International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Prediction of Diabetics using Machine Learning

G. Geetha, K.Mohana Prasad

Abstract: Around 50.9 Million People in India suffer from diabetics and Tamil Nadu stands second in the list of Indian states. The main objective of this paper is to develop prediction modeling of the given medical data of patients with and without diabetics. Through this paper, we aim to create hybrid models that can be easily used by doctors to treat patients with diabetics. Naïve Bayes and Random forest algorithms are used to predict whether a person having diabetics or not, by keeping his health conditions in mind. Thus this process enables doctors to easily group, classify and categorize the disease type accordingly treatment can be given to them.

We split the Dataset into 1) Training set and 2) Testing Set and perform analysis on them. The Pima Indian dataset was used to study and analyze the data, alongside with data mining techniques. It is the data obtained from the National Institute for Diabetics patients which contains n number of medical predictor variables and one target variable.

Initially, we replace the null values that are there in the dataset with the mean values of the respective columns. We then split the dataset into different ways to perform analysis on them: 85/15, 80/20, 70/30, 60/40. After procuring the data set, we apply Naïve Bayes and Random Forest algorithms on this. The Naïve Bayes algorithm is used here to find the probability of the independent features/columns. The data set is given as an input and the prediction takes place according to the NB Model. The Random Forest algorithm is used here in order to perform feature selection. It takes n inputs from the dataset and builds numerous uncorrelated decision trees during the time of training. It then displays the class that is the mode of all of the class outputs by individual trees.

Keywords : Diabetics, Machine learning, Random Forest, Data mining, Naïve Bayes Classification.

I. INTRODUCTION

A report by WHO (world health organization) says that there are 422 million people are suffering from diabetics and around 1.6 million leads to deaths in the year 2019. In 2012, 2.2 million people died due to a high blood glucose level. Many diseases are caused due to diabetics and they affect our kidneys, eyes, heart and also other organs. To understand diabetics, we first need to learn how the body works without diabetics. The food that we eat contains various kinds of components such as sugar, protein, fat, etc, The sugar we gain mainly comes from foods that contain carbohydrates which provides our body with energy. Foods rich in carbohydrates are bread, cereal, pasta, rice, fruit, dairy products, and vegetables. Such kinds of food, when consumed, are broken down into glucose by our body and they are supplied throughout by the means of our bloodstream. Mainly, glucose travels to the brain as it is required mainly for the body's thinking and functionality. The rest of the glucose is supplied to the rest of our bodies such as the cells, and the liver. Insulin is an important component that is required for the functionality of the human body. It is a hormone generated by beta cells in the pancreas. It permits the glucose to move from the bloodstream to the cells in our body. Since the pancreas is used to produce insulin, it needs enough glucose. If the pancreas cannot produce enough insulin, the glucose builds up and this is how diabetics is developed in an individual.

The signs or symptoms of diabetics can be listed as Blurred vision, Fatigue, Weight Loss, Increased Hunger and Thirst, Frequent Urination, Confusion, Poor Healing, Frequent Infections, Difficulty in concentrating.

Type 1 Diabetics: Type 1 diabetics arises when our immune system obliterates beta cells in your pancreas, this is the cells that construct the insulin in our body. The insulin builds up in our blood and as a result, our cells are in a state of starvation which causes diabetics. It occurs usually in people less than 30 years and about 5 - 10% of those with diabetics but can occur at any age.

Type 2 Diabetics: People who have type 2 diabetics secrete insulin, but their cells do not consume it as much as they should. The pancreas generates more insulin in order to obtain glucose into the cells since the cells do not make use of it properly sugar builds up in our bloodstream.

Data analytics is the identification of the hidden patterns from huge amounts of data for drawing concrete conclusions. In the health care domain, various machine learning algorithms are mostly used to examine the input medical data and to build learning models from where prediction can be done. In this paper, we are going to use techniques such as naive Bayes and RandomForest algorithms to predict diabetics with the help of PIMAdataset.

reference section. In the case of exclusion of references, it should be less than 5%.

