.1	0.3	0.3	0.3	0.1	0.0	0.2
Polyphagia	0.3	-0.2	0.4	0.3	0.2	0.2	1.0	-0.1	0.3	0.1	0.2	0.3	0.4	0.3	-0.1	0.0	0.3
Genital thrush	0.1	0.2	0.1	0.0	0.1	0.0	-0.1	1.0	-0.1	0.1	0.2	0.1	-0.2	-0.1	0.2	0.1	0.1
visual blurring	0.4	-0.2	0.2	0.3	0.1	0.3	0.3	-0.1	1.0	0.3	0.1	0.2	0.4	0.4	0.0	0.1	0.3
Itching	0.3	-0.1	0.1	0.1	-0.0	0.3	0.1	0.1	0.3	1.0	0.1	0.5	0.1	0.2	0.3	0.0	-0.0
Irritability	0.2	-0.0	0.2	0.2	0.1	0.1	0.2	0.2	0.1	0.1	1.0	0.1	0.2	0.2	0.0	0.1	0.3
delayed healing	0.3	-0.1	0.1	0.1	0.1	0.3	0.3	0.1	0.2	0.5	0.1	1.0	0.2	0.3	0.3	-0.1	0.0
partial paresis	0.2	-0.3	0.4	0.4	0.3	0.3	0.4	-0.2	0.4	0.1	0.2	0.2	1.0	0.2	-0.2	-0.0	0.4
muscle stiffness	0.3	-0.1	0.2	0.2	0.1	0.3	0.3	-0.1	0.4	0.2	0.2	0.3	0.2	1.0	0.0	0.2	0.1
Alopecia	0.3	0.3	-0.1	-0.3	-0.2	0.1	-0.1	0.2	0.0	0.3	0.0	0.3	-0.2	0.0	1.0	0.0	-0.3
Obesity	0.1	-0.0	0.1	0.1	0.2	0.0	0.0	0.1	0.1	0.0	0.1	-0.1	-0.0	0.2	0.0	1.0	0.1
class	0.1	-0.4	0.7	0.6	0.4	0.2	0.3	0.1	0.3	-0.0	0.3	0.0	0.4	0.1	-0.3	0.1	1.0

Fig 3.9: Heatmap

Finally In figure 3.6, the Heatmap is implemented to display the relationship of the attributes which are in the dataset.

CHAPTER 4

METHODOLOGY

4.1 Introduction

Through this section of my work, I will discuss the methodology which I have followed to accomplish this research work. The methodology is a technique that delivers the reciters a complete idea about how the process of the research was performed.

4.2 Used Algorithms

Diverse data mining Algorithms are used in this work to explore the dataset and establish the model to predict the diabetes malady.

4.2.1 Random Forest

Random Forest is the data mining algorithm that is applied for regression and classification intricacy. It is a supervised learning algorithm. Random Forest is structured by multiplex decision trees. It is a more advanced variant of the decision tree. It can be considered as an ensemble method. It does not work with just a single tree. It works by attaching different decision trees to provide easy solutions for very complicated problems while also obtaining better and stable accuracy. The numerous decision trees which are in the forest can confine the overfitting complication. Through selecting the little team of the input variables the random feature is constructed. During the time of the tree growth more randomness is added in the model.

Some phases are followed in the Random Forest to continue its work process. Those phases are given below:

Phase-1: In the first phase, RF started its work by choosing the random data points from a set which is known as a training set.

Phase-2: In the second phase, constructs multiple decision trees that are attached along with the randomly chosen data points.

Phase-3: In the third phase, to build the decision trees, the number of the decision trees has to be selected.

Phase-4: In this phase, phases 2 and 3 have to be repeated.

Phase-5: In this phase, the prognosis of every decision tree has to be detected. The class which conquers the maximal votes, data points has to be inserted in this class.

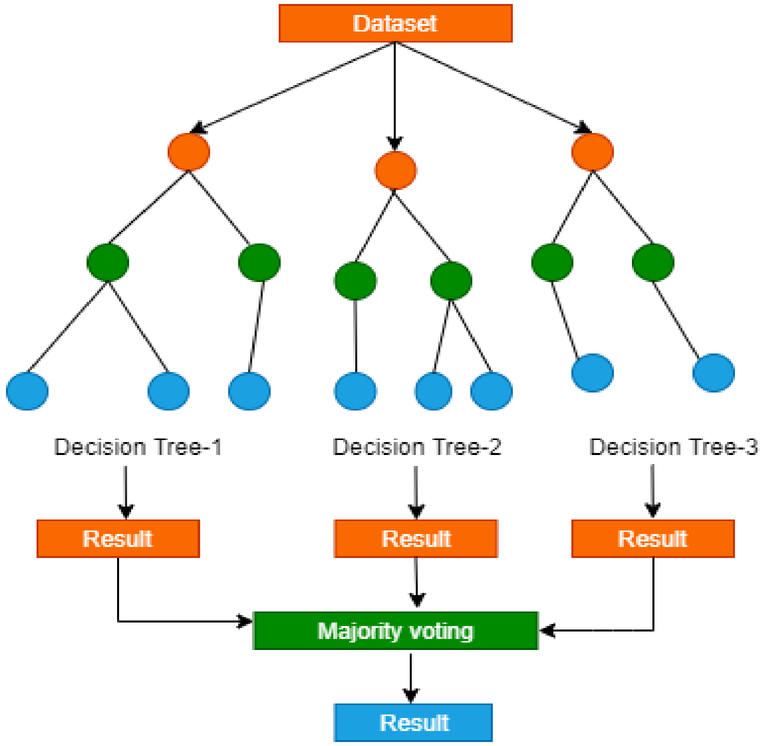


Fig 4.1: Random Forest

The Random Forest can be implemented in a diverse sector to efficiently predict the outcome. The name of the sectors is given below:

- ❑ **Banking Sector:** To forecast the hazard of the loan, Random Forest is very workable.
- ❑ **Medicine Sector:** The risk of diseases like Heart disease and Diabetes malady can be reliably predicted by the assistance of the Random Forest algorithm.
- ❑ **Marketing Sector:** To unroll the tendency of the marketplace Random Forest can be implemented.

4.2.2 Decision Tree

A decision tree is an algorithm that is used for the prediction. It is considered a very efficient tool for classification problems. The technique which is called dividing criteria is used by the decision tree for the classification problem. In the Decision tree the attributes of the data set are illustrated by the assist of internal nodes.

Three nodes are available in the decision tree these are:

1. Root node
2. Decision Node
3. Leaf node

The terminology of the Decision Tree is given below:

- ❑ **Root Node:** It is the starting node of the decision tree. The whole dataset is represented by this node.
- ❑ **Decision Node:** It is a kind of node of the decision tree where the sub-nodes are partitioned into other sub-nodes
- ❑ **Pruning:** The process which is implemented to remove the baggage node of the decision node is called pruning.
- ❑ **Sub tree:** In the decision tree, this tree is produced by anticipating the tree partition.
- ❑ **Leaf Node:** This node is the final node of the decision tree that represents the outcome.

