JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

# Diabetes Prediction using Machine Learning Techniques

**Gorisetty Nirosha**
Dept. of Computer Science and Engineering
Specialization in Information Technology
Andhra University College of Engineering
Visakhapatnam, Andhra Pradesh, India

**Dr. G.Sharmila Sujatha**
Dept. of Computer Science and Engineering
Andhra University College of Engineering
Visakhapatnam, Andhra Pradesh, India

***Abstract:-***Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat our body turns food into sugars, or glucose. At the point, one pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and Type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, various Machine Learning Techniques. Machine constructing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which are Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Logistic Regression (LR), GaussianNB (GNB). The accuracy is different for every model when compared to other models. The Project work gives the accurate of higher accuracy model shows that the model is capable of predicting diabetes effectively.

***Keywords*: *Diabetes, Machine, Learning, Prediction, Dataset, Ensemble***

## I. INTRODUCTION

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and soforth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particu- larly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabet- ics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it's tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

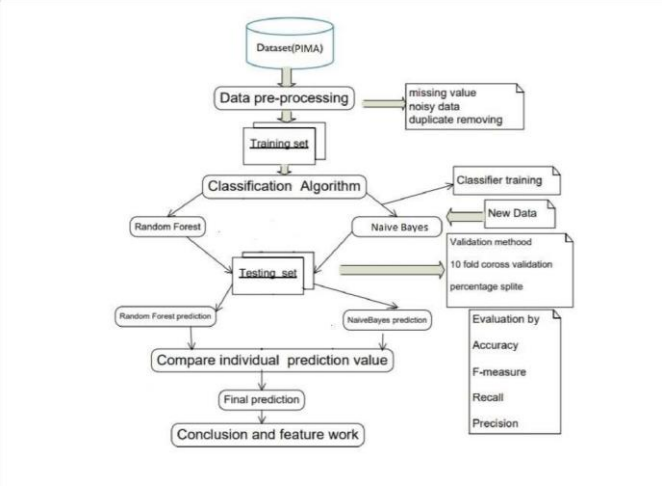
## II. LITERATURE REVIEW

K.VijiyaKumar et al. [11] proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in machine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al. [13] presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Tejas N. Joshi et al. [12] presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease. Deeraj Shetty et al. [15] proposed diabetes disease prediction using data mining assemble Intelligent Diabetes Disease Prediction System that gives analysis of diabetes malady utilizing diabetes patient's database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on diabetes patient's database and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar et al. [10] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algorithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much im- proved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research.

## III.SYSTEM DESIGN

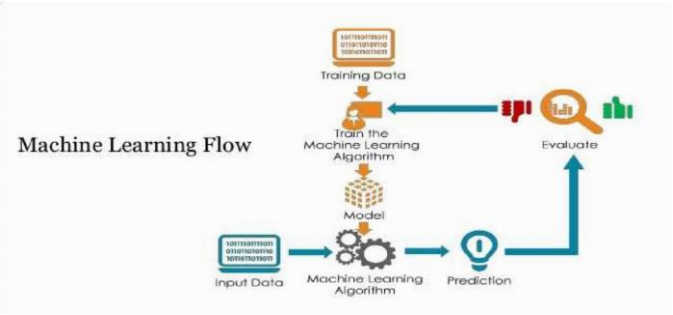



Proposed System Architecture

The proposed system architecture describes the workflow of the project we are working on. First, we procure the dataset, which is the PIMA Indian dataset. It is a dataset which is used mainly for diabetes prediction. The dataset contains up to 1000 rows and mainly depicts the features required for the prediction of diabetes.

We split the dataset into training and testing data where part of the dataset is trained and part of the dataset is used for testing. We train the dataset in order to find the accuracy of the percentage of people having and not having diabetes.