I. Literature Review

PriyankaIndoria, YogeshKumar Rathore (2018). "A survey for detecting and predicting diabetics using machine learning techniques" [1]. This paper focuses on machine learning techniques for improving the accuracy of perception and diagnosis of the diseases.

Manuscript published on January 30, 2020.
* Correspondence Author
G. Geetha *, Research Scholar, Sathyabama Institute of Science and Technology, Chennai
Dr.K.Mohana Prasad, Associate Professor, Sathyabama Institute of Science and Technology, Chennai

Supervised, unsupervised, reinforcement, semi-supervised, deep learning and evolutionary

learning algorithms are the various machine learning techniques that are used to classify the data sets. It also shows the comparison of the two methods namely, Naïve Bayes and Artificial Neural Networks (ANN).

The Bayesian Network applies the Naïve Bayes theorem which firmly assumes that the presence of any attribute in a class is not related to the presence of any other attribute, making it much more advantageous, efficient and independent

I. Mohammed Abdul Khaleel, Sateesh Kumar Pradhan, G.N Dash (2013). "A survey of data mining techniques for finding locally frequent diseases" [2]. This paper focuses on mining the required medical data to find frequently occurring diseases, such as breast cancer, heart illness, lung cancer and so on. Data mining techniques like Apriori and FPGrowth, linear genetic programming, decision tree algorithms, unsupervised neural networks, outlier prediction techniques, classification algorithm, NaïveBayesian and so on have been applied.

II. K. Vembandasamy, R. Sasipriya, E. Deepa (2015). "Aims on analyzing heart diseases using a Naïve Bayesian algorithm" [3]. The algorithm used here is Naïve Bayes, which firmly assumes that the presence of any attribute in a class is not related to the presence of any other attribute, making it much more advantageous, efficient and independent. The tools used are WEKA and classification is done by splitting data into 70% of the percentage split. The naïve Bayes technique used was able to produce 86.41% of the input data correctly and 13.58% of inaccurate instances. He uses a dataset collected from a leading diabetic research institute in Chennai which has about 500 instances or patients.

III. TawfikSaeed Zeki et al. "An expert system for diagnosing diabetics"[4]. They proposed rule-based IF-THEN system. They have used three modules for 3 stages, they are Block Diagram, Mockler Charts, and Decision Tables. After considering many factors, this system provides a diagnosis of diabetics. It was developed inVP-Expert.

IV. Vishali Bhandari and Rajeev Kumar, "Comparative Analysis of Fuzzy Expert Systems for Diabetic Diagnosis " [5] compared different fuzzy expert systems by using multiple parameters for diagnosing the diabetics. MATLAB fuzzy logic toolbox was used for the comparative study of these expert systems. Five parameters were used for comparison and results were generated.

V. IoannisKavakiotis and Olga Tsave, " Machine Learning and Data Mining Methods in Diabetics Research" [6]. They made a systematic survey on various machine learning algorithms and data mining techniques in diabetics prediction and diagnosis, complication and health care management on general characteristics of data. The methods which are a part of this approach are called filter methods and the feature set is filtered out before the model construction.

VI. Eka Miranda, EdyIrwansyah, AlowisiusY. Amelga, Marco M. Maribondang, MulyadiSalim, "Detection of cardiovascular Disease Risk's Level for Adults using naïve Bayes Classifier"[7]. This paper focuses on detecting cardiovascular disease risk levels using the Naïve Bayes classifier. The primary risk factors such as diabetics mellitus, coronary artery function, kidney function and the level of lipids in the blood are some of the characteristics of cardiovascular illness through this risk level of the disease that could be determined. Class labels were to be assigned according to the values of these risk factors: risk level 1, risk level 2 and so on. The evaluation of this method was done in three parameters namely accuracy, sensitivity, and specificity. The proposed model delivered 80% accurate results. The experiment was conducted by a variety of machine learning methods like naïve Bayes, decision trees, classification or clustering, and neural networks. The result showed that among all the naïve Bayes has the highest accuracy rate.

