

**Diabetics Prediction in Real Life Data Using Machine Learning Algorithm: A Case
Study in Bangladesh**

BY

Sobuj Mia

ID: 181-15-10699

Razib Hasan

ID: 181-15-10786

Md. Juweel Rana

ID: 181-15-10919

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

Supervised By

Md. Zahid Hasan

Assistant Professor

Department of CSE

Daffodil International University

Co-Supervised By

Ms. Nazmun Nessa Moon

Assistant Professor

Department of CSE



Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

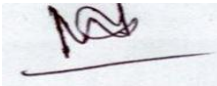
DHAKA, BANGLADESH

January 2022

APPROVAL

This Project/internship titled “**Diabetics Prediction in Real Life Data Using Machine Learning Algorithm: A Case Study in Bangladesh**”, submitted by **Sobuj Mia, Razib Hasan** and **Md. Juweel Rana**. ID No: **181-15-10699, 181-15-10786** and **181-15-10919** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 2nd January, 2022.

BOARD OF EXAMINERS



Chairman

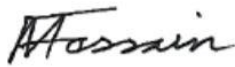
Dr. Md. Ismail Jabiullah

Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Dr. Md. Fokhray Hossain

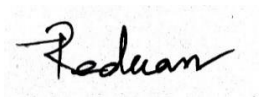
Internal Examiner

Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



Internal Examiner

Md. Reduanul Haque

Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University



External Examiner

Dr. Mohammad Shorif Uddin

Professor

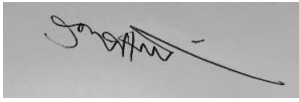
Department of Computer Science and Engineering

Jahangirnagar University

DECLARATION


We hereby declare that, this project has been done by us under the supervision of **Md. Zahid Hasan, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

SUPERVISED BY



Md. Zahid Hasan
Assistant Professor
Department of CSE
Daffodil International University

CO-SUPERVISED BY



Ms. Nazmun Nessa Moon
Lecturer
Department of CSE
Daffodil International University

SUBMITTED BY:



(Sobuj Mia)

ID: 181-15-10699

Department of CSE

Daffodil International University

Razib

(Razib Hasan)

ID: 181-15-10786

Department of CSE

Daffodil International University

Juweel Rana

(Md. Juweel Rana)

ID: 181-15-10919

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year Thesis successfully.

We really grateful and wish our profound our indebtedness to **Supervisor Md. Zahid Hasan, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to -----, -----, and Head, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

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

ABSTRACT

There are many diseases in Bangladesh where diabetes is ordinary disease of human body. A human is affected when his body sugar level over a period is pretty much high. It also causes of stroke, heart attack, kidney failure and blindness etc. It is possible to control if you understand the earlier stage and can save a human life. This is a disease which is in growing day by day. The motive of this study to do the measurement of the performances of some popular Machine Learning algorithms. In recent years, Machine Learning is a wonderful platform which has a huge impact on different corner of science and technology including medical sectors. There are many algorithms in Machine Learning. But in this paper work, we have used five popular ML algorithms which is Gaussian Naïve Bayes, Random Forest, Support Vector Machine, Logistic Regression and Decision Tree to find out the measurement of performances. These algorithms have been trained and tested on real data for diabetes patients in Bangladesh. The performances of these techniques are enlisted. We have used diabetes patient dataset, conducted in 2021, derived from Islami Bank Hospital and Diagnostic Center. The dataset consists of 485 data which contains 267 true class and the rest of data is false class. When we applied algorithms then we find that Random Forest based algorithm gives 97% accuracy. So, the RF based classifier is better than other algorithms.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii-iii
Acknowledgements	iv
Abstract	v
CHAPTERS	
CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	2
1.3 Rationale of the Study	2
1.4 Research Questions	2
1.5 Expected Outcome	3
1.6 Report Layout	3
CHAPTER 2: BACKGROUND	4-7
2.1 Introduction	4
2.2 Related Works	4-5
2.3 Research Summary	5-6
2.4 Scope of the Problem	7
2.5 Challenges	7

CHAPTER 3: RESEARCH METHOLOZY	8-20
3.1 Introduction	8
3.2 Research Subject and Instrumentation	8
3.3 Description of Machine learning Algorithm	9-10
3.3.1 Decision Tree (DT)	9
3.3.2 Random Forest (RF)	9
3.3.3 Support Vector Machine (SVM)	10
3.3.4 Logistic Regression (LR)	10
3.3.5 Gaussian Naïve Bayes (GNB)	10
3.4 Data Collection Procedure and Dataset description	11-12
3.5 Implementation Requirement	13
3.6 Data Visualization	14-15
3.7 Data Pre-Processing	16
3.8 Features Selection	17
3.8.1 Features Extraction	17-18
3.9 Build Model	19
3.10 Appling Algorithm	20
CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION	21-29
4.1 Introduction	21
4.2 Measurement Of Classification Techniques	21-23
4.3 Analysis Of The Results	23-24
4.4 Performance Evaluation	25-27
4.5 Receiver Operating Curve	28
4.6 Experimental Results	29
4.7 Descriptive Analysis	29

CHAPTER 5: CONCLUTION AND FUTURE SCOPE	30-32
5.1 Summary of the Study	30
5.2 Conclusion	31
5.3 Recommendations	31
5.4 Implication for Further Study	32
REFERENCES	33-35
APPENDIX	36-39
PLAGIARISM CHECKER SCREENSHOT	40

LIST OF FIGURES

FIGURES	PAGE NO
Figure 2.3.1: Whole Process of Finding Prediction	6
Figure-3.5.1: Research Procedure	13
Figure-3.6.1: Data Visualization	14-15
Figure-3.7.1: Visualization After Preprocessing	16
Figure-4.3.1: Heat Map for Correlated columns	23
Figure 4.3.2: Performance Measurement Accuracy	24
Figure 4.4.1: Decision Tree Confusion Matrix	25
Figure 4.4.2: Random Forest Confusion Matrix	25
Figure 4.4.3: Support Vector Machine Confusion Matrix	26
Figure 4.4.4: Logistic Regression Confusion Matrix	26
Figure 4.4.5: Gaussian Naïve Bayes Confusion Matrix	27
Figure 4.5.1: ROC Plot	28

LIST OF TABLES

TABLES	PAGE NO
Table 3.4.1: Description of dataset	12
Table 3.8.1: Shown Top Features Selection Wise X-train value	18
Table 4.4.6: Classification Performance Measurement	27
Table 4.6.1: Performance Measurement Accuracy	29

LIST OF ABBREVIATION

- DIU** – Daffodil International University
- CSE** – Computer Science and Engineering
- ML** – Machine Learning
- RQ** – Research Question
- SVM** – Support Vector Machine
- LR** – Logistic Regression
- RF** – Random Forest
- GNB** – Gaussian Naïve Bayes
- DT** – Decision Tree

CHAPTER 1

INTRODUCTION

1.1 Introduction

Diabetes is such a kind of prolonged persistent disease which has been affected in all ages of people. It is not depend on the age. This disease affects our health condition that our whole body from food into energy. Blindness, kidney failure, heart attack, stroke and lower limb imputation is a vital cause of diabetes patients. It has been growing more and more rapidly in low and middles income countries. But in high income countries have been not growing rapidly. About 642 million people are projected to have diabetes in 2040.

There are three kinds of diabetes such as 1) Immature or childhood diabetes (type 1 diabetes), 2) Type 2 or adult diabetes, 3) Gestational or type 3 diabetes. Lack of insulin production is occurred in type 1 diabetes and happens in the beginning of age. Type 1 symptoms are excessive excretion of urine (polyuria, much and more-thirsty (polydipsia), more hunger constantly (polyphagia), weight loss, visual blurring, itching and irritability or fatigue. Adult or type 2 is a very well-known form of diabetes. It contains a large number of man and women. Type 1 and Type 2 symptoms are very much similar. Gestational or type 3 diabetes is the common problems for human being.

In this research work, we took a step forward through combining patient's data from renowned hospital in Bangladesh. By doing the machine learned with the training dataset. As our model is based on diabetes patient's test report data. So, it will be perfect work in all terms and conditions. Then, we develop a supervised learning model to find out the accuracy of our model.

1.2 Motivation

In research, we think that health sector is a very interesting area. We are working on diabetes patients on basis of real time data. The most important things are that we are working on real time data. Most of the researchers do not work on real data. Then we clearly inspired to do this research.