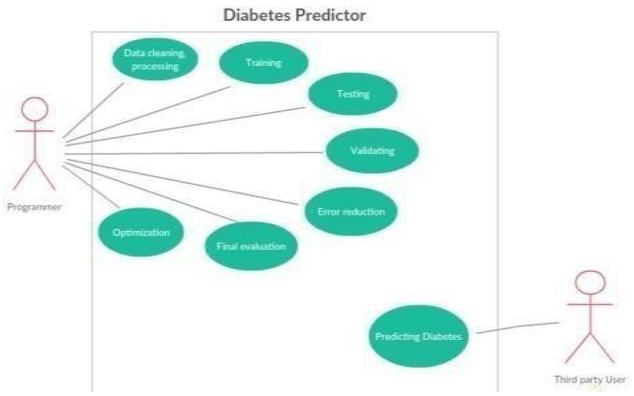
Many methods are used for the purpose of the prediction of diabetes such as Naïve Bayes, Random Forest, Logistic Regression, Decision Trees etc. We mainly focus on Naïve Bayes and Random Forest as these two are the most efficient in getting an efficient result for the prediction. We perform Naïve Bayes and Random Forest on the training and testing data and find the accuracy percentage of both the data for finding the best evaluation method among the two for the analysis of the dataset.




## IV.UML DIAGRAMS

**Use case diagram:**

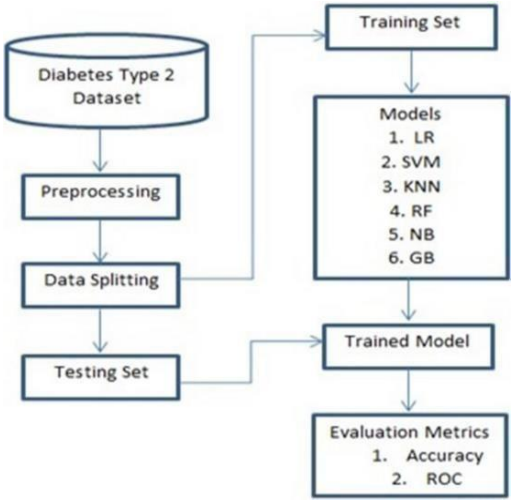
A use case diagram in the Unified Modelling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted. A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved. A use case diagram can identify the different types of users of a system and the different use cases and will often be accompanied by other types of diagrams as well. The use cases are represented by either circles or ellipses.



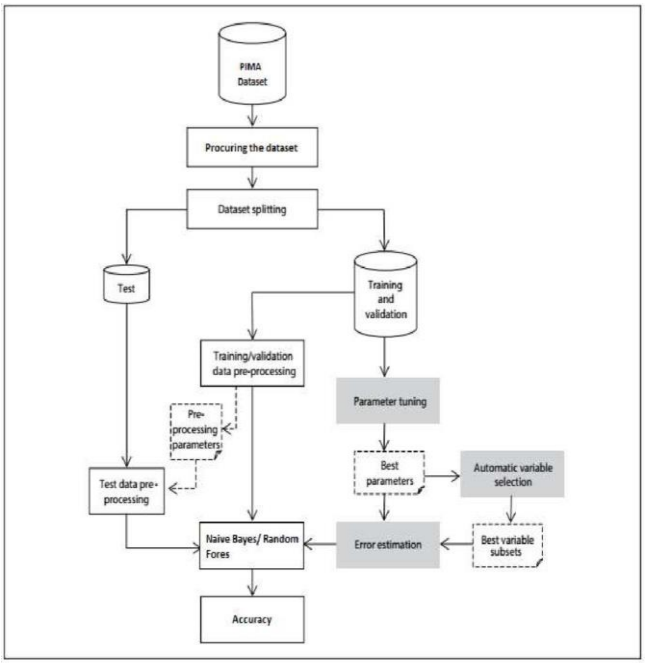

**Class diagram:**

In the design of a system, a number of classes are identified and grouped together in a class diagram that helps to determine the static relations between them. With detailed Modelling, the classes of the conceptual design are often split into a number of subclasses.