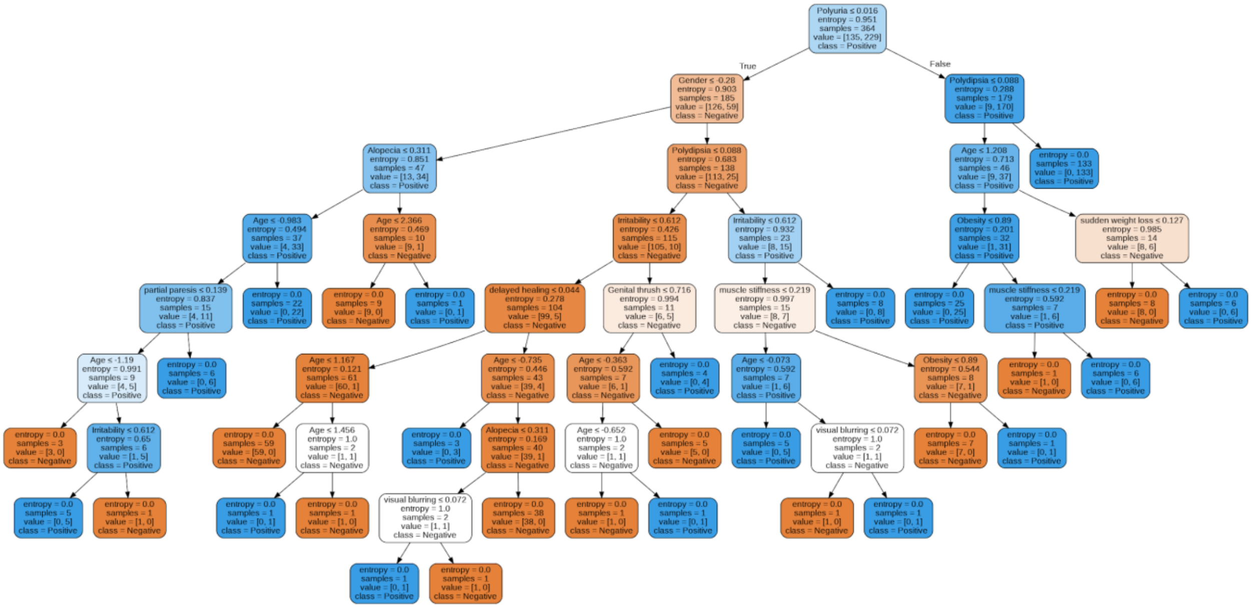

Polyuria ≤ 0.016 entropy = 0.951 samples = 364 value = [135, 229] class = Positive
True
False
Gender ≤ -0.28 entropy = 0.903 samples = 185 value = [126, 59] class = Negative
Polydipsia ≤ 0.088 entropy = 0.288 samples = 179 value = [9, 170] class = Positive
Alopecia ≤ 0.311 entropy = 0.851 samples = 47 value = [13, 34] class = Positive
Polydipsia ≤ 0.088 entropy = 0.683 samples = 138 value = [113, 25] class = Negative
Age ≤ 1.208 entropy = 0.713 samples = 46 value = [9, 37] class = Positive
entropy = 0.0 samples = 133 value = [0, 133] class = Positive
Age ≤ -0.983 entropy = 0.494 samples = 37 value = [4, 33] class = Positive
Age ≤ 2.366 entropy = 0.469 samples = 10 value = [9, 1] class = Negative
Irritability ≤ 0.612 entropy = 0.426 samples = 115 value = [105, 10] class = Negative
Irritability ≤ 0.612 entropy = 0.932 samples = 23 value = [8, 15] class = Positive
Obesity ≤ 0.89 entropy = 0.201 samples = 32 value = [1, 31] class = Positive
sudden weight loss ≤ 0.127 entropy = 0.985 samples = 14 value = [8, 6] class = Negative
partial paresis ≤ 0.139 entropy = 0.837 samples = 15 value = [4, 11] class = Positive
entropy = 0.0 samples = 22 value = [0, 22] class = Positive
entropy = 0.0 samples = 9 value = [9, 0] class = Negative
entropy = 0.0 samples = 1 value = [0, 1] class = Positive
delayed healing ≤ 0.044 entropy = 0.278 samples = 104 value = [99, 5] class = Negative
Genital thrush ≤ 0.716 entropy = 0.994 samples = 11 value = [6, 5] class = Negative
muscle stiffness ≤ 0.219 entropy = 0.997 samples = 15 value = [8, 7] class = Negative
entropy = 0.0 samples = 8 value = [0, 8] class = Positive
entropy = 0.0 samples = 25 value = [0, 25] class = Positive
muscle stiffness ≤ 0.219 entropy = 0.592 samples = 7 value = [1, 6] class = Positive
entropy = 0.0 samples = 8 value = [8, 0] class = Negative
entropy = 0.0 samples = 6 value = [0, 6] class = Positive
Age ≤ -1.19 entropy = 0.991 samples = 9 value = [4, 5] class = Positive
entropy = 0.0 samples = 6 value = [0, 6] class = Positive
Age ≤ 1.167 entropy = 0.121 samples = 61 value = [60, 1] class = Negative
Age ≤ -0.735 entropy = 0.446 samples = 43 value = [39, 4] class = Negative
Age ≤ -0.363 entropy = 0.592 samples = 7 value = [6, 1] class = Negative
entropy = 0.0 samples = 4 value = [0, 4] class = Positive
Age ≤ -0.073 entropy = 0.592 samples = 7 value = [1, 6] class = Positive
Obesity ≤ 0.89 entropy = 0.544 samples = 8 value = [7, 1] class = Negative
entropy = 0.0 samples = 1 value = [1, 0] class = Negative
entropy = 0.0 samples = 6 value = [0, 6] class = Positive
entropy = 0.0 samples = 3 value = [3, 0] class = Negative
Irritability ≤ 0.612 entropy = 0.65 samples = 6 value = [1, 5] class = Positive
entropy = 0.0 samples = 59 value = [59, 0] class = Negative
Age ≤ 1.456 entropy = 1.0 samples = 2 value = [1, 1] class = Negative
entropy = 0.0 samples = 3 value = [0, 3] class = Positive
Alopecia ≤ 0.311 entropy = 0.169 samples = 40 value = [39, 1] class = Negative
Age ≤ -0.652 entropy = 1.0 samples = 2 value = [1, 1] class = Negative
entropy = 0.0 samples = 5 value = [5, 0] class = Negative
entropy = 0.0 samples = 5 value = [0, 5] class = Positive
visual blurring ≤ 0.072 entropy = 1.0 samples = 2 value = [1, 1] class = Negative
entropy = 0.0 samples = 7 value = [7, 0] class = Negative
entropy = 0.0 samples = 1 value = [0, 1] class = Positive
entropy = 0.0 samples = 5 value = [0, 5] class = Positive
entropy = 0.0 samples = 1 value = [1, 0] class = Negative
entropy = 0.0 samples = 1 value = [0, 1] class = Positive
entropy = 0.0 samples = 1 value = [1, 0] class = Negative
visual blurring ≤ 0.072 entropy = 1.0 samples = 2 value = [1, 1] class = Negative
entropy = 0.0 samples = 38 value = [38, 0] class = Negative
entropy = 0.0 samples = 1 value = [1, 0] class = Negative
entropy = 0.0 samples = 1 value = [0, 1] class = Positive
entropy = 0.0 samples = 1 value = [1, 0] class = Negative
entropy = 0.0 samples = 1 value = [0, 1] class = Positive
entropy = 0.0 samples = 1 value = [0, 1] class = Positive
entropy = 0.0 samples = 1 value = [1, 0] class = Negative

Fig 4.2: Decision Tree

Some phases are followed in the Decision tree algorithm to continue the work process. Those phases are given below:

- **Phase-1:** The tree is started from the root node. The entire dataset is hold by the starting or root node.
- **Phase-2:** In the second phase, supreme attributes are being traced.
- **Phase-3:** In this phase, the subsets which contain the probable values of the supreme attribute, those subsets are partitioned from the root node.
- **Phase-4:** In this phase, the node that stores the supreme attribute is produced.
- **Phase-5:** The subsets of the third phase are used in this phase to construct the new decision trees repeatedly. This procedure is carried on till the state from where the leaf or final node arrives.

A decision tree is mainly applied to construct a training model which can be performed for the prediction of the outcome.

The fact which are the motive to use the decision trees are given below:

- ❑ During the decision making time the human ideas are copied by the decision trees. So the Decision tree is deliberated as a clear understanding model.
- ❑ The Decision tree is structured like a tree, that's why it is very comfortable to understand.

4.2.3 Extra Trees

Extremely randomized trees can be called Extra trees. This Extra Trees algorithm can effectively operate with many decision trees like the Random Forest algorithm. It assembles the forecast of each decision tree. Often, the Extra Trees algorithm can beat the Random Forest in terms of performance. Within the Extra Trees, different decision trees are manufactured from the pattern of training. In this algorithm, random node splitting occurs.

Extra Trees algorithm follows several rules those are given below:

- ❑ Numerous trees are constructed in this algorithm along with Bootstrap =Not True. This strongly indicates that Extra trees always sample by avoiding the replacement.
- ❑ It does not prefer the best split to splitting the tree nodes. However, it uses the random split to split the nodes of the trees.

