Bangladesh Army International University of Science & Technology (BAIUST)

Cumilla Cantonment, Cumilla

This thesis is submitted in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering.

MD. Ehsanul Haque

ID:1105004

Marzia khan Turna

ID:1105014

Supervised by

Md. Fazle Rabbi (Lecturer)

Department of Computer Science & Engineering (CSE)

Bangladesh Army International University of Science & Technology (BAIUST)

The thesis titled "Comparative Study in the prediction of Diabetes Using Machine Learning Technique" submitted by ID No. 1105004,1105014, Session 2020-2021 has been accepted as satisfactory in fulfillment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering (CSE) as B.Sc. Engineering to be awarded by the Bangladesh Army International University of Science & Technology (BAIUST).

Board of Examiners:

1 ………………………………………………………………… Chairman

Mohammad Asaduzzaman Khan
Associate Professor
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology (BAIUST)

2. ………………………………………………………………. Member

Md. Fazle Rabbi (**Supervisor)**
Lecturer
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology (BAIUST)

3. ……………………………………………………………….. Member

Dr. Nargis Parvin
Assistant Professor
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology (BAIUST)

4. ………………………………………………………………… Member

Robaitul Islam Bhuiyan
Lecturer
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology (BAIUST)

5. ………………………………………………………………… Member

Arifa Islam Champa
Lecturer
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology
(BAIUST)

6. ………………………………………………………………… Member

Rifat Hossain
Lecturer
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology
(BAIUST)

7. ………………………………………………………………… Member

Ahmed Arian Sajid
Lecturer
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology
(BAIUST)

8 …..……………………………………………………………… Member

Shimu Sultuna
Lecturer
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology
(BAIUST)

9 …………………………………………………………………….. Member

Mushfiqur Rahman Milton
Lecturer
Department of Computer Science & Engineering (CSE)
Bangladesh Army International University of Science & Technology
(BAIUST)

CANDIDATES DECLARATION

We thusly announce that this accommodation is our claim work and to the finest of our information it contains no materials already distributed or composed by another individual. It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma, or other qualifications.

MD. Ehsanul Haque

Date:

Marzia Khan Turna

Date:

APPROVAL

This thesis titled, “Comparative Study in the prediction of Diabetes Using Machine Learning Technique”, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in April 2021.

Group Members:

MD. Ehsanul Haque

Marzia Khan Turna

Supervisor:

____

Fazle Rabbi
Lecturer
Department of Computer Science and Engineering
Bangladesh Army International University of Science and Technology

ACKNOWLEDGEMENT

We are grateful to our creator Almighty ALLAH for all the blessings and mercy that has given upon us. We are indebted to our thesis supervisor Lecturer Fazle Rabbi sir for all his guidance, assistance, inspirations, motivations and suggestions.

At last, We thank our parents and family for the support they have always provided.

Cumilla Cantonment
April 2021

MD. Ehsanul Haque

Marzia Khan Turna

Contents

ABSTRACT

Diabetes is considered one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting a patient to a diagnostic center and consulting a doctor. Sometimes people call diabetes “borderline diabetes.” The motive of this study is to design a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore four machine learning classification algorithms namely Decision Tree, SVM, Logistic Regression, and Random Forest are used in this experiment to detect diabetes at an early stage. Experiments are performed on direct questionnaire from Sylhet Diabetes Hospital of Sylhet, Bangladesh.The performances of all the algorithms are evaluated on various measures like Precision, Accuracy, F-Measure, and Recall, K-Fold cross-validation, Confusion Matrix. Accuracy is measured over correctly and incorrectly classified instances. Results obtained show Random Forest outperforms with the highest accuracy of 99% comparatively other algorithms.These results are verified using Receiver Operating Characteristic (ROC) curves which is a graphical representation in a proper and systematic manner.

CHAPTER 1

Introduction

1.1 About Diabetes :

Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. It also causes by genes and environmental factors such as viruses that might trigger the disease. Diabetes is not only affected by excessive amounts of added sugars but also other negative factors on the liver like high risk of obesity, height, weight, hereditary factor, and insulin. But the major reason considered is sugar concentration among all factors. In addition to the general symptoms of diabetes, men with diabetes may have a decreased sex drive, erectile dysfunction(ED), and pair muscle strength. Women with diabetes can also have symptoms such as urinary tract infections, yeast infections, and dry itchy skin. Doctors treat diabetes with a few different medications. Some of these drugs are taken by mouth, while others are available as injections. [14] In 2018, 34.2 million Americans, or 10.5% of the population had diabetes. Undiagnosed of the 34.2 million adults with diabetes, 26.8 million were diagnosed, and 7.3 million were undiagnosed. The body is designed to regulate and buffer the amount of glucose dissolved in the blood to maintain a steady supply to meet cell needs. The pancreas, one of your body's many organs, produces, stores, and releases insulin into the bloodstream to bring glucose levels back down.

Insulin is the critical key to the cell's ability to use glucose. Problems with insulin production or with how insulin is recognized by the cells can easily cause the body's carefully balanced glucose metabolism system to get out of control. When either of these problems occurs, Diabetes develops, blood sugar levels surge and crash and the body risks becoming damaged.