VII. ZhengT, XieW, Xu L, He X, Zhang Y, You M, Yang G, Chen Y, "A Machine Learning-Based Framework to identify Type 2 Diabetics through Electronic Health Records" [8]. The aim of this paper is to identify type 2 diabetics (T2DM) using electronic health records (ERH). To identify diverse genotype-phenotype associations affiliated with diabetics type 2 through phenome-wide association(PheWAS) study and genome-wide association (GWAS) study and controls are to be identified (for example, via an Electronic Health Records). They developed a semi-automated framework using a machine-learning algorithm to improve the recall rate by keeping a low false-positive rate. They proposed a framework that identifies subjects with or without T2DM from ERH through engineering and machine learning. The following machine learning models including Random Forest, Naïve Bayes, Logistic Regression, K-Nearest- Neighbor and Support Vector Machine are evaluated and contrasted to measure the individual performance metrics. They have used a sample of 300 patients records randomly from 23,281 diabetics patients records of EHR repository.

VIII. Francesco Mercaldo, VittoriaNardone, AntonellaSantone (2017), "Diabetics Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques" [9]. The aim of this paper is to differentiate between the people affected with diabetics versus the ones that aren't affected by diabetics. They conducted hypotheses testing using different attributes of different people with diabetics and without diabetics. They used two methods to test for null hypotheses namely, Mann-Whitney (With p- level adhered to 0.05) and Kolmogorov-Smirnov (With p-level adhered to 0.05). They have chosen a 0.05 level of significance and six classification algorithms namely, Hoeffding Tree, Random Forest, JRip, Multilayer Perceptron (deep learning algorithm), Bayes Network. The metrics they used to evaluate the result is Precision, Recall, F-measure and ROC area.

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*

Fig.1. Features of Pima Indians dataset for Diagnosing Type 2 Diabetics Disease

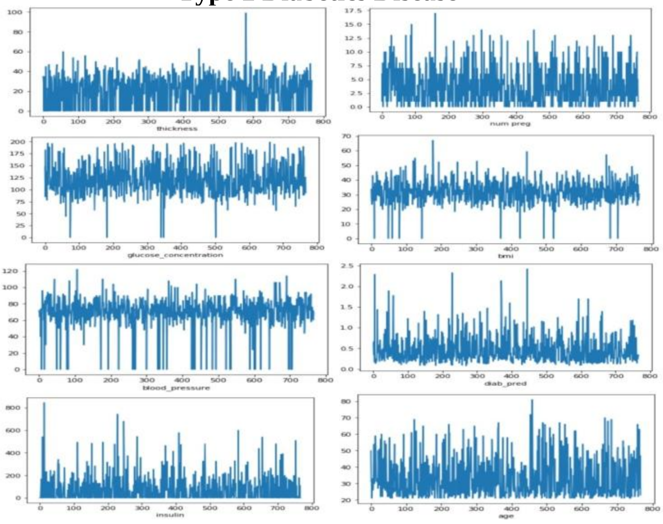

Fig. 2. Statistical Analysis for Mean and Standard Deviation in Pima Indians Diabetics Data Set

I. *Methodology*

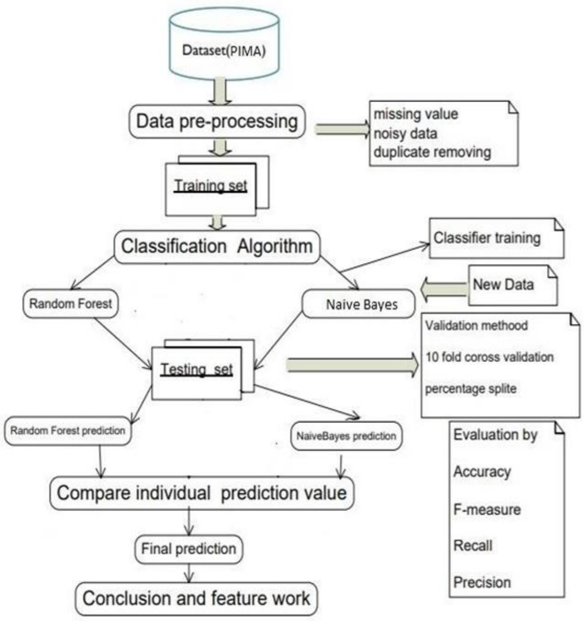

Fig. 3. Proposed System Architecture

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

Fig. 4. Dataset variables and their values

1. Procuring the dataset

The dataset used here is the PIMA Indian Dataset. It is the data obtained from the National Institute for Diabetics. It consists 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 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 to mainly 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 have 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. The 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.