The variance is considered as low in the Extra trees algorithm, that's why it is more applicable for the prediction in a static environment like traffic flow regression. The boosting ensemble method is mainly implemented by this Extra trees algorithm for the improvement of the accuracy and also for controlling the overfitting complexity.

Extra trees algorithm has some advantages like:

- ❑ Firstly, it can provide better accuracy than Random forest algorithms often.
- ❑ Secondly, it is quicker than the Random forest algorithm.
- ❑ Thirdly, in the appearance of the noisy feature, it can provide better performance.
- ❑ Lastly, It can control the overfitting problem.

4.2.4 XGBoost

The Extreme Gradient Boosting in a nutshell it can be called as a XGboost algorithm. Now days, it is very popular due to its performance and better speed. This algorithm is exploited to evolve classification problems. It can easily handle the missing values. It uses the Regularization technique to be rescued from the overfitting problem. By the inbuilt CV function, the XGboost algorithm becomes enabled. The XGboost algorithm has some targets to fill up.

The targets of the XGboost algorithm are:

- ❑ **Seizing the data set:** The XGboost algorithm efficiently regulates the data set which is structured.
- ❑ **Fastest speed**: The XGboost algorithm is much quicker than the other algorithms.

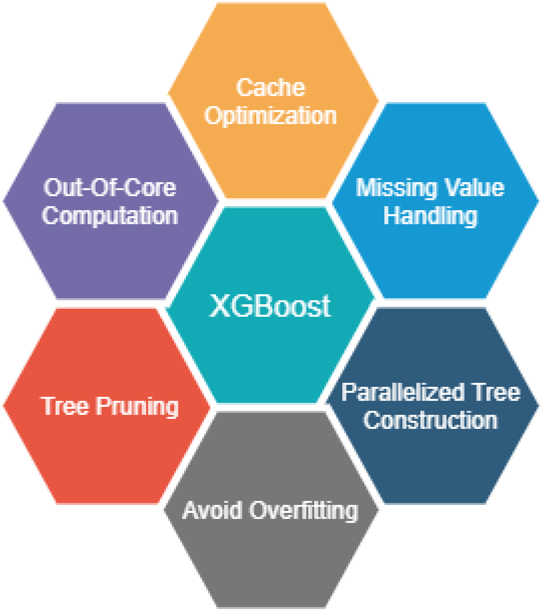


Fig 4.3: XGBoost Optimization

The system prominence of the XGboost algorithm is in the following:

Parallelization: In the training period, through using the parallelized impersonation the XGboost algorithm already enters the method of tree constructing which is sequential.

Pruning the tree: The XGboost algorithm implemented the max_depth by avoiding the criterion first and the trimming of the trees starts from the behind point.

Feasible use of hardware: The XGboost algorithm always ensures that the hardware components are used appropriately. This is performed by the cache optimization. During the time of managing the large data set which is unable to fit along with the memory. The out-of-core computation optimizes the space of the disk at that time.

Several facts that make the XGboost algorithm different from other algorithms such as:

1. XGboost algorithm can operate with diverse languages like Python, Java ,and C++.
2. It can be implemented for the multiplex predictions problem like user-identified and classification.
3. It can drive in Linux and Windows.
4. It works along with many ecosystems efficiently.

4.2.5 Bagging

Bagging is also called the Bootstrap Aggregation. Bagging is an ensemble method that works very effectively to convert the variance from a high to a low level. This is very efficient to prevent overfitting problems. In the bagging algorithm, the forecasts are prepared by using the average outcome of the trained models.

Some phases are followed in the Bagging algorithm to continue the work process. Those phases are in the following:

Phase-1: In the first phase, a pattern from the observation is elected by implementing random selection.

Phase-2: In the second phase, a model construction along with the feature subset and also with the pattern of the observation is started after choosing the subset of the features.

Phase-3: In the third phase, the feature of the subset which can efficiently deliver the best split on the data of the training such feature will be chosen.

Phase-4: Previous phases will occur repeatedly to construct the models. In this fourth phase, the models are trained parallelly.

Phase-5: In this phase, predictions of models are collected. Then aggregation of predictions which are from the models is made to get the final prediction.

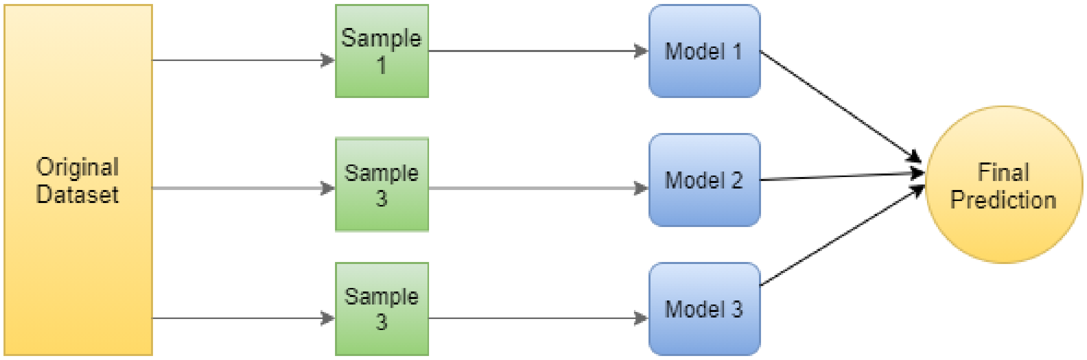


Fig 4.4: Bagging algorithm

In recent times, the Bagging algorithm is being extensively used in many sectors like the medical sector, fraud detecting sector, banking field, and network intrusion detection system. The bagging algorithm has some advantages such as in the high dimensional data it can provide the better performance, and the execution of this algorithm is not influenced by the missing values which are in the dataset.

4.3 Implementation Procedure

In this research work, diverse data mining algorithms have been used such as Random Forest, Decision Tree, Extra Trees, XGboost, and Bagging. The dataset of this work had been collected from a publicly accessible data gathering website. There was no inaccurate record in the dataset, so there was no necessity to clean the dataset. Afterward, checking the cleanliness of the dataset several libraries are imported into the system to import the dataset. Then the dataset was imported into the system. This dataset does not contain any missing values so the data imputation technique was skipped in our research work.

After performing the data preprocessing, the dataset was split by using the Percentage split and the K-fold cross-validation. In Percentage Split, 30% of the data are conducted for the testing and 70% of the data are conducted for the training purpose. In k-fold Cross-Validation 9 fold are used to split the dataset.

The workflow diagram of our whole process is being displayed in the following figure (fig: 4.15.)