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023



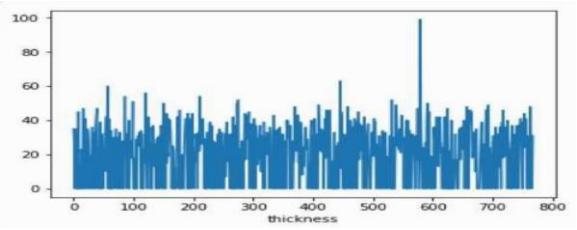

## V.PROPOSED METHODOLOGY



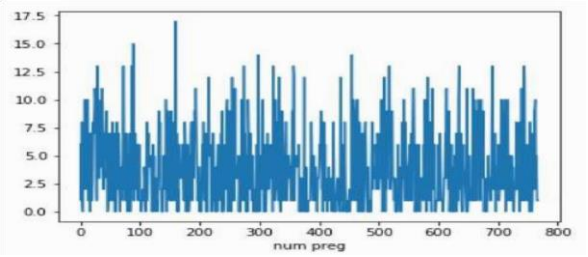

### PROCURING THE DATASET

The dataset used here is the PIMA Indian Dataset. It is the data obtained from the National Institute for Diabetes. It contains of several medical predictor variables and one target variable. The various medical variables are BMI, Glucose levels, Blood Pressure etc. It contains 768 rows and 9 columns. The columns that are present in the dataset are as follows:

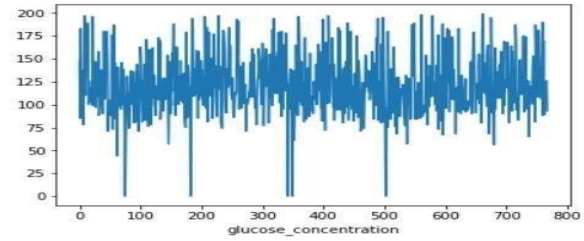
**Skin Thickness:** Skin thickness is a column in the dataset which denotes the thickness of an individual's skin. Skin thickness varies from person to person depending upon their health and various other factors which can affect the skin. A person's skin thickness can play a factor in denoting whether the person has diabetes or not, but in the dataset, there are a few rows where the skin thickness is set to 0. Skin thickness cannot be 0 for a person, so we try to avoid this column mainly to get the accurate results while performing prediction. While performing analysis, the skin thickness column is removed from the code we write so as to get a more accurate prediction result using Naïve Bayes and Random Forest.




**Number of Pregnancies:** When a woman gets pregnant, they may or may not go through gestational pregnancy. Gestational pregnancy is a common form of pregnancy where the woman develops diabetes. After the birth, the diabetes usually goes away. The diabetes is caused due to the high levels of sugar in the body which does not happen when the woman is not pregnant. This is due to the making of hormones by the placenta. The number of pregnancies plays a key factor when it comes to diabetes in women. So we record the number of pregnancies and if it is a male, the pregnancy is set to 0 in the dataset. It can also denote that a woman has not been pregnant during her life.



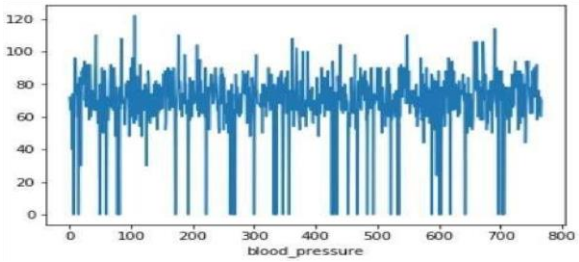

**Glucose concentration:** The glucose concentration is the level of glucose that is present in a person's blood. A teaspoon of glucose is required for a human body to function normally per day. The glucose present in the body travels through the bloodstream to other parts of the body. The glucose level is required to determine the amount of insulin present in the body. If the insulin is not able to handle the amount of glucose in the body, then this causes diabetes. The glucose levels in a person's body is an important factor in determining if the person has diabetes or not. In the dataset, we have a column to represent the glucose level of each person.