1.3 Rationale of the Study

We are working on Bangladeshi diabetes patient on basis of real time data. The challenges of the experimental way how they fulfill their goal by using the different learning methods. Here any researcher can find improving this area to achieve their desired goal. There is not enough research based on real data in Bangladeshi diabetes patient dataset. We have taken on hand that.

1.4 Research Question

We have selected some question as our research work which is being answered stepwise.

- 1) Why do we need research?
- 2) How will we perform the technique?
- 3) How do we evaluate the model?
- 4) What is the accuracy of the algorithm?
- 5) In term of accuracy, which is the best algorithm.
- 6) Which algorithm is the worst in terms of accuracy?

1.5 Expected Outcome

The measurement models performance will be tested currently that we have a constructed by way of usage of different types of machine learning algorithm. The things would have tested that how good algorithm response to our dataset.

The research work will deliver on awesome result or accuracy by using the model.

1.6 Report Layout

Chapter 1 Discusses our thesis motivation, the study's rationale, the research question, and the expected outcome.

Chapter 2 Introduce with the Background history of the research. It additionally gives us the facts of related works this research. Challenges also are included right here.

Chapter 3 Discusses the methodology used in our study project. Data mining and machine learning techniques are described in detail. Approximately the rate collection processes are also mentioned here.

Chapter 4 Discuss the specifics of the conclusion, as well as the ins and outs of that project's experiment and results.

Chapter 5 Discuss the future scope of our research and how to carry out the thesis.

CHAPTER 2

BACKGROUND

2.1 Introduction

The important process is to find out the patient's diabetes condition on basis of some symptoms. We can find out the positive and negative class or result on basis of according to data. We are studying when a patient is affected in a diabetes. Without background study we don't know about this.

But our technique is about the machine learned through different types of Machine Learning Algorithms. When the machine learned it is capable to measure the accuracy of the model in aspect of algorithms.

2.2 Related works

In under this section, we are discussing on a several momentous works are just related of the mentioned problems was discussed.

Diabetes is a tremendous problem most of the person in Bangladesh populace is laid low with this disorder. Various researchers have labored to expecting signs of diabetes with the aid of applying unique approaches which include machine learning. Few of them have also implemented neural community and genetic set of rules. On the grounds that the diabetes prediction problems is supervised in nature, this methods of gadget gaining knowledge of, Data Mining and also ANN were carried out by the using many.

A few intently associated works are mentioned on this phase. We are the used research have used real life patient's data for diabetes prediction. The different techniques applied by researchers can be extensively labeled as gadget getting to know famous computer algorithm. Long short-time period memory that it has been used by them and then functions had been extracted Support Vector Machine (SVM).to our result, they located by

completely excessive accuracy is 82%. Carried out 3 system studying strategies. There are Decision Tree (DT), Naïve Bayes (NB) and Support Vector Machine (SVM) on PIDD with a view to expecting the diabetes. Naïve Bayes algorithm turned into a result to be 97% correct. And the logistic regression, Artificial Neural Network (ANN) and Decision Tree (DT) are identify the danger of diabetes and before diabetes based totally on 20 hazard factors which blanketed age, gender, glucose level, blood pressure, serum uric-acid, serum creatinine, smoking, insulin result, body mass index (BMI), Diabetes pedigree, polyuria, polydipsia, weight loss, polyphagia, visual blaring, skin problem, feel's weakness, itching, irritate, class. Decision tree DT became located to provide satisfactory outcomes some of the three strategies. We have been involved in discovering chance of serious syndrome and diabetes. The prediction Naïve Bayes and choice tree model have been attached and the pattern of training set became accomplished by k-medoids sampling. of their take a look at, NB outperformed the others.

2.3 Research Summary

This Research study make us known that our research is based on Supervised Machine Learning algorithms. Classification Model Problem is used in our research work. In this paper work, we have learned very deeply about Machine Learning classification model and the related algorithms.

At first, handling the data and preprocessing the data consequently. Then we divided into two portion such as training and testing set. Then we are performing algorithms on this set and find out the diabetes prediction and accuracy of the model.

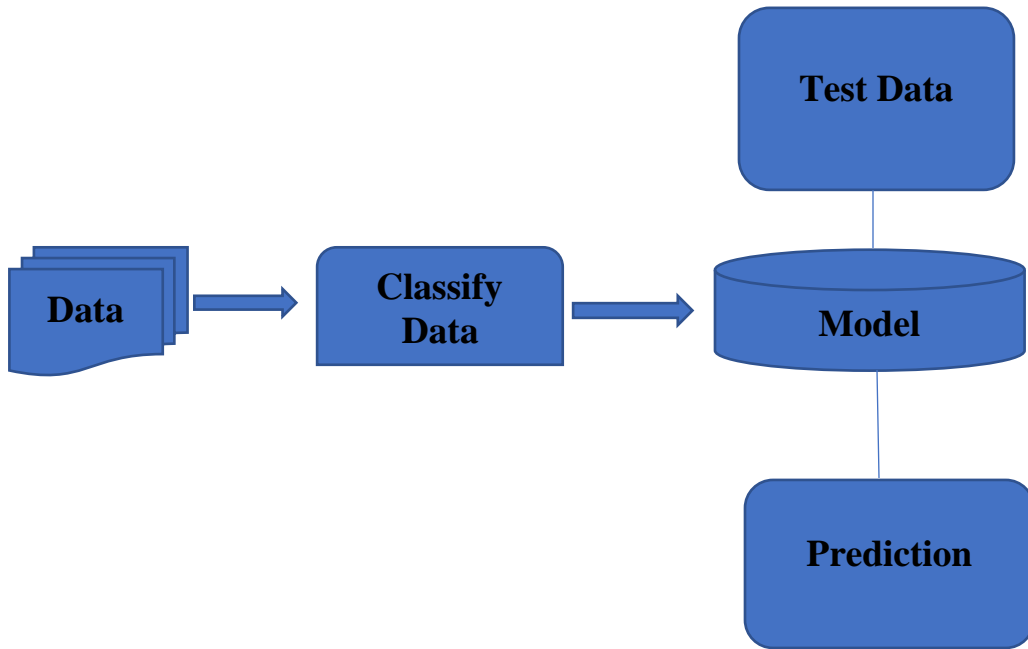


Figure 2.3.1: whole process of finding prediction

2.4 Scope of the Problem

The working opportunity of our research is absolutely very wide and broad. We can use Machine Learning algorithm for finding the diabetes disease prediction and accuracy of our model. As our dataset is almost very new. So, there is a big opportunity to detect the problems and solve it.

In future, researchers are working on this desired topic. Here is a lot of scope to do well day by day.

2.5 Challenges

The dare of our diabetes disease prediction research study is implementing the algorithm is a very hard task. At first, we have to acquire a knowledge about ML algorithms.

We collected a real time data from renowned hospital in Bangladesh. For data collection at first, we are discussing a hospital administration office to give us some data. Then the authority tells us to permission from our university. Then we took a permission from university via application. Then the permission recommendation letter shows in the hospital. The hospital authority checked the recommendation letter. Then they decided to permission for collecting the data. So, Data collection is also challenging work.

When a researcher goes for a research, basically he/she faces new things. So, keep patience for working can be a challenge.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

Nominal data is complex to read out in algorithm from numeric data. The machine learning algorithm does not know about text data. It just known about numeric data. So, we cannot use the text data immediately and it cannot provide us a numeric result. As all we know our research is based on finding accuracy. For this reason, we have to need a numeric data. So, text data transforms into a numeric data is also a challenging for us.

First of all, we implement the processing on nominal and transformed, the nominal data into the numeric data form, then we have used those our pre-processed data into the algorithms. We have used supervised Machine learning algorithms to find out the accuracy in particularly. The working procedure is discussed in the next.

3.2 Research Subject and Instrumentation

Our research subject and aim very simple but interesting topics. We are working on diabetes patient on basis of real data in Bangladesh. We have collected data from hospital. In our research, we are trying to find out the accuracy of existing algorithms based on our diabetes dataset.

At present, Machine Learning is a very familiar and also popular for the research. Many of the researchers are working on this Machine Learning algorithms. We are not exceptional. So, we have worked with some of ML algorithms. So, we learned about ML algorithms at first, we also learned python programming language.

Now, here in our research all Machine Learning Algorithms description.