2. Preparation of the Dataset

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, this column 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 diabetics and people who don't. Hence, we calculated the count for each case and the result turned out like this:

People with Diabetics: 268
People without Diabetics: 500
In the given data, around 35% of the people have been diagnosed with diabetics.

3. Splitting the Data

Dividing the dataset into training and test data is one of the crucial steps in analysis. This process is basically carried out to ensure test data is different from the training data because we need to test the model followed by the training process. First, the training data undergoes through learning and then, the data which is trained is generalized on the other data, based on which the prediction is made. The dataset in our case is split into multiple variants and prediction is performed accordingly. The dataset has multiple columns that are medical predictors and one target column, that of the diabetics outcome. The medical predictors are given as inputs to a variable and the target variable is given as input to another variable.

Using the inbuilt function, train_test_split, the dataset is split into arrays and is mapped to training and test subsets. In our case, we are performing splits of 80/20,70/30,75/25,60/40 and the accuracy of each is recorded. It was noticed that the dataset contain some null values, to streamline the analysis and the prediction, the null values were filled with the mean values of the respective columns.

4. Naïve Bayes

The most widely used type of Bayesian Network for classification is the Naïve Bayesian's, which has the highest accuracy value of up to 99.51% respectively. The Bayesian Network applies the Naïve Bayes theorem which firmly assumes that the occurrence of any particular attribute in a class is not related to the presence of any other attribute, making it much more advantageous, efficient and independent.

The Naïve Bayesian is based on the conditional probability (given a set of features, the probability of occurrence of certain results):

Naive Bayes classifiers can handle a subjective number of autonomous features to decide whether nonstop or all out. Given a lot of features, X = {x1,x2,x3...,xd}, the posterior probability for the occasion Cj could be build among a lot of conceivable results C = {c1,c2,c...,cd}, where X is the indicators and C is the arrangement of absolute dimensions of the needy variable. Utilizing Bayes' standard:

$$p\left(C_j \mid x_1, x_2, \ldots, x_d\right) \propto p\left(x_1, x_2, \ldots, x_d \mid C_j\right) p\left(C_j\right)$$

where p(Cj | x1,x2,x..., xd) is the posterior probability of class i.e., the probability that X has a place with Cj. Since Naive Bayes expect the restrictive probabilities of the autonomous features are exactly free we can decay the probability from the result terms:

$$p(X \mid C_j) \propto \prod_{k=1}^{d} p(x_k \mid C_j)$$

and revise the posterior as:

$$p(C_j \mid X) \propto p(C_j) \prod_{k=1}^{d} p(x_k \mid C_j)$$

By using Bayes' standard above, the variable X with a class level Cj that complete the most astounding posterior probability. The indicator (autonomous) factors simplifies the grouping task significantly, because it permits the class restrictive densities p(xk | Cj) can be determined autonomously for every factor, i.e., it decrease a multidimensional activity to various one-dimensional ones. Essentially, Naive Bayes eliminates a high-dimensional density estimation undertaking to a one-dimensional part density estimation. We import the model into a variable, input the training data, fit it accordingly to the Gaussian Naïve Bayes model using the function .fit(). Then a classification of the array present in each of the training and test variables is performed. This classification is the key operation taking place as it is performing the prediction of the input data according to the Naïve Bayes model. The accuracy of the model is then predicted by comparing the predicted model with the original model. This is done with the help of the metrics library present in sklearn.

5. Random Forest:

The Random Forests algorithm is a most powerful classification algorithm that can classify a large amount of data with high accuracy. Random Forest is a group learning method (it's a form of the nearest neighbor predictor) for classification and regression that construct several number of decision trees during training time and then finally the class with the higest number of votes will be consider as the predictor's output.