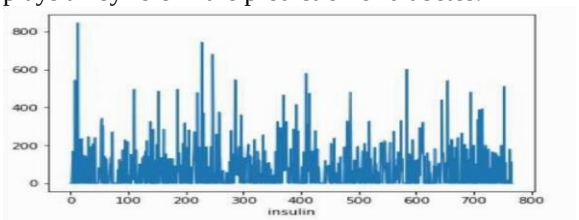

**Blood Pressure:** The blood in our body moves through our body by the means of blood pressure. It helps in the movement of oxygen and nutrients throughout our body through the blood. The white blood cells in our body are also delivered by the means of blood pressure. The normal blood pressure for a person is usually below 120 mm Hg systolic and 80 mm Hg diastolic. Variations in blood pressure can be a major cause of diabetes mellitus. So we take this factor in our

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

dataset for the prediction of diabetes.

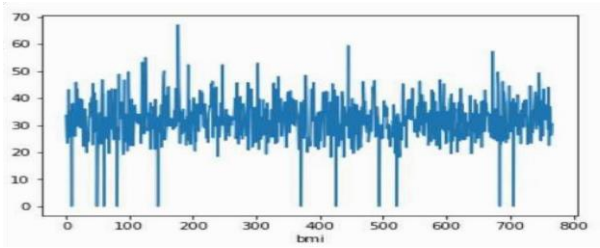



**Insulin:** To control the blood sugar in our body, insulin is required. Insulin is a hormone created by the pancreas to balance all the sugars in our body. The insulin controls the glucose concentration, which is a major factor in development of diabetes. If the insulin is not able to keep up with the levels of glucose in our body, it causes diabetes. Insulin also helps in the breaking down of fats and proteins in our body to form energy. Insulin resistance is the inability of insulin to exert its effects on the tissues in our body. In the dataset, the insulin level plays a key role in the prediction of diabetes.

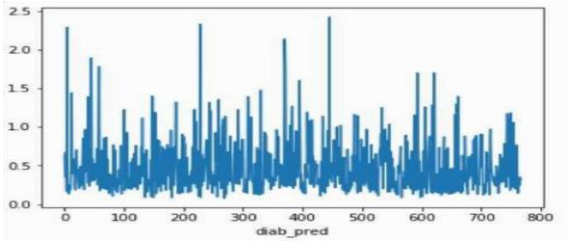



**Body Mass Index (BMI):**The Body Mass Index of a person can be defined as the person's weight divided by the square of the height. The weight is defined in kgs and the height is defined in meters. The BMI of a person varies according to the weight and height and it calculates whether the person is normal or obese. The following table denotes the various BMIs that distinguish a person into four categories:

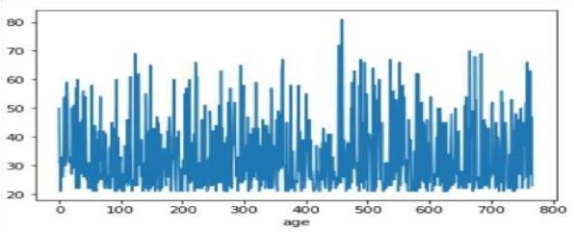
| BMI | Category |
|---|---|
| Under 18.5 kg/$m^2$ | Underweight |
| 18.5 to 25 | Normal Weight |
| 25 to 30 | Overweight |
| Over 30 | Obese |




**Diabetes Pedigree Function:**




**Age:** Age is a common factor for diabetes. When it comes to age, usually, people above the age of 40 are diagnosed with diabetes. But, sometimes, even people who are younger are diagnosed with diabetes. Type 1 diabetes usually occurs in people above the age of 40 but sometimes, people at ages as young as 15 – 16 can also be diagnosed. This all depends on factors such as family history, diet etc.




**Value of Diabetes Diseases:** In the dataset, this column is used to define if the person has diabetes or not. We define it using True or False. The dataset has predefined values for each person whether the person has diabetes or not and our project is to find whether the given values are accurate or not.

The following features are the key to finding whether a person has diabetes or not. There are various other factors as to determining diabetes, but in our project, we are mainly focusing on these features for the prediction.

The dataset file is in a .csv(Comma Separated Values) format. Using the help of Python's inbuilt library Pandas, which is a data frame library, we import the file into our Python environment. The other libraries that are imported into the environment are:

**Numpy** – a library that is used mainly to operate with large dimensional arrays and matrices, providing high level mathematical functionalities to work on data.