1.2 Motivation :

As we can observe the possible causes of diabetes in our intro part this disease arrives in case of high glucose level in blood. As also known, the exalted suger absorbing in blood cells. It is one's main source of energy gained from the food that is consumed. In this modern era of science, this disease can be reduced as the statistics show, [14] in 2019 an estimated 463 million people had diabetes worldwide (8.8% of the adult population).
In 2019, diabetes resulted in approximately 4.2 million deaths. It is the 7th leading cause of death globally. So, to predict the probability of this disease and for improving the result this probability is selected from the dataset of Sylhet Diabetes Hospital, Sylhet, Bangladesh, Which consists of 520 instances.

The specialty of this dataset is, it has been collected using direct questionnaires from the patients of that hospital. In order to find early predictors for diabetes, which in turn would enable the creation of policies for early prevention and adequate early treatment of the diabetes syndrome. Furthermore, this may undoubtedly have a highly beneficial impact on society. Sure enough this undertaking, in order to be fruitful, requires extensive medical records of elderly patients.

1.2Problem statement :

This problem states with the prediction of the possibility by having diabetes of a patient depending on some early symptoms using modern Machine learning techniques. Machine learning fundamentally is the “art of prediction”. And the major role of machine learning in the prediction of diseases. It helps to make good data-driven predictions. Through this, we will be able to determine the diabetes of a patient at an early stage.

CHAPTER 2

Literature Review

In this section, different research works that were envisioned to predict diabetes using data mining have been provided with their remarkable contribution.

In [1] , The author collected dataset using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh.They analyzed the dataset with Naive Bayes Algorithm, Logistic Regression Algorithm, and Random Forest Algorithm and after this applying tenfold Cross-Validation and Percentage Split evaluation techniques. The best result was achieved using Random Forest Algorithm where using tenfold cross-validation 97.4% instances were classified correctly and using percentage split technique, it could classify 99% of the instances correctly

In [3] , the authors used Pima Indians Diabetes Database (PIDD) which is sourced from UCI machine learning repository. In the feature selection process they consider height, weight, hereditary factor and insulin but the major reason considered is sugar concentration among all factors. They tested their dataset with Decision Tree, Support Vector Machine (SVM) and Naive Bayes. They found Naive Bayes outperforms with the highest accuracy of 76.30% comparatively other algorithms where SVM 65.10%, Decision Tree 73.82%.

In [4] , the authors collected 865 data with 9 attributes called Sex, Diastolic B.P, Plasma glucose, Skin fold thick, BMI, Diabetes Pedigree type, No. of times Pregnant, 2 h Serum Insulin and Diabetes probability and used WEKA 3.6.6 for the experiment. They found 100% accuracy with J48 (C4.5), 98.48% with the Decision Tree, 97.85% with the Neural Network, 96.54% with JRip and 95.85% with Naive Bayes algorithm. They also calculated the performance over time.

In [5] , the authors used collected Pima Indians diabetes dataset from the University of California, Irvine (UCI) Repository. For the data processing, the missing values are compensated by the Naïve Bayes (NB) method for data normalization , Adaptive synthetic sampling method (ADASYN) is adopted to oversample the dataset, increasing the number of the minor class so as to achieve a balance of classes. They used Naive Bayes (NB), Adaptive Synthetic Sampling Method (ADASYN), Random Forest (RF) for testing the dataset. The proposed DMP_MI algorithm has outperformed 87.10% other algorithms on accuracy and other classifier performance indicators, and has shown great potential for diabetes prediction.

In [6] , the authors collected UCI Pima Indian Diabetes dataset is needed for model training and testing. For the data processing they find Misssing value and outlier replacement and weighted feature selection. They used Random Forest Algorithm, Weighted Feature Selection Algorithm for the feature selection from dataset. XGBoost is an improvement of boosting algorithm based on Gradient Boosting Decision Tree (GBDT). RF-WFS and XGBoost accuracy 93.75%.

In [7] , the author has created a new model for type 2 diabetes patients treatment. He collected 318 medical records with 9 nominal attributes including the patient’s Gender, Age, Smoking, History of hypertension, Renal problem, Cardiac problem, and Eye problem. The duration of Diabetes Basic control was used as a class level attribute. He used the J48 algorithm and found an accuracy rate of 70.8% and ROC (Receiver operating characteristic) rate was 0.624.

CHAPTER 3

Methodology

3.1 Dataset Description:

This dataset [2] comprises medical detail of 520 instances which includes both male and female patients. The dataset also comprises numeric-valued 17 attributes where the value of one class ’0’ treated as tested negative for diabetes and the value of another class ’1’ is treated as tested positive for diabetes. It includes data about peoples including symptoms that may cause diabetes. This dataset has been created from a direct questionnaire to people who have recently become diabetic, or who are still nondiabetic but having few or more symptoms. This data has been collected from the patients using a direct questionnaire from Sylhet Diabetes Hospital of Sylhet, Bangladesh.

The data preprocessing has been conducted by handling the missing values following the technique of ignoring the tuples with complete value. After preprocessing Dataset description is defined by Table-3.1 and the Table-3.2 represents Attributes descriptions..