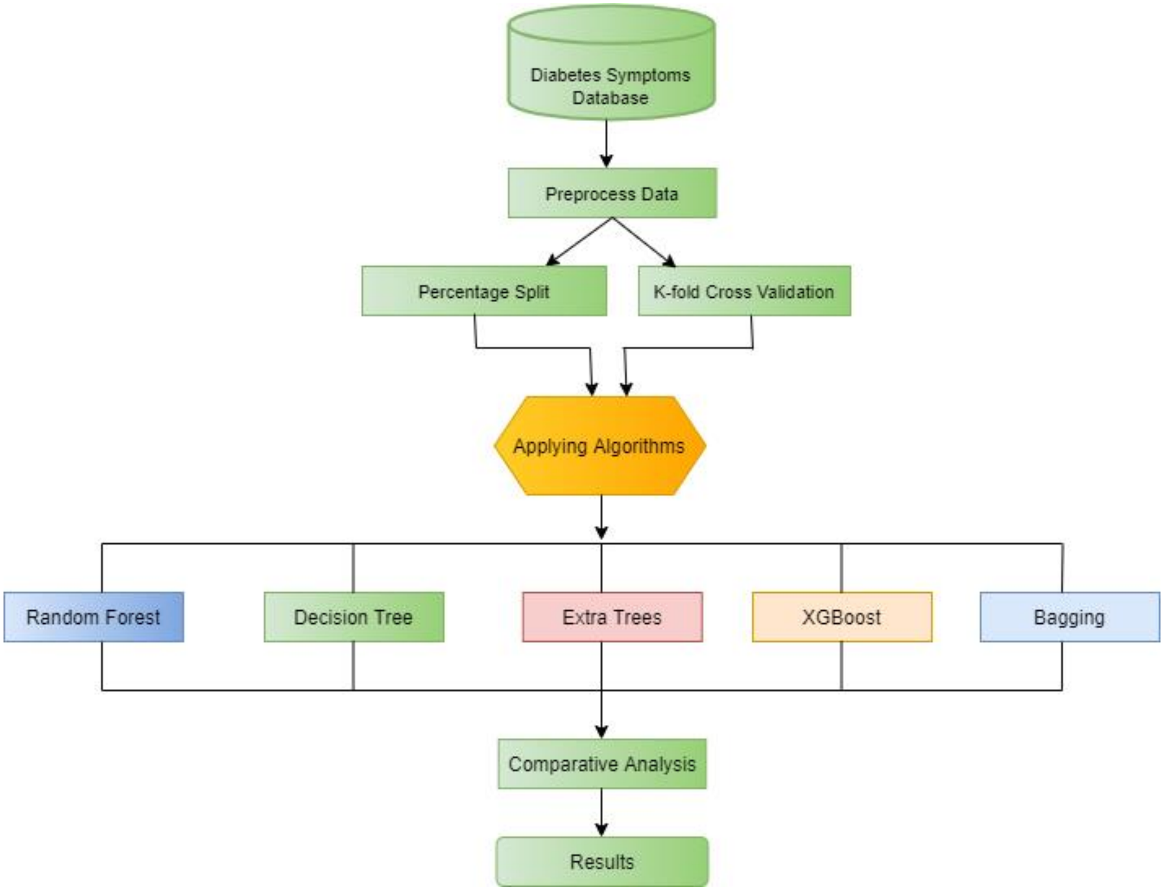


Fig 4.5: Workflow diagram

CHAPTER 5

ANALYSIS OF RESULTS

5.1 Introduction

In the preceding section details regarding the diverse algorithms like Random Forest, Decision Trees, Extra Trees, XGboost, and Bagging are explained. Now in this chapter, the results which are obtained from multiple algorithms will be discussed. In the chapter, there will be an analysis of the results of the algorithms, and the best accuracy-providing algorithm will be explored.

5.2 Result Analysis

In this research work, various data mining algorithms are implemented such as Random Forest, Decision Trees, Extra Trees, XGboost, and Bagging. The algorithms are compared to each other by accomplishing the calculation of the model accuracy, sensitivity, specificity, precision, F1-Score, and recall score.

5.3 Expository Analysis

The confusion matrix is gained to perform an assessment that evaluates the performance of the algorithms. The confusion matrix is used to display the number of instances that are assigned in every class. The dataset of this work contains two classes therefore the 2x2 confusion matrix is available for this experimental work.

The structure of the confusion matrix is in the following:

- ❑ A=Positive(Has Diabetes Disease)
- ❑ B=Negative(Hasn’t Diabetes Disease)

Table 5.1: A Primary Confusion Matrix

	A	**B**
A	**TP**	**FN**
B	**FP**	**TN**

Here:-

- ❑ **True Positive:** In a nutshell, it can be called TP. The TP indicates the digit of the records which are categorized as positive and they were also positive in real.
- ❑ **False Negative:** In a nutshell, it can be called FN. The FN indicates the digit of the records which are categorized as negative, but they were positive in real.
- ❑ **False Positive:** In a nutshell, it can be called FP. The FP indicates the digit of the records which are categorized as positive, but they were negative in real.
- ❑ **True Negative:** In a nutshell, it can be called TN. The TN indicates the digit of the records which are categorized as negative and they were also negative in real.

The Accuracy, Sensitivity, Specificity, Precision, F1 score, and Roc Curve those evaluation methods are implemented to evaluate the models for this research work.

The sensitivity is the presentation of the true positive rate. It is performed to discover the number of actual positives which were accurately diagnosed. It is also called the Recall.

$$\text{Sensitivity}= \frac{True\ \ Positive}{True\ Positive+False\ Negative} \times 100\%$$

The evaluation method which is implemented to find out the real negatives which were accurately diagnosed is called specificity. It is the representation of the true negative rate.

$$\text{Specificity}= \frac{True\ \ Negative}{True\ Negative+False\ Positive} \times 100\%$$

The precision is used to efficiently discover the true positive and the forecasted positive values. The precision is the path to calculate the rightness of the different models.

$$\text{Precision}= \frac{True\ Positive}{True\ Positive+False\ Positive} \times 100\%$$

The evaluation method used to combine the precision and the recall of the models is known as the F1 score.

$$\text{F1 Score}=\frac{2\times Precision\times Recall}{Precision+Recall} \times 100\%$$

5.4 Experimental Assessment

In our research work, five data mining algorithms are used to obtain the accuracy of each algorithm and to commit a comparison among the algorithms. Two techniques like Percentage split and 9-fold cross-validation are applied to assess the effectiveness of the algorithms. While using the percentage split technique, the Random Forest provided the highest accuracy score with 99.36% accuracy, the Extra Trees algorithm obtained the second maximal exactness percentage with 98.72% accuracy, the Bagging algorithm gained the third-maximal correctness with 98.08% accuracy, and the XGboost algorithm and the Decision Tree algorithm provided the lowest accuracy score with 97.44% accuracy

Where using the percentage split technique, the Extra trees algorithm provided the maximal exactness with 98.28% accuracy, the Random Forest algorithm provided the second-highest accuracy score with 98.08% accuracy, the Bagging algorithm obtained the third-maximal correctness with 96.74% accuracy, the XGboost algorithm achieved the fourth-maximal exactness which is 96.35% accuracy, and the Decision Tree algorithm provided the lowest accuracy algorithm provided the lowest accuracy score with 96.17% accuracy. .

The Confusion matrix has been used in this research work to assess the different models. The entire performance of the models can be demonstrated through an evaluation matrix which is recognized as the Confusion matrix. The confusion matrix of the diverse algorithms that are applied in this work is in the following tables.

Confusion matrix of the algorithms on Percentage Split technique:

Table 5.2: Random Forest Confusion Matrix

	A	**B**
A	**91**	**0**
B	**1**	**64**

Table 5.3: Decision Tree Confusion Matrix

	A	**B**
A	**89**	**2**
B	**2**	**63**

Table 5.4: Extra Trees Confusion Matrix

	A	**B**
A	**90**	**1**
B	**1**	**64**

Table 5.5: XGBoost Confusion Matrix

	A	**B**
A	**88**	**3**
B	**1**	**64**

Table 5.6: Bagging Confusion Matrix

	A	**B**
A	**88**	**3**
B	**0**	**65**

Confusion matrix of the algorithms on K-fold Cross-Validation technique:

Table 5.7: Random Forest Confusion Matrix

	A	**B**
A	**315**	**5**
B	**5**	**195**

Table 5.8: Decision Tree Confusion Matrix

	A	**B**
A	**308**	**12**
B	**8**	**192**

Table 5.9: Extra Trees Confusion Matrix

	A	**B**
A	**316**	**4**
B	**5**	**195**

Table 5.10: XGBoost Confusion Matrix

	A	**B**
A	**308**	**12**
B	**7**	**193**

Table 5.11: Bagging Confusion Matrix

	A	**B**
A	**311**	**9**
B	**8**	**192**

The evaluation metrics have been implemented in this research work to assess the effectiveness of the algorithms efficiently. The evaluation method is very necessary to review the performance of the various models. It provides the evaluation numbers. By analyzing those evaluation numbers the performance of the models can be improved.