**Matplotlib** – the library that provides Python with the functionality of plotting graphs and plots. It works in tandem with NumPy. Pandas has a function named read_csv(), which essentially reads a file of the format (.csv). Once the dataset is loaded into the environment, we can check the dimensions of the dataset by the function .shape() which returns the number of rows and columns. Basic lookup of the data is done, by using the inbuilt commands .head() and .tail() which print the number of rows from the start of the dataset and the bottom of the dataset respectively.

| *Number of Features* | *Features* | *Descriptions and Features values* |
|---|---|---|
| *1* | *Number of times a person was pregnant* | *Numeric value* |
| *2* | *Glucose Concentration* | *Numeric value* |
| *3* | *Blood Pressure* | *Numeric value (in mm Hg)* |
| *4* | *Skin Thickness* | *Numeric value (in mm)* |
| *5* | *Insulin* | *Numeric value* |
| *6* | *Body Mass Index (BMI)* | *Numeric value (weight in kg/(height in m)$^2$)* |
| *7* | *Diabetes Pedigree Function* | *Numeric value* |
| *8* | *Age* | *Numeric value* |
| *9* | *Value of Diabetes Diseases* | *Yes = True No = False* |

Features of Pima Indians Diabetes for Diagnosing Diabetes Disease Type 2

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

| Number of Attributes | Attributes Name | Mean | Standard Deviation |
|---|---|---|---|
| 1 | Number of times a person was pregnant | 3.8 | 3.4 |
| 2 | Glucose Concentration | 120.9 | 32.0 |
| 3 | *Blood Pressure* | 69.1 | 19.4 |
| 4 | *Skin Thickness* | 20.5 | 16.0 |
| 5 | *Insulin* | 79.8 | 115.2 |
| 6 | *Body Mass Index (BMI)* | 32.0 | 7.9 |
| 7 | *Diabetes Pedigree Function* | 0.5 | 0.3 |
| 8 | *Age* | 33.2 | 11.8 |

Statistical Analysis for Mean and Standard Deviation in Pima Indians Diabetes Data set

After procuring the dataset, we see if we can make any changes to the dataset. Operations such as initialization of the variables, cleansing the data, making appropriate labels for the data takes place. In our case, the dataset contains a parameter skin thickness, which is column that has a weak correlation to the contribution of a person being diabetic. Hence, we remove the column for our analysis. In this stage, we can calculate the numeric aspects of the data, such as the average of a particular column, number of cases of the column based on conditions etc.The dataset contains the values for the people having diabetes and people who don't. Hence, we calculated the count for each case, the result turned out like this:

People with Diabetes: 268

People without Diabetes: 500

In the given data, around 35% of the people have been diagnosed with diabetes.

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

## VI. EXPERIMENTAL RESULTS

### *TABULATED RESULTS*

After performing the Random Forest and Naive Bayes algorithms, we are generating the following results for the different splits of training and testing data:

Prediction Using Naïve Bayes

| Train | Test | Train_Result (% value) | Test_Result (% value) |
|---|---|---|---|
| 60 | 40 | 75.22% | 77.27% |
| 70 | 30 | 75.98 | 74.89 |
| 75 | 25 | 75.87 | 74.48 |
| 80 | 20 | 75.57 | 77.27 |

In the above table, we can see that for the four different splits, we get results that are close to 75% in the training set and 74-77% in the test results.

This depicts that the training set has been trained up to 75% accuracy which means that the data that has been trained has been used to predict the test results which have a 75% average accuracy in the analyzing of the dataset.

For each split, the percentage of test results depicts that74-77% of the dataset prediction is accurate and rest of the 25% approx. cannot be predicted due to various other reasons.