Table 3.1 Description of dataset

	No. of attributes	No. of instances
Diabetes symptom dataset	16	520

Table 3.2 Description of attribute

Attributes	**values**
Age	1.20-35, 2.36-45, 3.46-55, 4.56-65, 6 above 65
Sex	1.Male, 2.Female
Polyuria	1.Yes, 2.No
Polydipsia	1.Yes, 2.No
Sudden weight loss	1.Yes, 2.No
Weakness	1.Yes, 2.No
Polyphagia	1.Yes, 2.No
Genital thrush	1.Yes, 2.No
Visual blurring	1.Yes, 2.No
Itching	1.Yes, 2.No
Irritability	1.Yes, 2.No
Delayed healing	1.Yes, 2.No
Partial paresis	1.Yes, 2.No
Muscle stiffness	1.Yes, 2.No
Alopecia	1.Yes, 2.No
Obesity	1.Yes, 2.No
Class	1.Positive, 2.Negative

3.2 Preprocessing :

Time to time processing data and reduction of dimensionality becomes indispensable techniques in recent knowledge of finding scenarios, controlled by increasingly large datasets. The method works to help to reduce the complexity inherent to real-world datasets so that they can be easily processed by current data mining solutions. In the field of modern data science a wide range of new and sophisticated computational methods as well as its tools for building predictive models and performs enhanced data analysis. Benefits are also observed as this approaches include, a faster and more precise learning process, the comparatively more understandable structure of raw data. However, in the context of data preprocessing techniques for data streams have a long road ahead of them, despite online learning is growing in importance thanks to the development of the Internet and technologies for massive data collection.

Throughout this survey, we summarize, categorize and analyze those contributions on data preprocessing that cope with streaming data. This work also takes into account the existing relationships between the different families of methods (feature and instance selection, and discretization). For clinical importance, these methods are used to offer support in tasks such as decision-making based on the patient's data. Previous collected patient data can be used to build a predictive model which provides a prediction for the clinical outcome. Clinicians can act on this information and promptly react to possible or likely adverse events. Data analytics can of course also be applied to analyze retrospective clinical data of the aging population which can be crudely separated into healthy and frail people.

At first, the data were converted from nominal to numeric form and checked. If necessary then corrected. Furthermore, in order to reduce the influence of mixed features, filtering is very inevitable to reduce a large dataset as we can remove the feature which contains constant values in the dataset are known as Constant Features. These are redundant data available in the dataset. Quasi-constant features, as the name suggests, are the features that are almost constant. In other words, these features have the same values for a very large subset of the outputs. Moreover, duplicates are an extreme case of nonrandom sampling, and they bias-fitted model. Including them will essentially lead to the model overfitting. They are not real data that coincidentally have values that are identical. The presence of this feature has no effect on the target, so it is good to remove these features from the dataset. But no Duplicate Features were founded in the dataset. So " Class " features were selected from the dataset as this feature is positively and negatively correlated with others.

3.3 Classification Algorithm

3.3.1 Random Forest :

Random forest [8] is a pliable supervised machine learning algorithm that generates an excellent outcome mostly even without hyper-parameter tuning, a. A liable algorithm because of its clarity and diversity. Both classification and regression tasks can effortlessly be terminated with this methodology. The "forest" term goes with the idea by using an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. while growing the trees. While growing the trees in lieu of searching for the most important feature by splitting a node, it hunts for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in a random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. Fig 3.1 shows a sample of graphical representation:

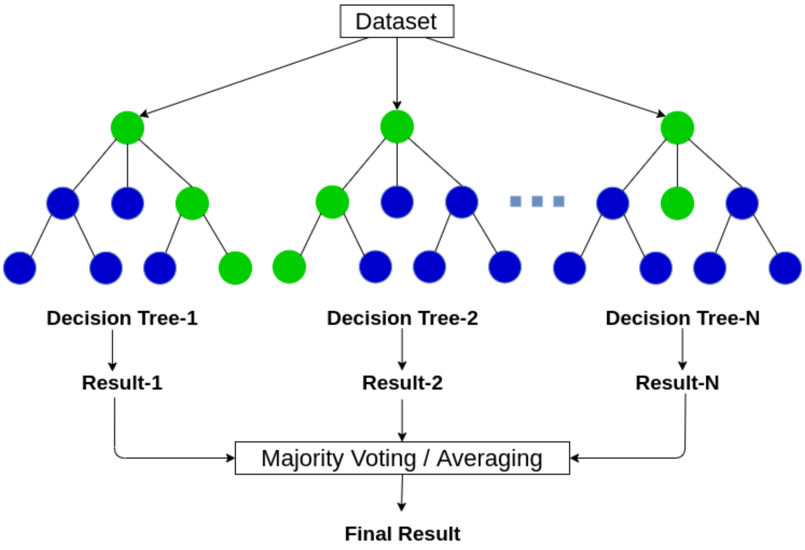


Fig : 3.1 Random Forest[9]

3.3.2 Logistic Regression :

This method [16] used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to Linear Regression except that how it is used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. Here we are dealing with classification issues so the technique is included here. Fig 3.2 shows a sample of graphical representation:

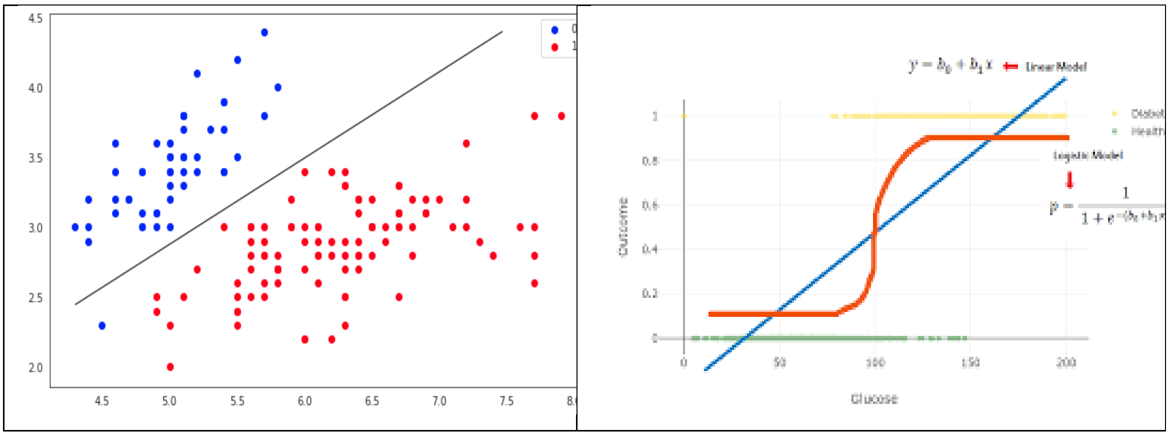


Fig : 3.2 Logistic Regression[15][16]

3.3.3 Support Vector Machine (SVM):

Support Vector Machine is a learning methodology based on Vapnik’s statistical learning theory. It is a set of related supervised learning methods used for classification and regression and applicable for both linear and nonlinear data. A support vector machine constructs a hyperplane or set of hyperplanes in a high dimensional space by 4 basic concepts of SVM-separating hyperplane, maximum margin hyperplane, soft margin, the kernel function.

Given a two-class training sample, the aim of a support vector machine is to find the best highest-margin separating hyperplane between the two classes. For better generalization hyperplane should not lies closer to the data points belong to the other class. Hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors. The SVM finds the optimal separating hyperplane by maximizing the distance between the two decision boundaries. Mathematically, we will maximize the distance between the hyperplane which is defined by wT x + b = −1, and the hyperplane defined by wT x + b = 1 This distance is equal to 2 w. This means we want to solve max 2 w. Equivalently we want min w| 2. The SVM should also correctly classify all x(i), which means yi (wT xi + b) >= 1, ∀i ∈ {1, ¢¢, N}. Fig 3.3 shows a sample of graphical representation:

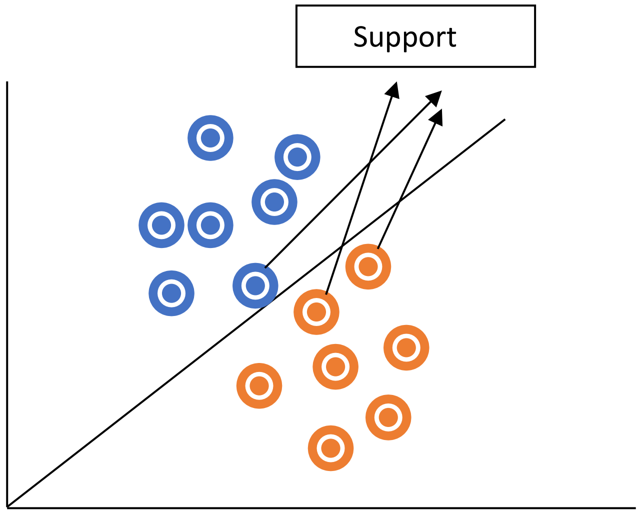


Fig: 3.3 Support Vector Machine (SVM)

3.3.4 Decission Tree:

A tree consists of roots, branches, and leaves. Decision Tree [10] followed this structure of root node, branches, and leaf nodes. Testing an attribute is on every internal node, the outcome of the test is on a branch, and the class label, as a result, is on the leaf node. The parent is the "root node" of all nodes as it is the topmost node in the Tree. A decision tree is a tree where each node shows a feature (attribute), each link (branch) shows a decision (rule) and each leaf shows an outcome (categorical or continuous value). As decision trees mimic the human level thinking interpretations. The entire data and process a single outcome at every leaf. Fig 3.4 shows a sample of graphical representation:

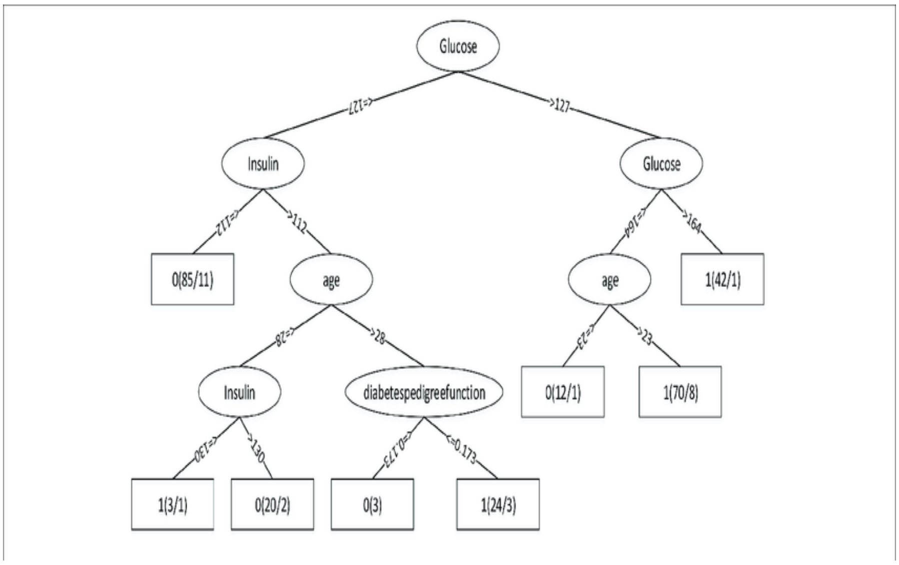


Fig: 3.4 Decision Tree[11]

3.4 Evaluation Measure :

Evaluation provides a study of a program, practice, intervention, or initiative to understand how well it achieves its goals also helps determine what works well and what could be improved in a program or initiative. In order to evaluate our model, we use the methods of accuracy, confusion matrix, and ROC curve.