Table 5.12: Performance evaluation score on Percentage Split

Algorithms	Sensitivity	Specificity	Precision	F1 Score
Random Forest	100%	98.46%	98.91%	99.45%
Decision Tree	97.80%	96.92%	97.80%	97.80%
Extra Trees	98.90%	98.46%	98.90%	98.90%
XGBoost	96.70%	98.46%	98.87%	97.77%
Bagging	96.70%	100%	100%	98.32%

Table 5.13: Accuracy of different data mining algorithms on Percentage Split

Algorithms	Accuracy
Random Forest	99.36%
Decision Tree	97.44%
Extra Trees	98.72%
XGBoost	97.44%
Bagging	98.08%

The horizontal bar graph has been used to display the accuracy of diverse data mining algorithms .The accuracy of different data mining algorithms is shown through the below graph:

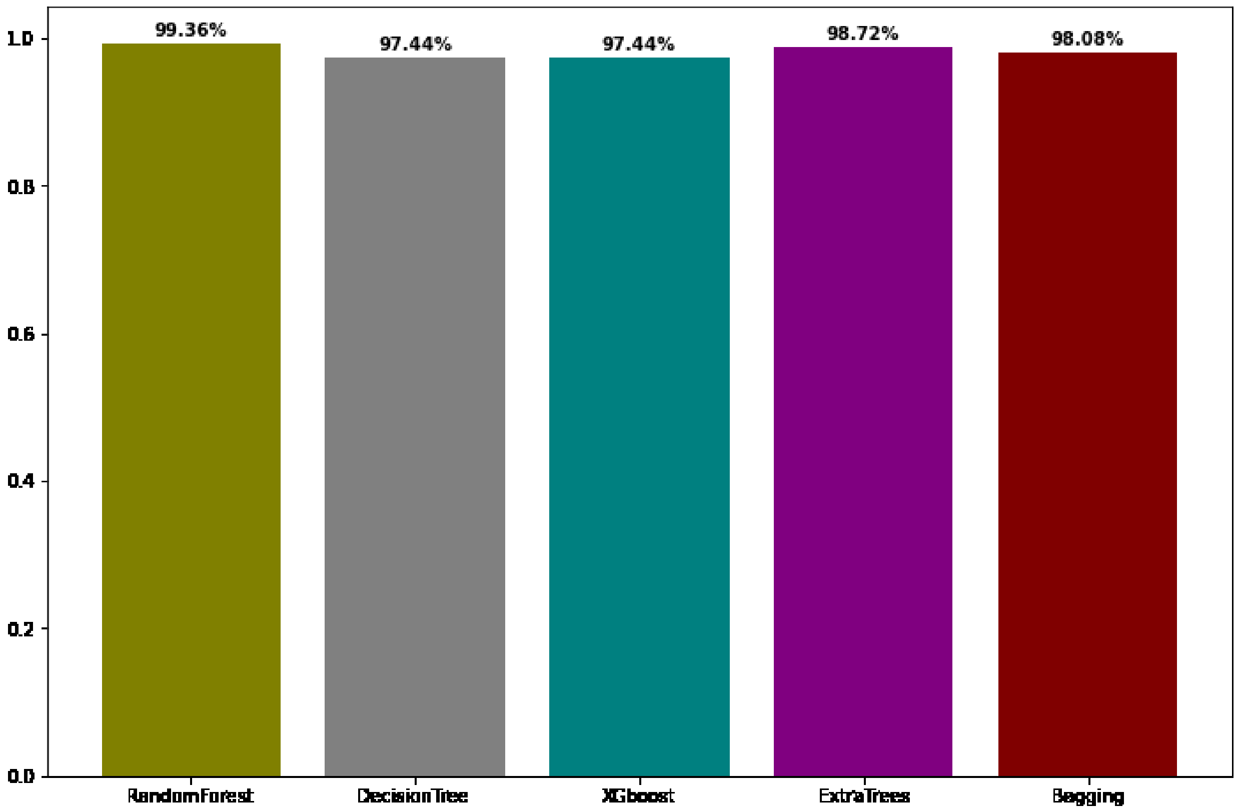


Fig 5.1: Performance of algorithms using Percentage Split

The algorithms are ranked in the above graph based on their accuracy. By analyzing the graph it is evident that Random Forest is capable of providing the best accuracy amongst all the algorithms. The Random Forest obtained the 99.36% accuracy whole using the percentage split technique.

Table 5.14: Performance evaluation score on K-fold Cross-Validation

Algorithms	Sensitivity	Specificity	Precision	F1 Score
Random Forest	98.43%	97.5%	98.43%	98.43%
Decision Tree	96.25%	96%	97.46%	96.85%
Extra Trees	98.75%	97.5%	98.44%	98.59%
XGBoost	96.25%	96.5%	97.77%	97%
Bagging	97.18%	96%	97.49%	97.33%

Table 5.15: Accuracy of different data mining algorithms on K-fold Cross-Validation

Algorithms	Accuracy
Random Forest	98.08%
Decision Tree	96.17%
Extra Trees	98.28%
XGBoost	96.35%
Bagging	96.74%

The accuracy of different data mining algorithms while using K-fold Cross-validation is shown through the below graph:

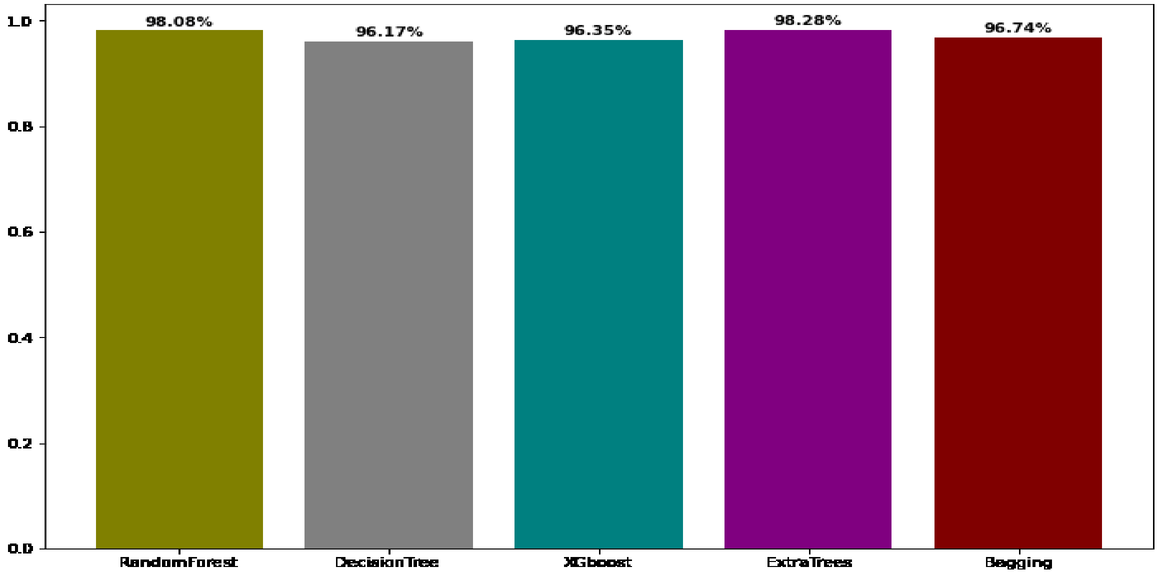


Fig 5.2: Performance of algorithms using K-fold Cross-Validation

In the above graph, the algorithms are ranked by their accuracy. By observing the graph it can be assured that the Extra Trees algorithm is proficient in providing the highest accuracy among all the algorithms. The Extra trees achieved 98.28% accuracy while using the K-fold Cross-Validation.

AUC-ROC Curve: The curve is used to operate a measurement for the classification problem that is called the AUC-ROC curve. The probability curve is efficiently represented by the ROC (Receiver Operator Characteristic) and AUC (Area Under the Curve) represents the antithesis of the classes.