Prediction Using Random Forest

| Train | Test | Train_Result (% value) | Test_Result (% value) |
|---|---|---|---|
| 60 | 40 | 97.61 | 77.27 (best) |
| 70 | 30 | 98.70 | 73.16 |
| 75 | 25 | 98.61 | 72.92 |
| 80 | 20 | 98.37 | 74.68 |

In the above table, we can see that for the four different splits, we get results that are close to 98% in the training set and 72-77% in the test results.

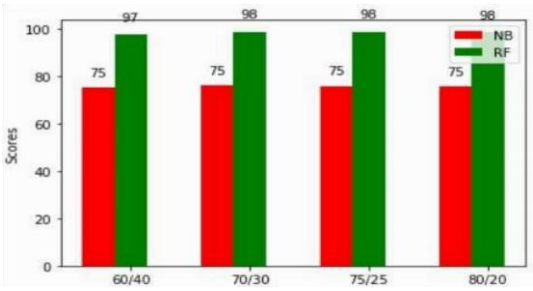
This depicts that the training set has been trained up to 98% accuracy which means that the data that has been trained has been used to predict the test results which have a 75% average accuracy in the analysing of the dataset.

For each split, the percentage of test results depicts that 72-77% of the dataset prediction is accurate and rest of the 25% approx. cannot be predicted due to various other reasons.

While analyzing both the tables, we can understand that the Random Forest algorithm has a better training set result which in turn gives a better accuracy of the prediction and analysis. The dataset is trained to the maximum accuracy where all variables are taken into aspect without excluding missing data as Random Forest algorithm will make sure that there is no missing data in large datasets.

Naïve Bayes algorithm tends to ignore missing data which does not provide accurate results while performing analysis. From the tables, we can find out that the best prediction result is giving by the 60/40 split while performing Random Forest.

### *COMPARISON GRAPHS*



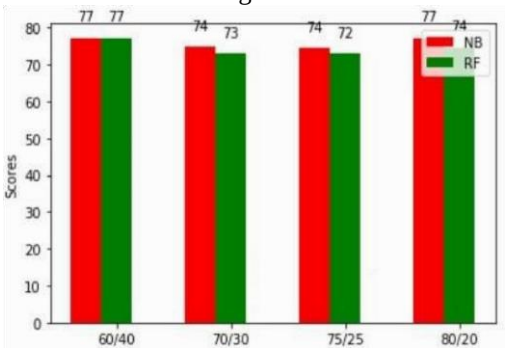

Comparison of Training results for various splits

The above graph depicts the comparison graph for the training results for both Naïve Bayes and Random Forest for various splits. We can understand that the Random Forest training results are more accurate when compared to that of Naïve Bayes as it gives a 98% accuracy when it comes to training the dataset.

The training result of Naïve Bayes is very low compared to that of Random Forest as there are errors that occur in the Naïve Bayes algorithm while performing training. Sometimes, it cannot detect missing data so there are fluctuations and errors in the accuracy of the result, but in the case of Random Forest, it gives the proper accuracy

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

even when it comes to large datasets like PIMA dataset.




Comparison of Test results for various splits

The above graph depicts the comparison graph for the testing results for both Naïve Bayes and Random Forest for various splits. We can understand that the Random Forest and Naïve Bayes test results are almost the same and they differ by 2-3%.

Even though the Naïve Bayes testing results are greater compared to the Random Forest results, the training result for Naïve Bayes was lesser than that of Random Forest, so the accuracy of the results when compared, is greater for Random Forest since the training data was much more accurate when compared to Naïve Bayes.

After analyzing the results, we can come to the conclusion that the Random Forest algorithm is a more efficient method to analyze the dataset using means of splitting it into training and testing sets. It serves as a more accurate method of prediction of diabetes.

## VII. CONCLUSION

Diabetes is one of the most chronic and the largest growing disease in India. According to the World Health Organization (WHO), India had 69.2 million people living with diabetes as of 2015. A study conducted by the American Diabetes Association states that India will see a great increase in the number of people diagnosed with diabetes by the 2030. Identifying diabetes or predicting the upcoming of a diabetic life can be propelled by using various machine learning techniques like Naïve Bayesian Network, Random Forest etc.