3.4.1 Accuracy :

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

Here the formula is evaluated as given below:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Where ,

FP=False Positive Rate

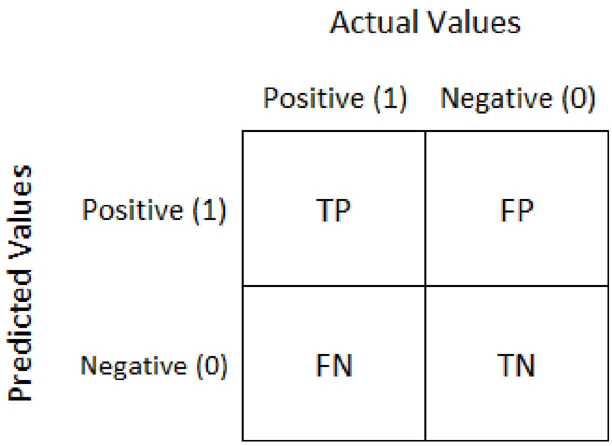
TP=True Positive Rate

FN=False Negative Rate

TN=True Negative Rate

3.4.2 Confusion Matrix:

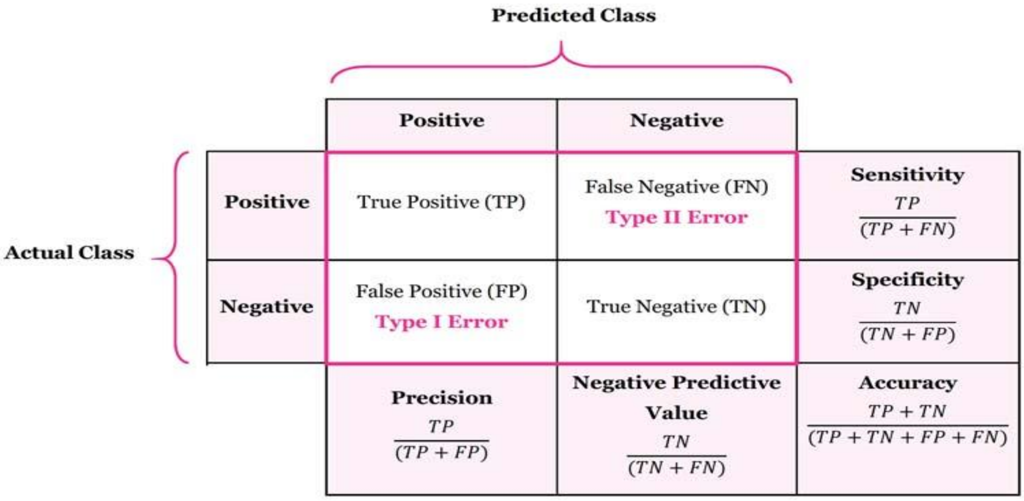
A Confusion matrix [13] is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.The rows represent the predicted values of the target variable. This matrix defines accuracy, precision, recall,fi-score, and support for our dataset. The visual representation Fig:3.5 can be:



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig : 3.5 Sample Confusion Matrix [17]

From Confusion Matrix we get the results of precision, recall, fi-score, support, accuracy, sensitivity, specificity. A sample visual matrix representation Fig:3.6 is given below:



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Fig : 3.6 Sample Confusion Matrix [13]

3.4.3 ROC (Receiver Operating Characteristic Curve)

ROC curves [12] are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The ROC does not depend on the class distribution. This makes it useful for evaluating classifiers predicting rare events such as diseases or disasters. In contrast, evaluating performance using accuracy (TP +TN)/(TP + TN + FN + FP) would favor classifiers that always predict a negative outcome for rare events. Fig 3.7 shows a sample of graphical representation:

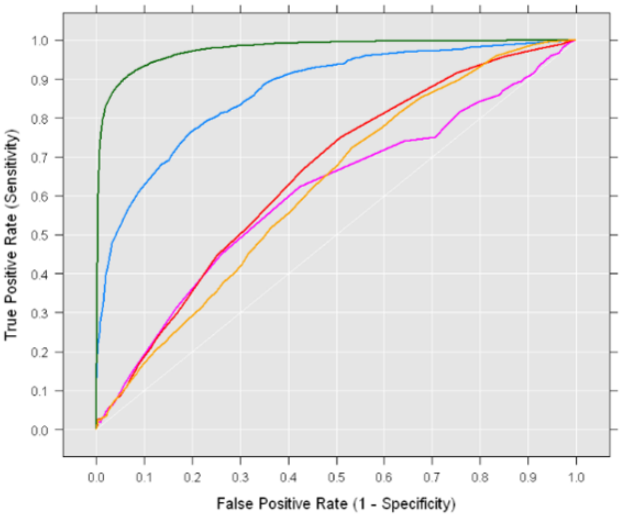


Fig:3.7 ROC Curve[12]

CHAPTER 4

Result & Descussion

This section covers the whole structural design and performance of our model. Sorted with the parts of the environment, data preprocessing & featuring, filtering, classification methods, comparison, accuracy, and graph plotting.