AUC-ROC Curve of the algorithms on Percentage Split Technique:

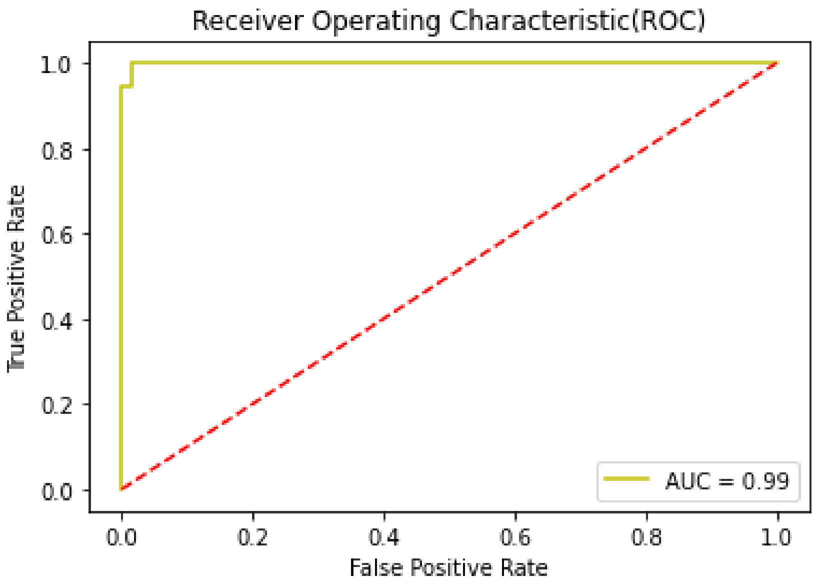


Fig 5.3: AUC-ROC for Random Forest

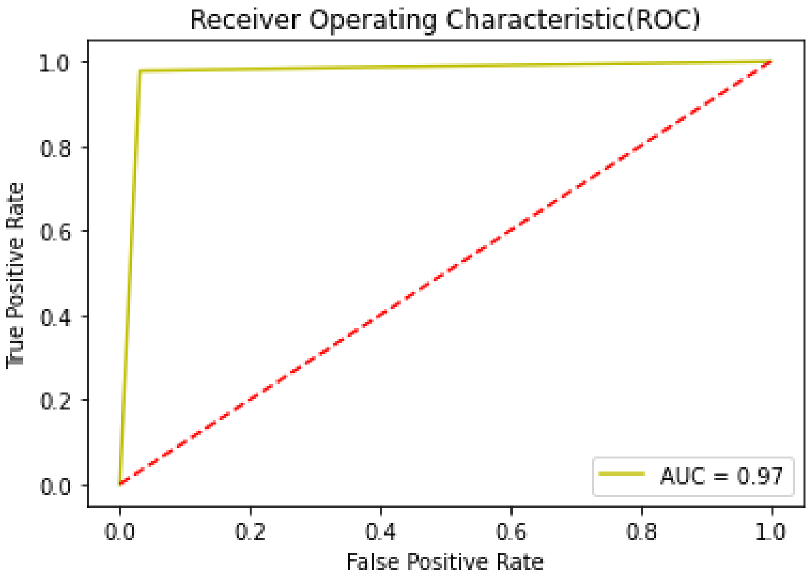


Fig 5.4: AUC-ROC for Decision Tree

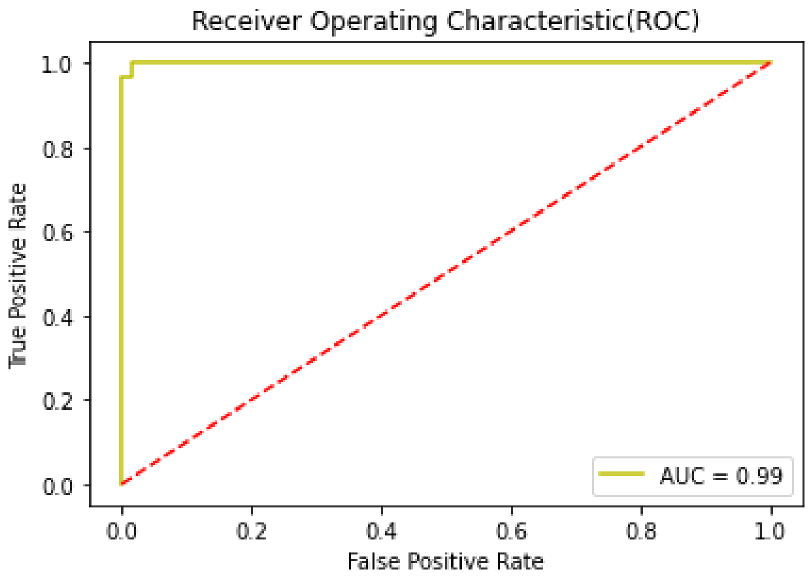


Fig 5.5: AUC-ROC for Extra Trees

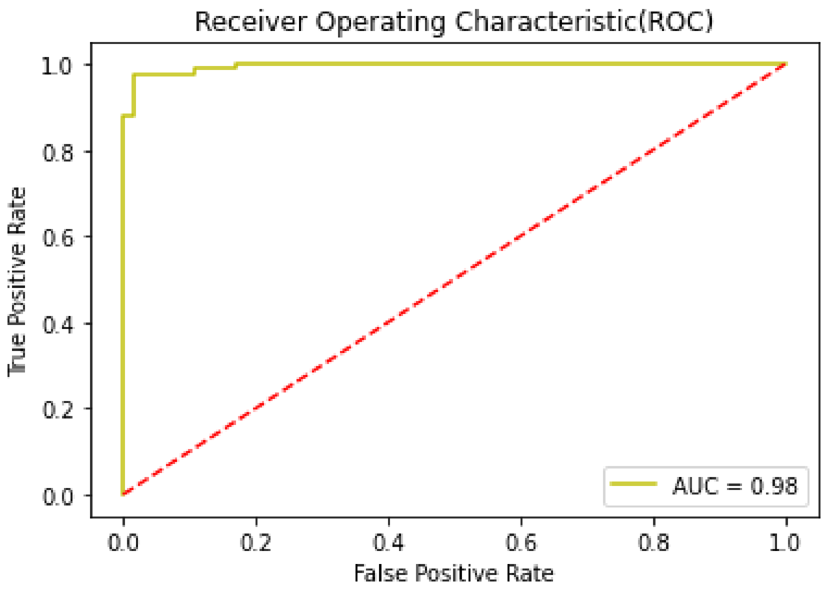


Fig 5.6: AUC-ROC for XGBoost

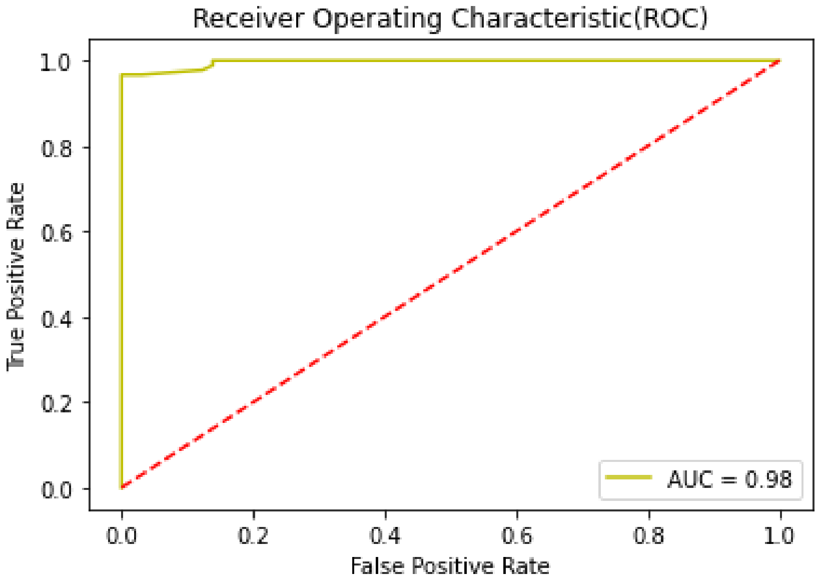


Fig 5.7: AUC-ROC for Bagging

AUC-ROC Curve of the algorithms on K-fold Cross-Validation Technique:

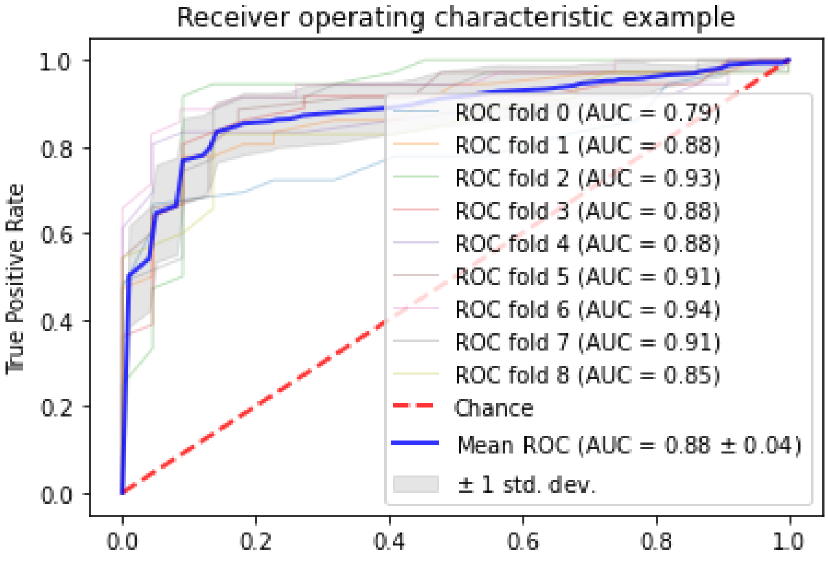


Fig 5.8: AUC-ROC for Random Forest

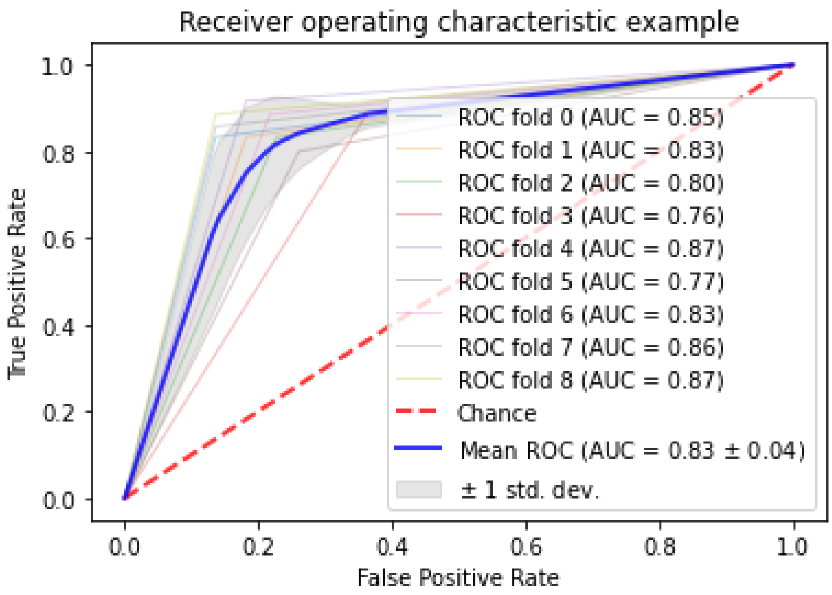


Fig 5.9: AUC-ROC for Decision Tree

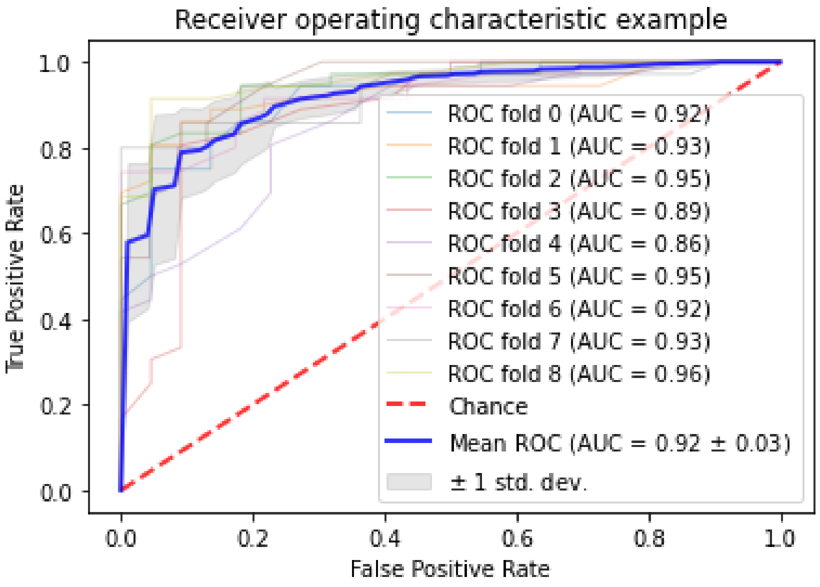


Fig 5.10: AUC-ROC for Extra Trees

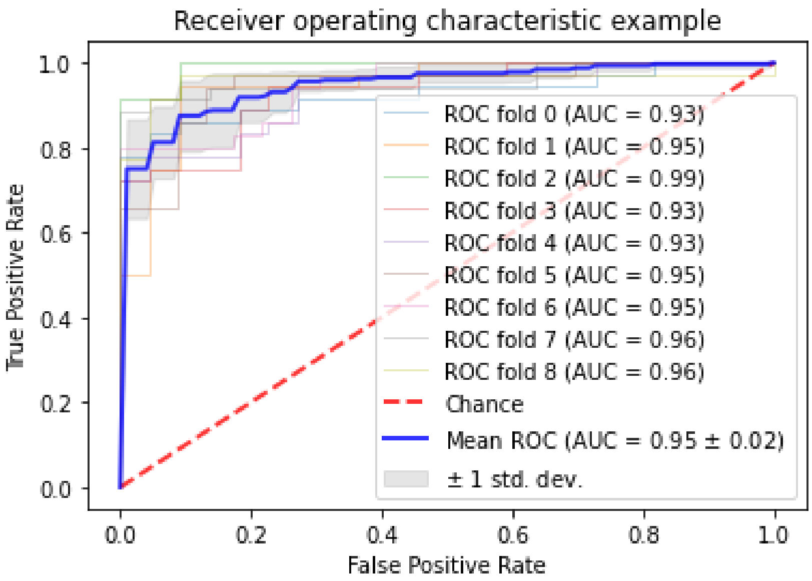


Fig 5.11: AUC-ROC for XGBoost

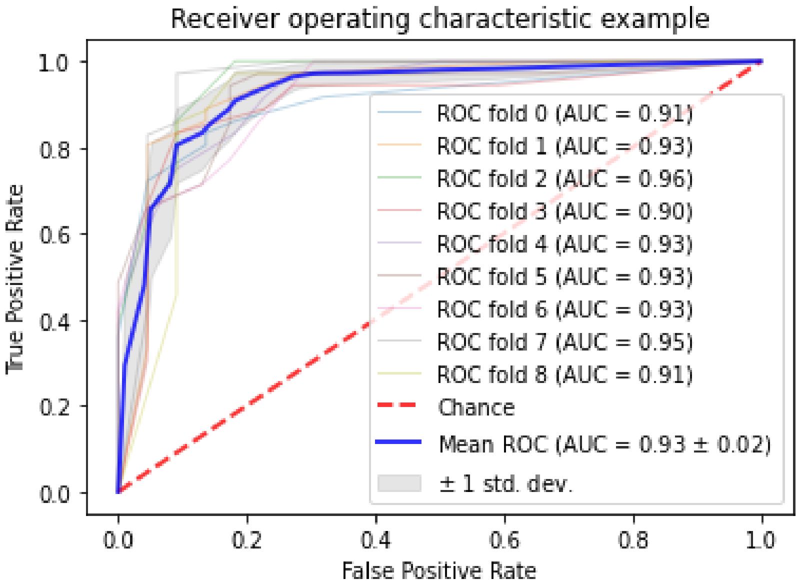


Fig 5.12: AUC-ROC for Bagging

5.5 Comparison of Accuracy with the Existing work

Table 5.16: Comparison of accuracy using Percentage Split technique

Work	Random Forest	Decision Tree	Extra Trees	XGBoost	Bagging
(Islam et al., 2019)	99.4%	95%	-	-	-
(Malik et al., 2020)	98.8%	93.88%	-	-	93.47%
(Mujumdar & Vaidehi, 2019)	91%	86%	91%	-	90%
(Soni & Varma, 2020)	77%	75%	-	-	-
(Shuaibu et al., 2021)	-	-	-	77%	-
(Khanam & Foo,2021)	77.14%	73.14%	-	-	-
This work	**99.36%**	**97.44%**	**98.72%**	**97.44%**	**98.08%**

Through analyzing the above table it is obvious that these proposed work algorithms like Random Forest, Decision Tree, Extra Trees, XGBoost, and bagging achieved better accuracy of 99.35%,97.43%,98.71%,97.43%, and 98.07% respectively using Percentage split than the other works algorithms shown in the table.