3.3 Describe of Machine Learning Algorithms

3.3.1 Decision Tree (DT)

It has a large-scale forecasting technique with applications in a variety of places. In process, Decision Trees are created use for algorithms method. Discovers several path into divide a data set that's it is called on distinct criteria conditions. DT means Decision tree is a part of supervised learning. This is a machine learning algorithm. It has also every use classification problem and regression problem. In this algorithm that create a one model base to their target value. In algorithm to solve this problem. To work this algorithm for training set. It is appropriate because it does not necessitate any parameter settings. For the sake of knowledge discovery. The rules that the decision tree follows. The statements that follow are often in the form of if-then-else sentences. Decision trees are used to classify data without the use of rules.

3.3.2 Random Forest (RF)

The Random Forest algorithm from RF Means is a well-known supervised algorithm that can perform both regression and classification. Leo was the first to propose the RF mean's Random Forest. In the process, RF provides Decision Trees, which are used in conjunction to obtain more accurate and useful predictions. Bagging is given an extra degree of arbitrariness using these procedures. Otherwise, the Random-Forest algorithm generates an arbitrary subset of predictors, preferably from the node where the trees are separated.

3.3.3 Support Vector Machine (SVM)

Support Vector Machine, or SVM for short, is a supervised learning technique based on linear classification. SVM works well for a wide range of issues and can solve both linear and non-linear problems. SVM outperforms all other classification techniques when it comes to appropriately solving regression and classification problems. Otherwise, the Support Vector Machine algorithm attempts to divide a dataset into two classes by passing a linearly section of hyperplane through it.

3.3.4 Logistic Regression (LR)

LR stands for Logistic Regression, which is currently widely employed in biological research and applications. The algorithm Logistic Regression (LR) is one of the most widely used machine learning methods when the target variable is categorical. In today's world, LR is one of the most effective strategies for solving any binary classification problem. In addition, it has shown a binary product between 0 and 1.

3.3.5 Gaussian Naïve Bayes (GNB)

GNB stands for Gaussian Naive Bayes classifier, which is a basic but effective classification technique. Gaussian Naive Bayes is a statistical procedure that works with training datasets to make a conditional independence speculation. The Gaussian Naive Bayes classifier, on the other hand, is an acceptable class strategy for determining a first-class answer for a dataset from a pool of different things.

3.4 Data Collection Procedure and Description of dataset

In this work, we take the diabetes disease data from Islami Bank Hospital and Diagnostic center, Dhaka, Bangladesh. Data collection is also a challenging task. At first, we are talking in hospital administration office to take a permission for collecting data. Then they do not take permission. Then we wrote an application to our honorable head for recommendation letter. Then our application is granted. Then, we again go to the hospital administration office and they give us a permission. Then, we collect the data.

485 patient data are included in our dataset and the dataset consist on many attributes. We have summarized the attributes and corresponding values in table 3.4.1

Table 3.4.1- Description of dataset

Serial No.	Attributes	Type	Values
1	Age (Years)	Numeric	(1 to 100)
2	Gender	Nominal	(Male, Female)
3	Glucose Level	Numeric	(3 to 40)
4	Blood pressure	Nominal	(High, Normal, Low)
5	Serum Uric Acid	Numeric	(1 to 15)
6	Serum Creatinine	Numeric	(0 to 4)
7	Smoking	Nominal	(Yes, No)
8	Insulin Result	Numeric	(50 to 550)
9	BMI level	Numeric	(15 to 50)
10	Diabetes Pedigree	Nominal	(Yes, No)
11	Polyuria	Nominal	(Yes, No)
12	Polydipsia	Nominal	(Yes, No)
13	Weight Loss	Nominal	(Yes, No)
14	Polyphagia	Nominal	(Yes, No)
15	Visual blurring	Nominal	(Yes, No)
16	Skin problem	Nominal	(Yes, No)
17	Feel's weakness	Nominal	(Yes, No)
18	Itching	Nominal	(Yes, No)
19	Irritate	Nominal	(Yes, No)
20	Class	Nominal	(True, False)

3.5 Implementation Requirement

For acquiring our goal, we have followed some steps to do our research.

Fig 3.5.1 Shows the research procedure.