4.1 Environment Setup:

Domains we choose to complete our findings are Python & Jupytar Notebook. Python is a dynamically typed and easier interpreted language that can generate an output without any delay and straightforwardly. The platform Jupyter Notebook allows to create and share documents that contain live code, equations, visualizations, and explanatory text including data cleaning and transformation, numerical simulation, statistical modeling, machine learning, and much more.

4.2 Data Preprocessing:

The dataset we have assembled is categorized by converting nominal data into numerical data, then features are selected using filtering method and removed constant, quasi constant, and duplicate data. filtering is very inevitable to reduce a large dataset as we can remove the feature which contains constant values in the dataset are known as Constant Features. These are redundant data available in the dataset. Quasi-constant features, as the name suggests, are the features that are almost constant. In other words, these features have the same values for a very large subset of the outputs. Moreover, duplicates are an extreme case of nonrandom sampling, and they bias-fitted model. Including them will essentially lead to the model overfitting.

4.3 Accuracy of classification mathod and AUC

By applying SVM, Decision Tree, Logistic Regression, Random Forest machine learning algorithm the accuracy of predicting from the given dataset resulted. A “Hold out” results in the dataset that is called “Hold out accuracy “. We get the accuracy of Random Forest is 99%, Logistic regression is 93%, SVM is 92%, the Decision tree is 98% that is the total can variable. It can not give the actual value because the random folder is selected here. That’s the reason we applied a confusion matrix to get the right accuracy. Table 4.1 shows the likelihood accuracy of classification mathod.

Table 4.1 Comparison of perfformance parameters using percentage

Algorithm	**Accuracy**	**AUC**
Random Forest	99%	**-**
Logistic Regression	93%	0.927
Support Vector Machine	92%	0.917
Decission Tree	98%	0.974

4.4 K-fold Cross-Validation mathod and AUC

With a view to getting the more accurate value, we applied the K-fold Cross Validation method. The whole dataset was K-fold taking the value of K=5 as well as K=10. Some results from applying this method are found that are For K=5 we get the accuracy of each fold is 84%, 88%, 91%, 81%, 92% whereas the average accuracy is 87%. Taking K=10 the accuracy results are 87%, 88%, 86%, 89%, 98%, 94%,78%, 96%, 92%, 98% of each fold whereas the average accuracy is 91% . Table 4.2 and 4.3 shows the comparison between ten-fold and five-fold cross-validation

Table 4.2 Comparison of ten-fold cross-validation

Feature Drop	k-fold	Accuracy	AUC
No feature	K-10	91%	0.987
One feature		90%	0.987
Two features		89%	0.9625

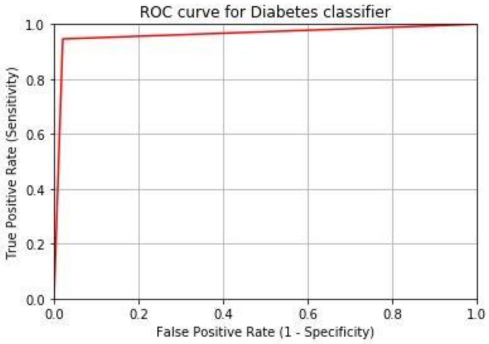
Table 4.3 Comparison of five-fold cross-validation

Feature Drop	k-fold	Accuracy	AUC
No feature	K-5	90%	0.962
One feature		88%	0.953
Two feature		87%	0.922

Eventually, we get more accurate results without dropping any feature so all features have taken to get the best accuracy. The literature review shows that with 10 K-folds we get the finest quality of the model.

4.5 ROC Curve (receiver operating characteristic curve)

Besides all these, we also clarify our results with a ROC curve that allows the diversified solution to define the best model. The actual graph results we get from our model,



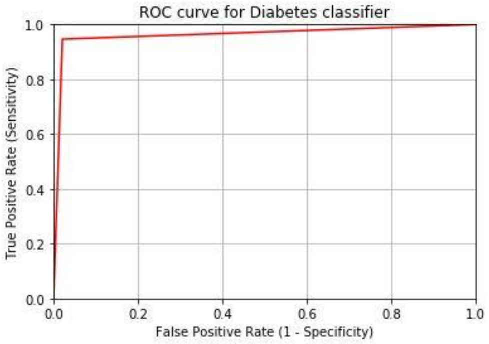
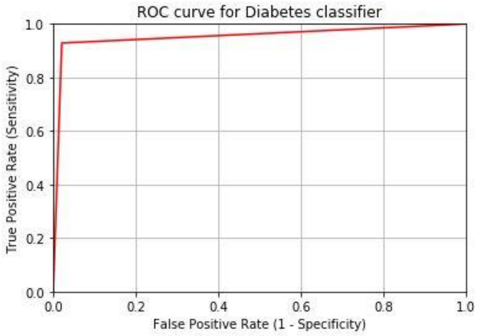


Fig:4.1 K-fold cross validation with all feature for k=10

Fig:4.2 K-fold cross validation with all feature for k = 5

After dropping one feature we get this average value for k=5 is 88% and k=10 is 90%