From this project, we can conclude that the best method of prediction of diabetes is Random Forest. This method gives us an approximate result after the splitting and analysis of the training and testing data. The efficiency of this method is much better compared to that of Naïve Bayes. The analysis done from the PIMA dataset is really important. The aim of splitting the dataset is to find the highest/best accuracy of the Algorithms and as to how they would respond if the data split set is varied. Procuring the dataset is done to make sure that there are no empty values In the data set so that the accuracy of our prediction model is high. Pre-processing of the dataset makes sure that all the attributes (columns) are taken into account while predicting. From the above prediction and analysis, we can observe that the results obtained using Random Forest Algorithm give us an accuracy of 98%. The several decision trees that are part of Random Forest are used to result in this maximum efficiency value. There by, we can conclude that it is more efficient than Naïve Bayes. Hence this proposed method will give us an efficient method for both analysis and prediction of diabetes.

## VIII.REFERENCES

[1] Priyanka Indoria, Yogesh Kumar Rathore (2018). A survey: Detection and Prediction of diabetes using machine learning techniques. IJERT

[2]Khaleel, M.A., Pradhan, S.K., G.N Dash (2Ol3). A Survey of Data Mining Techniques on Medical Data for Finding frequent diseases. IJARCSSE.

[3]K. Vembandasamy, R. Sasipriya, E. Deepa (2015). Heart Diseases Detection using Naïve Bayes Algorithm. IJISET.

[4]Tawfik Saeed Zekia, Mohammad V. Malakootib, Yousef Ataeipoorc, S. Talayeh Tabibid. An Expert System for Diabetes Diagnosis. American Academic & Scholarly Research Journal Special Issue Vol. 4, No. 5, Sept 2012.

[5]Vishali Bhandari and Rajeev Kumar. Comparative Analysis of Fuzzy Expert Systems for Diabetic Diagnosis. International Journal of Computer Applications (0975 – 8887) Volume 132 – No.6, December 2015.

[6]Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, "Machine Learning and Data Mining Methods in Diabetes Research", Jan 8, 2017.

[7]Eka Miranda, Edy Irwansyah, Alowisius Y. Amelga, Marco M. Maribondang, Mulyadi Salim (2016). Detection of cardiovascular Disease Risk's Level for Adults using naïve Bayes Classifier, The Korean Society of Medical informatics (KOSMI).

[8]Zheng T, Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y (2017). A Machine Learning-Based Framework to identify Type 2 Diabetes through Electronic Health Records, International Journal of medical informatics (IJMI) Vol 9, pages 120-127.

[9]Francesco Mercaldo, Vittoria Nardone, Antonella Santone (2017). DiabetesMellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques, Procedia Computer Science 112 (2017) 2519-2528.

[10]Rahul Joshi, Minyechil Alehegen (2017). Analysis and Prediction of Diabetes Disease using Machine Learning Algorithm: Ensemble Approach, International Research Journal of Engineering and Technology (IRJET) Volume 04 Issue 10, e-ISSN: 2395-0056.

[11]Jimmy Singla, Dinesh Grover (2017). The Diagnosis of Diabetic Nephropathy using Neuro-Fuzzy expert system, Indian Journal of Science and Technology (IJST) Vol 10(28) ISSN (online) 0974-5645.

[12]Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, Leila Shahmoradi (2017). An Analytical method for Disease Prediction using Machine Learning Techniques, Computers and Chemical Engineering 106 (2017) 212-223.

[13]Saba Bashir, Usman Qamar, Farhan Hassan Khan, Lubna Naseem (2016). HMV: A Medical Decision Support Framework using Multi-layer Classifiers for disease prediction, Journal of Computational Science 13 (2016) 10-25.

[14]Mekruksavanich, S. (2016). Medical Expert System Based Ontology for Diabetes Disease Diagnosis. In

Software Engineering and Service Science (ICSESS), 7th IEEE International Conference Pages 383-389, IEEE.

[15]Rajeswara Rao, D., Vidyullata Pellakuri, SathishTallam, Ramya Harika, T. (2015). International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 6 (2), 1103-1106.