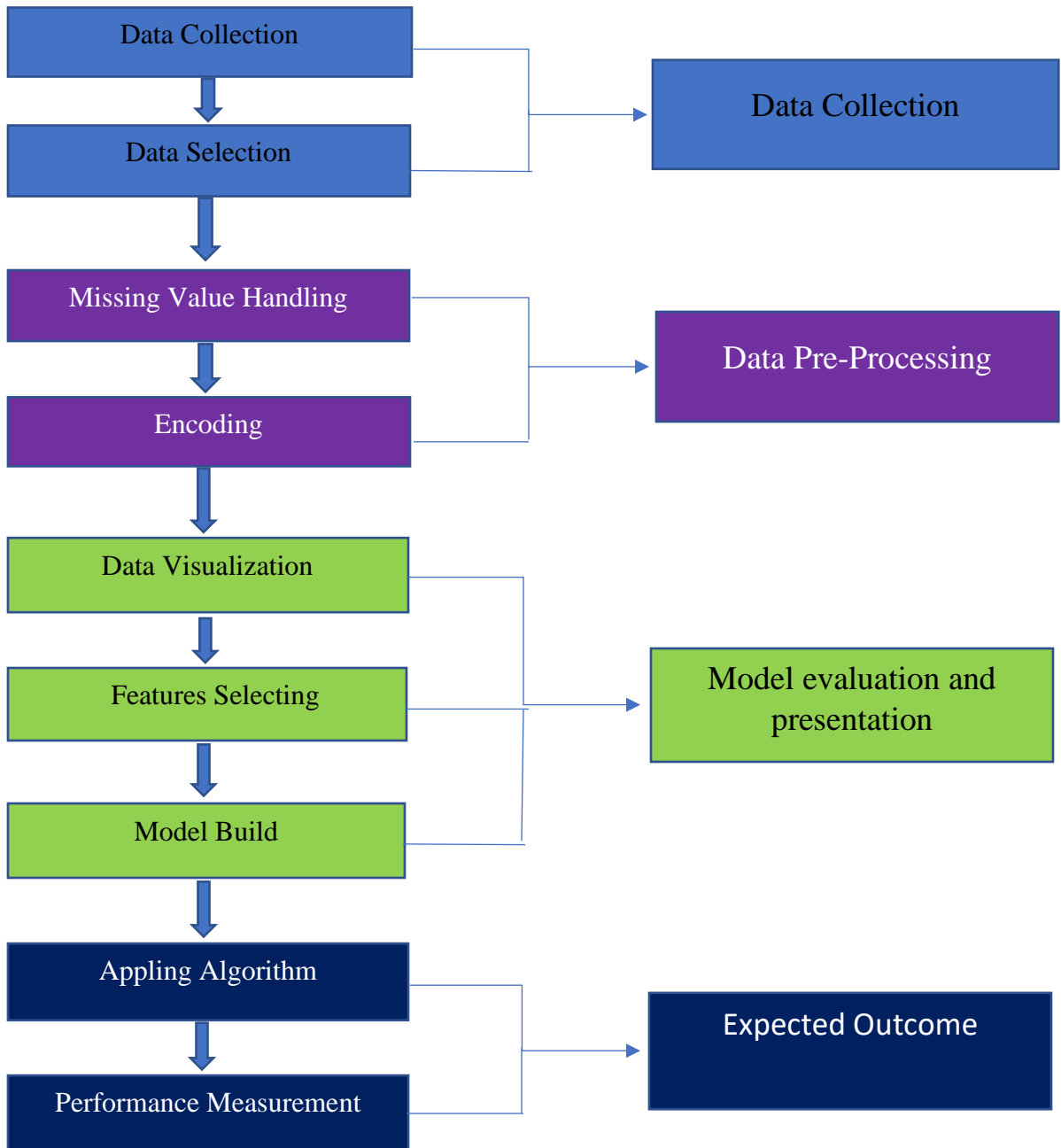


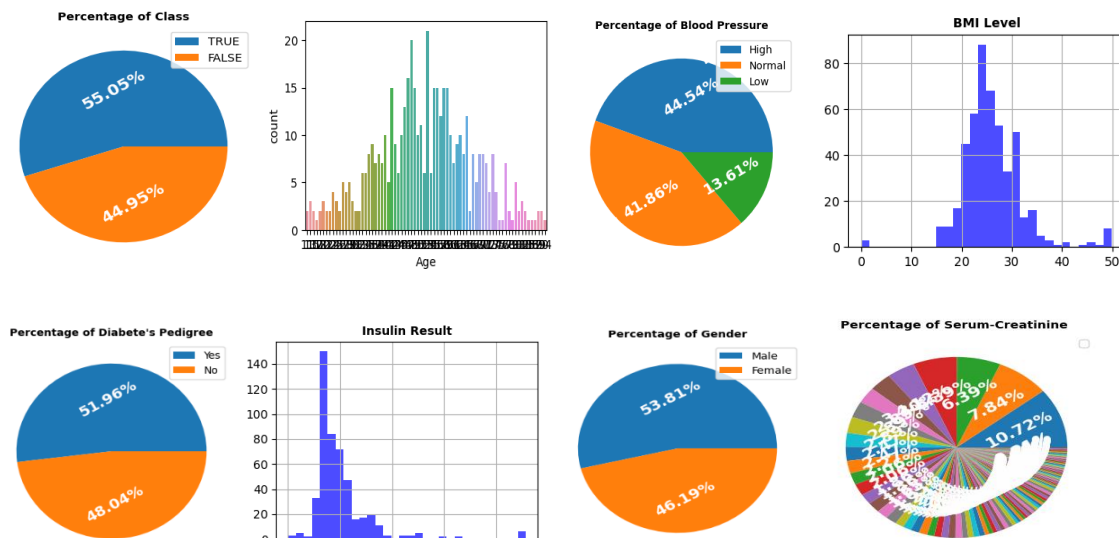
Figure 3.5.1- Research Procedure

3.6 Data Visualization

First of all, we try to find out the data in our data shape and also find out numbers of row and columns. There we find that 485 rows and 20 columns. Then we will find out the data type of our data. Then we are checking the duplicate row. Three duplicate rows are here. We will visualize overall scenario of our dataset. At first, we visualized the data for finding missing values. Then we visualized individually every column, for finding the number of True and False class.

We will visualize each and every column data percentage and number of data individually. In our figure, we show that percentage of our True and False in between total class column. The number of True data is 267 and False data is 218. The percentage of True data is 55.05% and the restore data is False.

In figure, 3.6.1 shows the data visualization



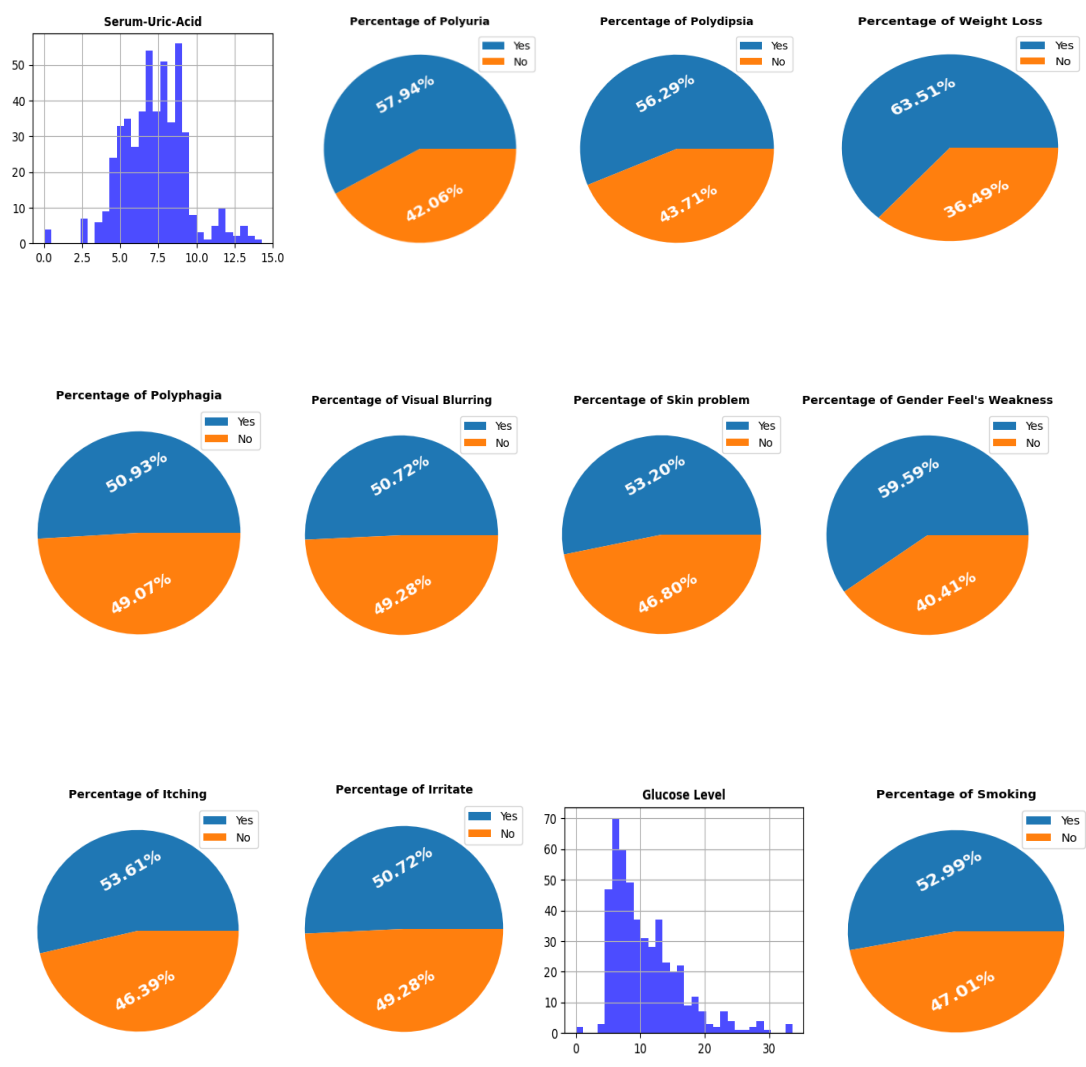


Figure 3.6.1- data visualization

3.7 Data pre-processing

In preprocessing, at first, we handle the missing values in our dataset. We check missing value in which column. Then we have seen that three columns have missing values. Which is glucose level, serum-creatinine and insulin result column. Then we handle this column. We handle the missing values in aspect of missing procedure. Then we find the mean value of individual missing columns. We will fill up the missing portion according to this column mean value.

We used the label encoder to transform data from string to nominal to numeric. After encoded the data we applied the histogram of data visualization.

In figure 3.7.1, we show that visualization after preprocessing.

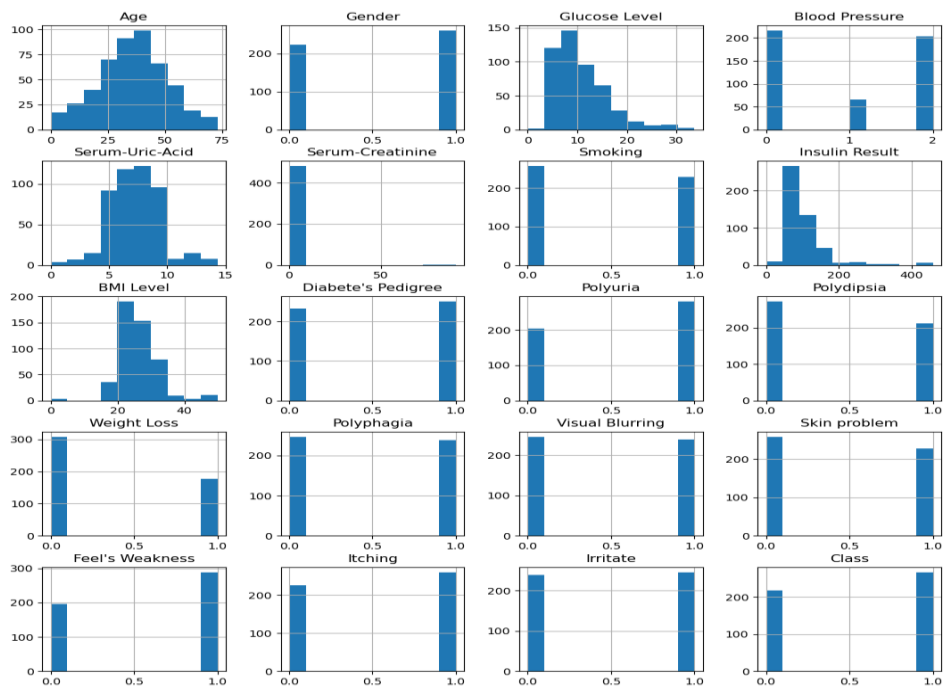


Figure-3.7.1-Visualization after Preprocessing

3.8 Features Selection

Features selection from the preprocessing dataset is a very important task. It helps a lot to implement our algorithm. In our algorithm we are working on 20 features.