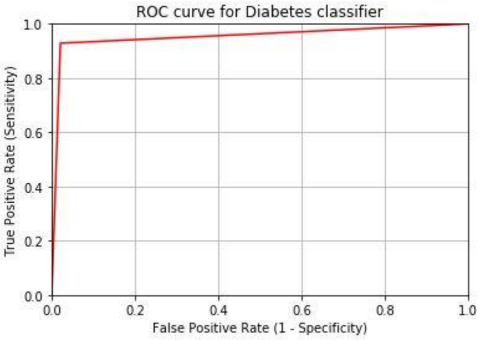
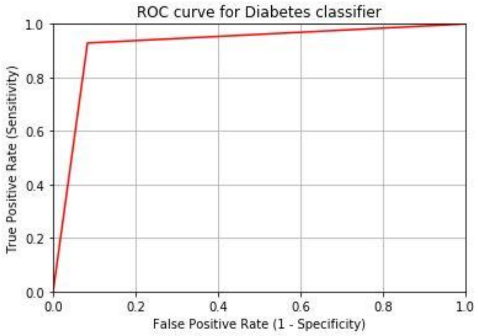


Fig:4.3 K-fold cross validation with One feature drop for k=10

Fig:4.4 K-fold cross validation with One feature drop for k = 5

and dropping two this average value goes for k=5 is 87% and k=10 is 89% .



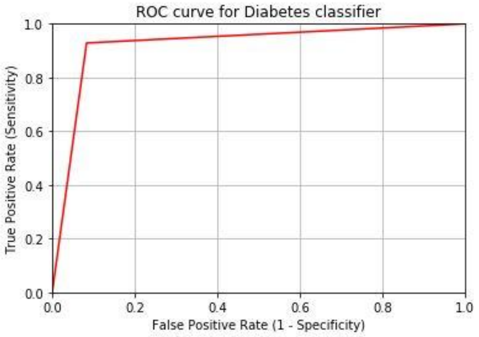


Fig:4.5 K-fold cross validation with Two feature drop for k=10

Fig:4.6 K-fold cross validation with Two feature drop for k = 5

The AUC value for For Random Forest is 98%, Logistic Regression is 92%, SVM is 91%, Decision Tree is 95%. The resulted ROC graphs are given below for each machine learning classification:

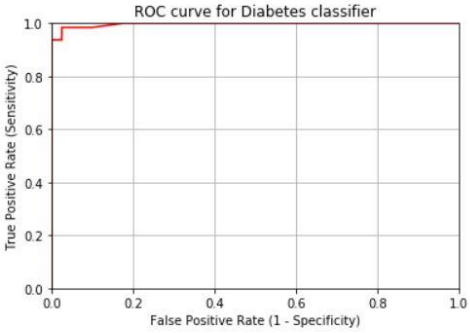


Fig:4.7 ROC Curve For Random Forest Algorithm

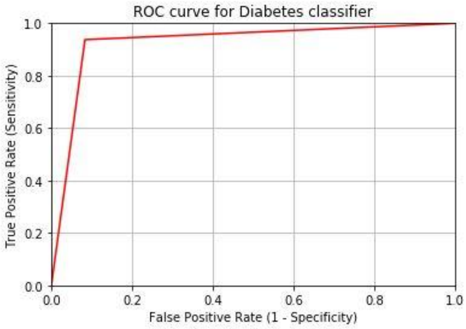


Fig:4.8 ROC Curve For Logistoc Regression Algorithm

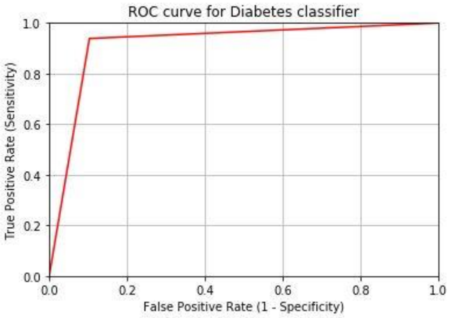


Fig:4.9 ROC Curve For SVM Algorithm

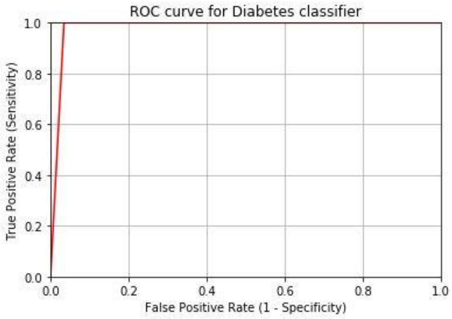


Fig:4.10 ROC Curve For Decision Tree Algorithm

4.6 Performance measure between models

Author Name	Dataset	Methodology	Result
Islam et al.	Sylhet Diabetes Hospital of Sylhet	Naive Bayes Algorithm, Logistic Regression Algorithm, Random Forest , tenfold Cross-Validation, Percentage Split evaluation techniques	Random Forest Algorithm where using tenfold cross-validation 97.4% instances were classified correctly and using . percentage split technique, it could classify 99% of the instances correctly
Sisodiaa et al.	Pima Indians Diabetes Database (PIDD)	Decision Tree, Support Vector Machine (SVM) and Naive Bayes	Naïve Bayes 76.30% SVM 65.10%, Decision Tree 73.82%
Kumar et al.	A five year sample dataset is created. contains 865 instances with 9 attributes	J48 (C4.5), Decision Tree, Neural Network, Jrip, Naive Bayes	J48 (C4.5) 100%, Decision Tree 98.48%, Neural Network 97.85%, Jrip 96.54%, Naive Bayes 98.48%
Wang et al.	Pima Indians Diabetes Database (PIDD)	Naive Bayes (NB), Adaptive Synthetic Sampling Method (ADASYN), Random Forest (RF) DMP_MI algorithm	DMP_MI algorithm 87.10%
Xu et al.	UCI Pima Indian Diabetes dataset	Random Forest , Weighted Feature Selection, XGBoost	XGBoost 93.75%
Ahmed et al.	JABER ABN ABU ALIZ clinic center diabetes Database	J48	J48 70.8%,
Our model	Sylhet Diabetes Hospital of Sylhet	Random Forest, Logistic Regression, Support Vector Machine (SVM), Decission Tree	Random Forest 99%, Logistic Regression 93%, Support Vector Machine (SVM) 92%, Decission Tree 98%