Table 5.17: Comparison of accuracy using K-fold Cross-Validation

Work	Random Forest	Decision Tree	Extra Trees	XGBoost	Bagging
(Islam et al., 2019)	97.4%	95.6%	-	-	-
(Alphan & Ilgi,2020)	97.5%	95.96%	96.15%	-	-
(Chaves & Marques, 2021)	96.92%	-	-	-	-
(Peker et al.,2018)	91.34%	93.03%	-	-	-
(Zou et al., 2018)	80.84%	78.53%	-	-	-
(Bhulakshmi & Gandhi,2020)	-	-	-	81%	-
(Khanam & Foo,2021)	74.96%	74.24%	-	-	-
(Shuja et al.,2020)	-	92.50%	-	-	94.14%
This work	**98.08%**	**96.17%**	**98.28%**	**96.35%**	**96.74%**

By analyzing the above table it is noticeable that these proposed work algorithms like Random Forest, Decision Tree, Extra Trees, XGBoost, and bagging achieved better accuracy of 98.08%,96.16%,98.27%,96.35%, and 96.73% respectively using K-fold Cross-Validation than the other works algorithms shown in the table.

5.6. Discussion

In this chapter, the performance of the various algorithms is discussed. Different evaluation metrics are implemented to efficiently evaluate the effectiveness of the models. The comparison of the model's effectiveness of this work with the existing work is also shown in this chapter.

CHAPTER 6
CONCLUSON AND FUTURE WORK

6.1 Conclusion

Numerous peoples are becoming affected by chronic diabetes disease frequently. The peoples can reduce the risk of developing diabetes disease by maintaining a healthful lifestyle that includes following a healthy diet plan, doing exercise regularly, and avoiding smoking. The fast detection of diabetes disease can provide the peoples opportunity to start their treatment as early as possible. Data mining performs a very significant role in diagnosing undiagnosed diseases. It also can decrease the cost to predict the diabetes malady. By analyzing the effectiveness of data mining in the medical sector, this research work involves data mining algorithms to predict the diabetes malady. The main motto of this work is to discover the best algorithms by analyzing the performance of the data mining algorithms to predict the diabetes malady. It is observed that the Random Forest algorithm is the most effective while using the percentage splitting technique.

6.2 Future Work & Scope of this research work

In the future, the most efficient model of this work will be applied to build a diabetics prediction application for the desktop and the mobile. The application will be very much user-friendly. This application will be suitable for both apple and android users. People will quickly get the diabetes prediction result by using this application. The models of this work will also be improved by assembling more data to implement on the various algorithms.

REFERENCES

[1] Watson, S. (2020, February 27). *Diabetes: Symptoms, Causes, Treatment, Prevention, and More*. Healthline. https://www.healthline.com/health/diabetes.

[2] Klöppel, G., Löhr, M., Habich, K., Oberholzer, M., & Heitz, P. U. (1985). Islet Pathology and the Pathogenesis of Type 1 and Type 2 Diabetes mellitus Revisited. *Pathology and Immunopathology Research*, *4*(2), 110–125.

[3] World Health Organization. (2021, April 13). *Diabetes*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/diabetes.

[4] Elflein, J. (2021, February 18). *Topic: Diabetes*. Statista. https://www.statista.com/topics/1723/diabetes/.

[5] Diabetes Australia. (2018, July 8). *Failure to detect type 2 diabetes early costing $700 million per year*. https://www.diabetesaustralia.com.au/mediarelease/failure-to-detect-type-2-diabetes-early-costing-700-million-per-year/.

[6] Ramachandran, A., Snehalatha, C., Shetty, A. S., & Nanditha, A. (2012). Trends in prevalence of diabetes in Asian countries. *World Journal of Diabetes*, *3*(6), 110–117.

[7] Ramachandran , A. (2014). Know the signs and symptoms of diabetes. *The Indian Journal of Medical Research*, *140*(5), 579–581.

[8] Amalarethinam, D.G., & Vignesh, N. A. (2015). Prediction of Diabetes mellitus using Data Mining Techniques: A Survey. *International Journal of Applied Engineering Research*, *10*(82), 24–31.

[9] *What is data mining?* SAS. (n.d.). https://www.sas.com/en_us/insights/analytics/data-mining.html.

[10] *Diabetes - long-term effects*. Better Health Channel. (n.d.). https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects.

[11] Diabetes Mellitus, Fasting Glucose, and Risk of Cause-Specific Death. (2011). *New England Journal of Medicine, 364*(9), 829–841.

[12] LeRoith, D., Biessels, G. J., Braithwaite, S. S., Casanueva, F. F., Draznin, B., Halter, J. B., Hirsch, I. B., McDonnell, M. E., Molitch, M. E., Murad, M. H., & Sinclair, A. J. (2019). Treatment of Diabetes in Older Adults: An Endocrine Society Clinical Practice Guideline. *The Journal of Clinical Endocrinology & Metabolism*, *104*(5), 1520–1574.

[13] Zhuo, X., Zhang, P., Barker, L., Albright, A., Thompson, T. J., & Gregg, E. (2014). The Lifetime Cost of Diabetes and Its Implications for Diabetes Prevention. *Diabetes Care, 37*(9), 2557–2564.

[14] Mahboob Alam, T., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Imtiaz Baig, T., Hussain, A., Malik, M. A., Raza, M. M., Ibrar, S., & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, *16*, 100204.

[15] Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2019). Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques. *Computer Vision and Machine Intelligence in Medical Image Analysis*, 113–125.

[16] Alpan, K., & Ilgi, G. S. (2020). Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach. *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).*

[17] Chaves, L., & Marques, G. (2021). Data Mining Techniques for Early Diagnosis of Diabetes: A Comparative Study. *Applied Sciences*, *11(5),* 2218.

[18] Iyer, A., S, J., & Sumbaly, R. (2015). Diagnosis of Diabetes Using Classification Mining Techniques. *International Journal of Data Mining & Knowledge Management Process, 5(1),* 01–14.

[19] Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*, 1578–1585.

[20] Silva, L. H. S.D, Pathirage, N., & Jinasena, T. M. K. K. (2016). Diabetic Prediction System Using Data Mining. *International Research Conference-KDU,* 66–72.

[21] Peker, M., Özkaraca, O., & Şaşar, A. (2018). Use of Orange Data Mining Toolbox for Data Analysis in Clinical Decision Making. *Expert System Techniques in Biomedical Science Practice,*143-167.

[22] Malik, S., Harous, S., & El-Sayed, H. (2020). Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women. *Modelling and Implementation of Complex Systems*, 95–106

[23] Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292–299.

[24] Soni , M., & Varma, D. S. (2020). Diabetes Prediction using Machine Learning Techniques.*International Journal of Engineering Research & Technology (IJERT), 9(9),* 921–925.

[25] Pradhan, N., Rani, G., Dhaka, V. S., & Poonia, R. C. (2020). Diabetes prediction using artificial neural network. *Deep Learning Techniques for Biomedical and Health Informatics*, 327–339.

[26] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, *9*, 1–10.

[27] Shuaibu, A. R., Muhammad, N. A., & Nwojo, N. A. (2021). Comparative Analysis of Diabetes Mellitus Classification Using Support Vector Machine and Extreme Gradient Boosting Algorithms . *International Conference on Electrical Engineering Applications (ICEEA'2020)*, 65–71.

[28] Bhulakshmi, D., & Gandhi, G. (2020). The Prediction of Diabetes in Pima Indian women Mellitus Based on XGBOOST Ensemble Modeling using data science. *EasyChair*.

[29] Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*.

[30] Shuja, M., Mittal, S., & Zaman, M. (2020). Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE. *Advances in Computing and Intelligent Systems*, 195–211.

APPENDIX A