Here we divided our dataset into two labels. Class column is a dependent label and the rest of the column are independent label. We are working on aspect of this type of feature selection.

3.8.1 Feature Extraction

In features extraction, we are working only independent features. In our research work without class column rest of the column in independent features. From each independent feature we are selecting 13 features as a top feature.

We find the top features in terms of dependent features. From sklearn.feature_selection we used True object which is selectKBest and f_classif to calculate the top features from existing features.

After selecting top features than we split our dataset into training and test. Now, we show the X-train blew.

In figure, the top 13 features position and extraction which is evaluated by sklearn library is given below figure 3.8.1

Table 3.8.1 Shows Top Features Selection Wise X-train Value

NR	GL	BP	SU	SM	IR	BL	DP	PL	PD	WL	PH	FW	IT
290	6.96	2	7.300	0	234.0	23.20	0	0	1	0	0	0	0
39	8.58	2	5.050	0	122.0	23.66	1	0	0	0	1	0	0
347	6.17	0	7.800	1	152.0	25.90	1	1	0	0	1	0	0
224	9.70	1	4.600	1	79.0	22.20	0	1	0	0	0	0	0
305	33.60	0	12.900	1	65.0	50.90	1	0	0	0	0	1	1
....
255	5.35	0	0.000	1	89.0	21.30	1	0	1	0	1	0	0
72	6.34	2	6.021	0	86.0	22.36	0	0	0	0	0	1	0
396	13.79	2	7.120	0	120.0	25.90	0	1	0	0	1	1	0
235	5.28	1	5.050	0	80.0	21.39	0	0	0	1	0	0	1
37	8.80	2	9.700	1	170.0	24.77	0	0	0	0	0	0	1

339 rows * 13 columns

Where,

NR – Number of Row, GL- Glucose Level, BP- Blood Pressure, SU- Serum-Uric-Acid, Sm- Smoking,

IR- Insulin Result, BL- BMI Level, DP- Diabetes Pedigree, PL- Polyuria, PD – Polydipsia,

WL- Weight Loss, PH- Polyphagia, FW- Feel's Weakness, IT- Irritate.

3.9 Build Model

For building our model, we have selected 13 top features from our existing 19 independent features. In top features of included polyuria, glucose level, serum uric-acid, blood pressure, irritate, feel weakness, polydipsia, weight loss, BMI level, polyphagia, diabetes pedigree, insulin result, smoking according to their values respectively. Then we build up our final model. To applied some Machine Learning algorithms for diabetes disease prediction. There are Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and Gaussian Naïve Bayes algorithms included. It is the important phase in our work.

Algorithm: Diabetes Prediction using pipeline

Step 1: Included required libraries

Step 2: Included in our real life diabetes datasets

Step 3: Taking missing our dataset

Step 4: Handling missing value in our dataset

Step 5: Encoded our nominal data and pre-processed our data.

Step 6: Visualized the pre-processed data

Step 7: Features scaling and labeling

Step 8: Fit the data for training dataset.

Step 9: Calculated the accuracy

Step 10: Identify the most accurate model based on test dataset.

We have followed this step to build up our model and find most accurate model of diabetes prediction.

3.10 Applying Algorithm

We have used 485 data in our research work. We divided the data in between training set and testing of our dataset. We have included 70% of data for training set and rest of the data for test set. Then applying our dataset based on some best machine learning algorithm.

Our research in based on supervised learning. So, we have used some classification algorithms there are various types of the tools available for using the algorithm and the procedures are different, in our sklearn library become different from other.

Different algorithm shows different result. So, selecting the proper algorithm is also a several task.

We have implemented Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and Gaussian Naive Bayes algorithm. The using these algorithms, we have used the classifier model.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

Diabetes is also common persistent diseases that is create a various problem to human health. Basically, the diabetes characteristic by high blood sugar then the ordinary degree. It is as result of faulty insulin selection. Other then it has created a biological impact of human health. Diabetes can damage the various human organ and tissues. Heart, kidney, eye, blood vessels and nerves also are damage this disease. Diabetes are also 3 parts that is type-1 (T1D), type-2 (T2D) and type-3 (T3D). If patient with type-1 (T1D) are general and normal less than 30 year. And patient with type-2 (T2D) is called the medical language for high blood sugar. And Type-3 are also called it. This type of diabetes can't be successful with the medicinal drugs and every patient are required insulin therapy. Kind this type diabetes is more harmful to our carrier at age human being. So, It's frequently associated the prevalence of obesity, hypertension and any diseases.

4.2 Measurement of Classification Techniques

It is the final step of prediction model here we evaluated the prediction result using various evaluation metrics like classification accuracy, confusion matrix and f1 score.

Accuracy formula: The proportion of True positive and True negative predictions to the total number of predictions. It is as follows: –

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

Here, TP= True Positive, TN= True Negative, FP= False Positive and FN= False Negative.

Confusion Matrix Diagram: A confusion matrix is concise of predict result on a classification on our dataset values. We use a confusion matrix to evaluate the resulting model and describes the complete performance of the model. The model given below:

Actual class\Predicted class	C1	\neg C1
C1	TP	FN
\neg C1	FP	TN

Where, TP: True Positive

FP: False Positive

FN: False Negative

TN: True Negative

Accuracy: Accuracy is the measurement of performance based on our confusion matrix values in true positive and false negative in divided by the total number of prediction values.

$$\text{Accuracy} = \frac{TP+TN}{ALL}$$

Where, ALL: Total number of samples

Precision: It is the % of instances that the classifier predicted as positive that are actually positive based on confusion matrix values. In this way we calculate our precision result. Precision mathematical form it is given below-

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

Recall: It is the % of positive instances that the classifier predicted correctly as positive based on confusion matrix values. In this way we calculate our precision result. Recalls mathematical form it is given as-

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

F1 score: We used the F1 score to assess the accuracy of a test on a dataset. Although the perfect score for both is 1.0, there is frequently a trade-off between Precision and Recall. The F1 score is a harmonic mean of precision and recall. The mathematical form of the F1 score is as follows:

$$F1 \text{ score} = 2 * \frac{(\text{precision} * \text{recall})}{(\text{precision}) + (\text{recall})}$$

4.3 Analysis of The Results

In this section, we have use various classification techniques to measure the most popular 5 machines learning classification algorithm for diabetes prediction.

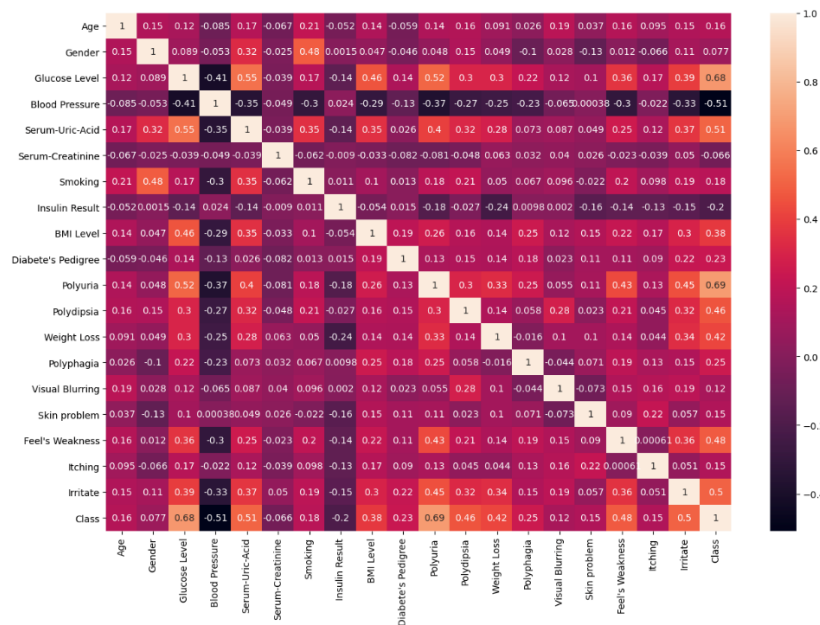


Figure 4.3.1- Heat map for correlated columns

From our dataset 267 true values and 218 false samples taken for analysis. We divided the dataset into 2 portion of dataset, where the training set contains 70% and test set contains remaining 30% of data. So, the dataset has been checked verify the co-related features in order to drop the redundant columns. The heat map shows in figure 4.3.1 appear to have no correlated column.