CHAPTER 5

Conclusion and Future work

Diabetes disease has an adverse effect on human life if it is not detected early and treated at the immediately at the early stage.This causes unimaginable increase of glucose (blood sugar) in human’s body. Rises in blood sugar result to inadequatereproduction of insulin in the body or failure of the body to respond to the produced insulin. The basic symptoms of diabetesare intensify thirst, hunger and frequent urination. Diabetes can occur due to several factors such as unhealthy consumption of food substance, heredity and obesity. This dataset comprises medical detail of 520 instances which includes both male and female patients. The dataset also comprises numeric-valued 17 attributes where the value of one class ’0’ treated as tested negative for diabetes and the value of another class ’1’ is treated as tested positive for diabetes. It includes data about peoples including symptoms that may cause diabetes. This dataset has been created from a direct questionnaire to people who have recently become diabetic, or who are still nondiabetic but having few or more symptoms. This data has been collected from the patients using a direct questionnaire from Sylhet Diabetes Hospital of Sylhet, Bangladesh.Where there are 314 positive values and 186 are negative values. The dataset assembled are categorized by converting nominal data into numerical data, then features are selected using filtering method and removed constant, quasi constant, and duplicate data. By applying SVM, Decision Tree, Logistic Regression, Random Forest machine learning algorithm the accuracy of predicting from the given dataset resulted. With a view to getting the more accurate value, we applied the K-fold Cross Validation method. The whole dataset was K-fold taking the value of K=5 as well as K=10. Results obtained show Random Forest outperforms with the highest accuracy of 99% comparatively other algorithms.

This work, we introduced the factors to develop a trustworthy predictive healthcare diabetes system. The association rule-based on supporting algorithms provide a unique manner of selecting attributes and classification tech-niques.We are ardent to make an effort on different types of classifiers and build a model to enhance the accuracy of diabetes prediction. This paper goes after to reach high prediction accuracy. Machine Learning (ML) technologies that are applied in the last nine years including Deep Learning (DL) classifiers were reviewed regarding their frequency of use and accuracy results.Finally we get more accurate results without dropping any feature so all features have taken to get the best accuracy. The literature review shows that with 10 K-folds we get the finest quality of the model.

As future work, the non-used classifiers can also be applied to the given datasets as well as other datasets in a combined model to enhance further the accuracy of predicting the Diabetes disease. The datasets are not balanced in our model that can be updated by using oversampling, under-sampling, or other techniques.

References:

1. Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125. (5 September, 2020)
2. Early stage diabetes risk prediction dataset. Data Set: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. (9 September, 2020)
3. Deepti Sisodiaa , Dilip Singh Sisodiab : Prediction of Diabetes using Classification Algorithms. Deepti Sisodia et al. / Procedia Computer Science 132 (2018) 1578–1585 (5 july, 2020)
4. Kumar, V., Valide, L.: A data mining approach for prediction and treatment of diabetes disease. Int. J. Sci. Invent. Today (2014). ISSN 2319-5436 (8 july, 2020)
5. Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng and D. N. Davis, "DMP_MI: An Effective Diabetes Mellitus Classification Algorithm on Imbalanced Data With Missing Values," in IEEE Access, vol. 7, pp. 102232-102238, 2019, doi: 10.1109/ACCESS.2019.2929866 (6 july, 2020)
6. Z. Xu and Z. Wang, "A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier," 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI), 2019, pp. 278-283, doi: 10.1109/ICACI.2019.8778622 (5 july, 2020)
7. Ahmed: Developing a predicted model for diabetes type 2 treatment plans by using data mining (2016b) (7 july, 2020)
8. A Complete Guide to the Random Forest Algorithm: https://builtin.com/data-science/random-forest-algorithm#how (11 october, 2020)
9. A Simple Analogy to Explain Decision Tree vs. Random Forest: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ (11 october, 2020)
10. DecisionTree: https://www.researchgate.net/publication/330138092_Study_and_Analysis_of_Decision_Tree_Based_Classification_Algorithms (5 january, 2021)
11. Study and Analysis of Decision Tree Based Classification Algorithms: https://www.researchgate.net/figure/Decision-tree-structure-by-using-all-features-and-Pima-Indians-dataset-From-this_fig2_328766758 (5 january, 2021)
12. What is a ROC Curve and How to Interpret It: https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/ (25 january, 2021)
13. What is Confusion Matrix and Advanced Classification Metrics: https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html (10 March, 2021)
14. Diabetes facts & figures: https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html (13 March, 2021)
15. Logistic Regression using Python: https://www.c-sharpcorner.com/article/logistic-regression/ (27 March, 2021)
16. Predicting Diabetes using Logistic Regression with TensorFlow.js: https://towardsdatascience.com/diabetes-prediction-using-logistic-regression-with-tensorflow-js-35371e47c49d (27 March, 2021)
17. Understanding Confusion Matrix: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62 (9 April, 2021)