In figure 4.3.2 shows the performance measurement of 5 Supervised Machine Learning Algorithm for diabetes prediction. Here, RF and LR outperformed the other classification algorithms. In terms of accuracy acquiring the highest accuracy as 97% and 95% respectively. However, the support vector machine exhibits the lowest performance than the other algorithms.

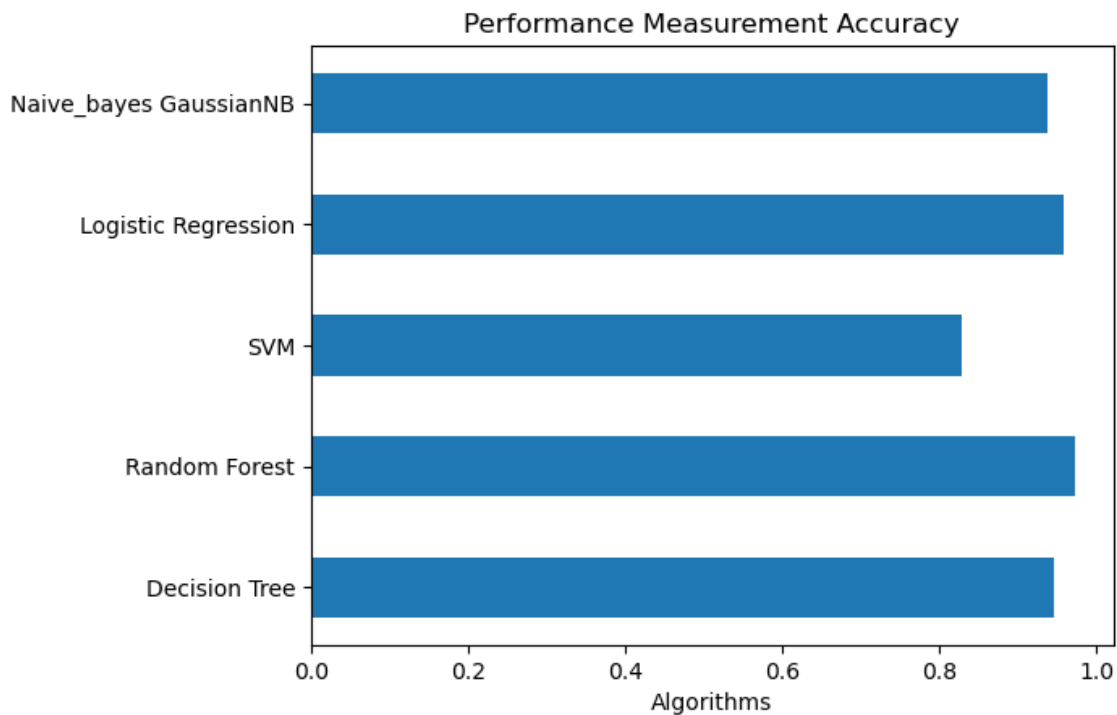


Figure- 4.3.2- Performance Measurement Accuracy

4.4 Performance Evaluation

In this portion, we measurement the performance of the prediction model. In below we have discussed our 5 algorithms accuracy, precision, recall and F1 score.

In decision tree algorithms we find out the accuracy, precision, recall, and F1 score in terms of confusion matrix. In this algorithm the accuracy is 0.93, precision is 0.92, recall is 0.92 and the f1 score is 0.92. In figure 4.4.1 shows the decision tree confusion matrix.

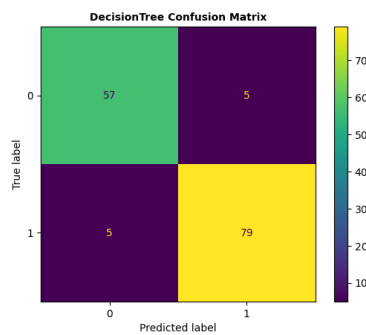


Figure 4.4.1- Decision Tree Confusion Matrix

In random forest algorithms we calculate the accuracy, precision, recall, and F1 score in terms of confusion matrix. In this algorithm the accuracy is 0.97, precision is 0.94, recall is 0.98 and the f1 score is 0.96. In figure 4.4.2 Shows the random forest confusion matrix.

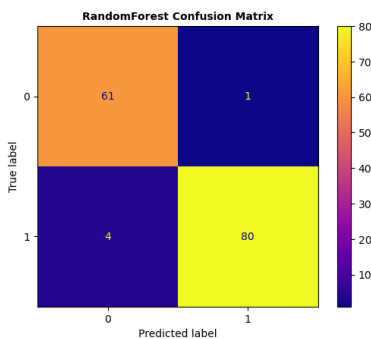


Figure 4.4.2- Random Forest Confusion Matrix

In support vector machine algorithms, we find out the accuracy, precision, recall, and F1 score in terms of confusion matrix. In this algorithm the accuracy is 0.82, precision is 0.76, recall is 0.87 and the f1 score is 0.81. In figure 4.4.3 shows the support vector machine confusion matrix.

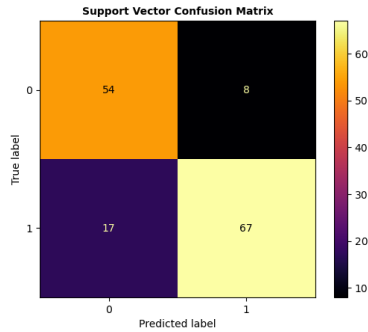


Figure 4.4.3- Support Vector Machine Confusion Matrix

In logistic regression algorithms we find out the accuracy, precision, recall, and F1 score in terms of confusion matrix. In this algorithm the accuracy is 0.95, precision is 0.94, recall is 0.97 and the f1 score is 0.95. In figure 4.4.4 shows the logistic regression confusion matrix.

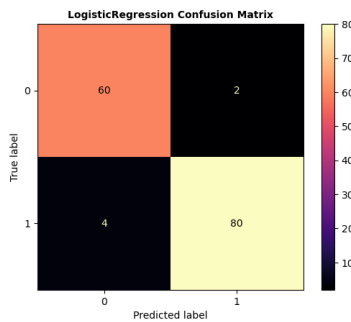


Figure 4.4.4- logistic regression Confusion Matrix

In Gaussian Naïve Bayes algorithms, we find out the accuracy, precision, recall, and F1 score in terms of confusion matrix. In this algorithm the accuracy is 0.93, precision is 0.92, recall is 0.94 and the f1 score is 0.93. In figure 4.4.5 shows the Gaussian naïve Bayes confusion matrix.

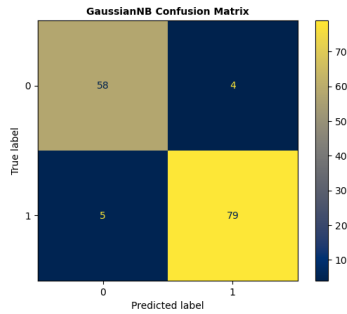


Figure 4.4.5- Gaussian Naïve Bayes Confusion Matrix

Now, we constructed a table for above description.

Table 4.4.6 shows the classification performance measurement

Table 4.4.6 Classification performance Measurement

Measurement Techniques	DT	RF	SVM	LR	GNB
Accuracy	0.93	0.97	0.82	0.95	0.93
Precision	0.92	0.94	0.76	0.94	0.92
Recall	0.92	0.98	0.87	0.97	0.94
F1	0.92	0.96	0.81	0.95	0.93

4.5 Receiver Operating Curve

As the receiver operating curve show our model distribution our classes. The according to our five algorithms value of AUC values represent of all figure. A useful device whilst predicting the opportunity of a binary outcome is the Receiver working feature curve, or ROC curve. The model's overall success is also shown by the micro and macro averages. As a result, the micro-average ROC adds up each true positive, false positive, and false negative value and plots it on a graph. Where macro-average uses the average of precision and recall to map a value on a graph. The micro-average ROC, on the other hand the dataset is severely skewed, it is taken into account. The Balanced Dataset Result: There are a variety of approaches for balancing data sets. Balanced datasets, for example, penalized Models and Anomaly Detection.

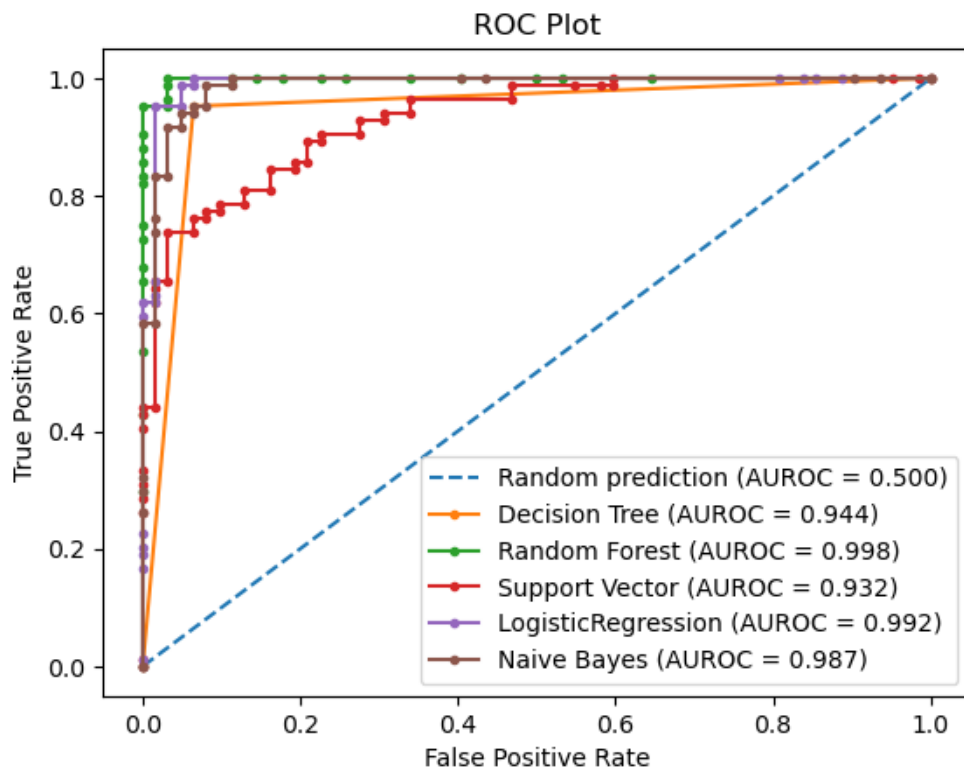


Figure-4.5.1- ROC Plot

4.6 Experimental Results

To acquire expected our aim, we have need here 5 different algorithms. After applying the algorithm then we evaluated the models. Then we find that Random Forest algorithm gives us the best accuracy which is around 97%.

In table 4.6.1 represents the accuracy of all Model Performances.

Table- 4.6.1- Performance Measurement Accuracy

Number	Algorithm	Result/Accuracy
1	Decision Tree	92%
2	Random Forest	97%
3	Support Vector machine	82%
4	Logistic Regression	95%
5	Gaussian Naive Bayes	93%

4.7 Descriptive Analysis

After measuring the overall performances, we find that the Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and Gaussian Naïve Bayes accuracy is 92%, 97%, 82%, 95% and 93% respectively. So, we say that the best accuracy in Random Forest Algorithm and accuracy is 97%. The lowest accuracy in Support Vector Machine and accuracy in 82%.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

According to our report, in this section we have explain about important of our work aspects and describe properly for our research work that are the concise summary of research, conclusion, recommendation, implication of the future work. Basically, step by step of all our work describe given below:

5.1 Summary of the Study

Many diseases have most detrimental for our health. Now a day's, diabetes is one of the most one. Diabetes is called the root of all other diseases. In this diabetes diseases that occurs in our high blood sugar. Blood sugar is source of all energy and comes from eating food. Blindness is one of the most cause of diabetes. A hormone called by insulin made by the pancreas collect glucose from our eating food get into our cells to be used for energy. Now a day's most of the people in the world are affected by this type of diseases. In this cause, a large amount of patient is suffering from blindness. Sometimes our body doesn't make insulin properly. In this cause, our glucose level very fast high in our body and make health problem. Diabetes has no cure, follow some rule's that manage your diabetes and stay safe.

There are some types of diabetes that type (1, 2) and gestational diabetes. For age of most affected people between 60-65. About 90-95% people of cause of adult people affected by type 2 diabetes.

In other that many sides problem in high blood pressure. That is heart diseases, stroke, kidney diseases, eye problem and lower limb etc.

5.2 Conclusion

Diabetes has no cure but following some rules that come back easily. Exercise is one of them. Exercise is reduced of severity for diabetes and every similar disease. The long-term complications of diabetes.

Is reduce for exercise. In it's a well-planned and regular exercise for help to reduce our diabetes. In our life if made a part and parcel of our everyday life that helps our good health. Exercise is helped the controlling the blood sugar levels without any medicine. There is no blood sugar but also control our weight and blood pressure as well as our exercise level. And lowers the bad cholesterol also. Exercise can reduce the heart diseases and never damage the risk of kidney failure.

Headaches from diabetes can range. But, the most familiar co-morbid situations consist of amputations, cardiovascular sickness, obesity, high blood pressure, hypoglycemia, dyslipidemia, and risk of coronary heart attack or stroke.

The type of diabetes, gestational diabetes, consequences pregnant girls and is characterized via raised blood sugar.

5.3 Recommendations

Even though we declaration some related research but its amount is only a few and properly there has no work like us primarily based on Bangladeshi statistics. We have Understand all their research system and work style after that we started to restore our research aim. After a tough effort with the aid of doing all of the work little by little ultimately, we're at a level what may be stated it is our anticipated studies goal. So, for making this kind of studies work it need a remarkable work for directing us via the right path of studies.

We've experienced some one-of-a-kind troubles that were inside the beginning of our studies. We've also stuck with the getting to know of the large area of facts mining and device mastering. With the entire journey if this studies paintings our manager Md. Zahid Hasan sir helped us a lot and guided us for making this Studies challenge successful.

5.4 Implication for Further Study

Diabetes is measurement of blood glucose level that microvascular and cardiovascular complications. If we lead a quality life that to reduce our diabetes that is harm him. It's helps for insulin. Type 1 diabetes is brief reliance on insulin for survival and comprises 10% of all cause of diabetes. In type 2 is more prevalent from a diabetes that 90% affected of all people with diabetes. In that type insulin are not make our health. In cause our sugar level grew up an affiliation among the complications of diabetes and accelerated blood glucose levels become postulated in the early a part of this century.

We hope our research serves to help that create a better life for this impacted by diabetes. Our studies also help how to investigational and broken down and diabetes prediction and remove from the body in this disease. Off all type from diabetes. We recognize that running toward extra complete remedy for diabetes will enable researchers to enhance the lifestyles of the thousands and thousands of folks that struggle with diabetes each day. Type 1 diabetes is a persistent situation characterized with the aid of a loss of insulin production. This kind of diabetes is regularly existence-long, and cannot be avoided with the aid of lifestyle alternatives or workout habits.

REFERENCES

- [1] Othmane Daanouni, Bouchaib Cherradi and Amal Tmiri “ Predicting Diabetes Diseases Using Mixed Data and Supervised Machine Learning Algorithms” Proceedings of the 4th International Conference on Smart City Application Article No.: 85 Pages 1–6, October 2019,
- [2] Amani Yahyaoui, Akhtar Jamil, Jawad Rasheed and Mirsat Yesiltepe e “A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques.” Proceedings of the 2019 1st International Informatics and Software Engineering Conference (UBMYK), November 2019, <https://ieeexplore.ieee.org/document/8965556>
- [3] Ms. K Sowjanya, Dr. Ayush Singhal and Ms. Chaitali Choudhary “MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices.” Proceedings of the 2015 IEEE International Advance Computing Conference (IACC), June 2015, <https://ieeexplore.ieee.org/document/7154738>
- [4] Kucharlapati Manoj Varma and Dr B S Panda “Comparative analysis of Predicting Diabetes Using Machine Learning Techniques.” June 2019 https://www.researchgate.net/publication/338402143_Comparative_analysis_of_Predicting_Diabetes_Using_Machine_Learning_Techniques
- [5] S M Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed, Sheak Rashed Haider Noori and Md Nazirul Islam Sarkar” Machine Learning Based Unified Framework for Diabetes Prediction.” 2018 International Conference on Big Data Engineering and Technology. August 2018, https://www.researchgate.net/publication/334988298_Machine_Learning_Based_Unified_Framework_for_Diabetes_Prediction
- [6] Badiuzzaman Pranto, Sk. Maliha Mehnaz, Esha Bintee Mahid, Imran Mahmud Sadman, Ahsanur Rahman and Sifat Momen “Evaluating Machine Learning Methods for

Predicting Diabetes among Female Patients in Bangladesh” July 2020
<https://www.mdpi.com/2078-2489/11/8/374>

[7] M. Tech. Scholar Arvind Aada and Prof. Sakshi Tiwari “Predicting Diabetes in Medical Datasets Using Machine Learning Techniques. ” International Journal of Scientific Research & Engineering Trends Volume 5, Issue 2, Mar-Apr-2019, <https://www.ijser.org/researchpaper/Predicting-Diabetes-in-Medical-Datasets-Using-Machine-Learning-Techniques.pdf>

[8] D. Vigneswari, N. Komal Kumar, V. Ganesh Raj, A. Gugan and S. R. Vikash. “Machine Learning Tree Classifiers in Predicting Diabetes Mellitus. ” 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), June 2019, https://www.researchgate.net/publication/333651256_Machine_Learning_Tree_Classifiers_in_Predicting_Diabetes_Mellitus

[9] Debadri Dutta, Debpriyo Paul and Parthajeet Ghosh “Analysing Feature Importances for Diabetes Prediction using Machine Learning” 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), November 2018, <https://ieeexplore.ieee.org/abstract/document/8614871>

[10] Lejla Alic, Hasan T. Abbas, Marelyn Rios, Muhammad AbdulGhani, and Khalid Qaraqe ” Predicting Diabetes in Healthy Population through Machine Learning.” 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), June 2019, <https://ieeexplore.ieee.org/document/8787404>

[11] Md. Maniruzzaman, Md. Jahanur Rahman, Benojir Ahammed and Md. Menhazul Abedin “Classification and prediction of diabetes disease using machine learning paradigm.” Article in Health Information Science and Systems · January 2020 <https://link.springer.com/article/10.1007/s13755-019-0095-z>

[12] Aishwarya Mujumdar and Dr. Vaidehi V “Diabetes Prediction using Machine Learning Algorithms.” 2019 INTERNATIONAL CONFERENCE ON RECENT

TRENDS IN ADVANCED COMPUTING , Volume 165, 2019, Pages 292-299,
November 2019 <https://www.sciencedirect.com/science/article/pii/S1877050920300557>

[13] S.Saru and S.Subashree “ ANALYSIS AND PREDICTION OF DIABETES USING
MACHINE LEARNING.” International Journal of Emerging Technology and Innovative
Engineering Volume 5, Issue 4, April 2019
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3368308

Appendix





Thesis Project Name: Diabetes Prediction

Dept.: Computer Science and Engineering

Supervised By:


Md. Zahid Hasan, Assistant. Professor, DIU

We are the student of Daffodil International University in department of Computer Science and Engineering (CSE). We are working for research on Diabetes patient using Machine Learning. So, we have to need some information in diabetes patient test report. This information is needed as follows:

- Age
- Gender
- Glucose
- Blood Pressure
- Serum-Creatinine
- Serum-Sodium
- smoking
- Skin Thickness
- Insulin
- BMI
- Diabetes Pedigree Function
- Polyuria
- Polydipsia
- Sudden weight loss
- Weakness
- Polyphagia
- Genital Thrush
- Visual Blurring
- Itching
- Irritability
- Delayed healing
- Partial Paresis
- Muscle stiffness
- Alopecia
- Obesity
- Class

Student Info:

1. Sobuj Mia- 181-15-10699
2. Razib Hasan- 181-15-10786
3. Md. Juweel Rana- 181-15-10919


31/10/2021
Dr. Sheek Haidar Noor
Associate Professor & Associate Head
Department of CSE, DIU
Daffodil International University



Thesis Project Name: Diabetes Prediction

Dept.: Computer Science and Engineering

Supervised By:

Md. Zahid Hasan, Assistant. Professor, DIU

Diabetes prediction questionnaire:

- What is the patient's age?
- What is the patient's gender?
- How much the patient's glucose level?
- How much the patient's blood pressure?
- How much the patient's Serum-Creatinine?
- How much the patient's Serum-Sodium?
- Do the Patient's the smoking?
- What is the patient's insulin result?
- How much the patient's BMI level?
- Is there diabetes pedigree in your family?
- Is the patient's suffering in polyuria?
- Is the patient's suffering in polydipsia?
- How much patient weight loss?
- Is the patient's suffering in polyphagia?
- If the patient's feel visual blurring?
- Is your skin problem?
- Are you feel weakness?
- Are you suffering from itching?
- Are you feels irritate?
- Are you suffering from alopecia?

Student Info:

1. Sobuj Mia- 181-15-10699

2. Razib Hasan- 181-15-10786

3. Md. Juweel Rana- 181-15-10919


31.10.2021
Dr. Sheak Rashed Haider Noor
Associate Professor & Associate Head
Department of CSE, DIU
Daffodil International University

০২-১১-২০২০
বঙ্গাব্দ

সুশাসনিক

স্বাধীনতা সংগ্রাম সংগঠন
মতিমিন, ঢাকা।

ক্রিয়: জাতিবৈজ্ঞানিক রোগীর তথ্য দিয়ে আর্থিক আবেদন।

জনাব,

যথাযথ অস্বাস্থ্যকর পরিস্থিতিতে বিনীত নিবেদন, যেহেতু
আমরা জাতিবৈজ্ঞানিক আন্তর্জাতিক বিশ্ববিদ্যালয়ের দুর্ভাগ্যবশত হওয়া
আমরা জাতিবৈজ্ঞানিক রোগের উপর বিশ্ববিদ্যালয় আবেদন
পূর্বক গবেষণা করছি। এরপরও, আমাদের জাতিবৈজ্ঞানিক
রোগীর ক্ষিপ্র তথ্য প্রয়োজন।

অতএব, জনাবের নিকট আবেদন যে, আমাদের
জাতিবৈজ্ঞানিক রোগের প্রয়োজনীয় তথ্য-দিয়ে আর্থিক
করত সুমতি হন।

বিনীত নিবেদন,

- ১। ডাঃ জুয়েল রানা - 181-15-10717
- ২। অরুণ সিন্ধা - 181-15-10677
- ৩। রাফিক হোসেন - 181-15-10786

জাতিবৈজ্ঞানিক আন্তর্জাতিক বিশ্ববিদ্যালয়
গ্রেড কাম্পাস, ঢাকা।

কি কি বিষয়ে ডাক্তার
এই প্রশ্নের উত্তরে
এই জাতিবৈজ্ঞানিক রোগের
স্বাধীনতা সংগ্রাম সংগঠন
এই জাতিবৈজ্ঞানিক রোগের
স্বাধীনতা সংগ্রাম সংগঠন
এই জাতিবৈজ্ঞানিক রোগের
স্বাধীনতা সংগ্রাম সংগঠন
এই জাতিবৈজ্ঞানিক রোগের
স্বাধীনতা সংগ্রাম সংগঠন

স্বাধীনতা
স্বাধীনতা

Plagiarism Checker Screenshot

Report

ORIGINALITY REPORT

24% SIMILARITY INDEX	17% INTERNET SOURCES	18% PUBLICATIONS	11% STUDENT PAPERS
--------------------------------	--------------------------------	----------------------------	------------------------------

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	S. M. Hasan Mahmud, Md Altab Hossin, Md. Razu Ahmed, Sheak Rashed Haider Noori, Md Nazirul Islam Sarkar. "Machine Learning Based Unified Framework for Diabetes Prediction", Proceedings of the 2018 International Conference on Big Data Engineering and Technology - BDET 2018, 2018 Publication	3%
3	F J Damanik, D B Setyohadi. "Analysis Of Public Sentiment About Covid-19 In Indonesia On Twitter Using Multinomial Naive Bayes And Support Vector Machine", IOP Conference Series: Earth and Environmental Science, 2021 Publication	1%
4	nebula.wsimg.com Internet Source	1%

www.doria.fi