The tree predictors variables for each tree depend on the values of a random vector sampled independently with the same distribution for all trees in the forest. Random Forests solves this problem of high variance and high bias by finding a natural balance between the two extremes. They also have a mechanism to estimate the error rates (Out of the Bag error). Many machine learning models, like linear and logistic regression, are easily impacted by the outliers in the training data. Outliers are changes in the system behavior and can also be caused by human error, instrument error. There are chances for a given sample to be contaminated. These outliers or extreme values do not impact model performance/accuracy. RF Algorithm overcomes and solves this problem.

In our case, we have split the labels into two variables, and these are the input to the classifier. One of the greatest strengths of Random Forest classifiers is the ability to use with any kind of data, especially with feature selection.

In our case, we use the RandomForestClassifier() function in sklearn library to perform prediction on the data. The input training and test data are fitted to the model using fit(). The training data is then classified into arrays during prediction. The accuracy of the model is obtained by comparing the predicted values against the original set of values.

6.Finding the accuracy

First, the accuracy of the training data are checked by feeding the arguments for the training data split. After that,
accuracy of the testing data is done by the same way with the testing data as the parameters. By comparing these two, we can construct a confusion matrix. The main intention of confusion matrix is to evaluate the accuracy of the classification. By definition a confusion matrix C is that, Ci,j represents the number of observations known to be in group i but predicted to be in group j. Hence in binary classification, the count of true negatives are C0,0, false negatives are C1,0, true positives are C1,1 and false positives are C0,1.

```
Confusion Matrix
[[38 17]
 [18 81]]

Classification Report

              precision    recall  f1-score   support

           1       0.68      0.69      0.68        55
           0       0.83      0.82      0.82        99

   micro avg       0.77      0.77      0.77       154
   macro avg       0.75      0.75      0.75       154
weighted avg       0.77      0.77      0.77       154
```

Classification Report for Naïve Bayes

```
Confusion Matrix
[[37 18]
 [19 80]]

Classification Report

              precision    recall  f1-score   support

           1       0.66      0.67      0.67        55
           0       0.82      0.81      0.81        99

   micro avg       0.76      0.76      0.76       154
   macro avg       0.74      0.74      0.74       154
weighted avg       0.76      0.76      0.76       154
```

Classification Report for Random Forest

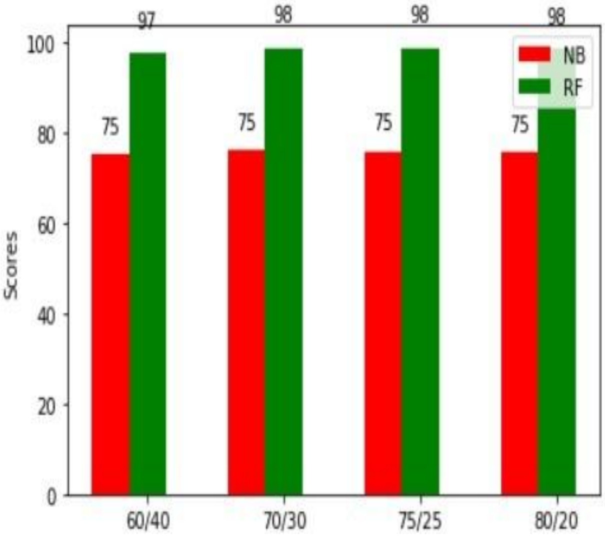

Fig. 6. Prediction using Naive Bayes

II. 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:

In figure 5, 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.

In figure 6, 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

Fig. 5. Comparison of Training results for various splits

Train	Test	Train Results(%)	Test Results(%)
60	40	75.22	77.27
70	30	75.98	74.89
75	25	75.87	74.48
80	20	75.57	77.27

While analyzing both the tables, we can understand that the Random Forest algorithm has a better training set result which in turn gives better accuracy of the prediction and analysis.

The data set is trained to the maximum accuracy where all variables are taken into aspect without excluding missing data as the 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 the result is giving by the 60/40 split while performing Random Forest.

Figure 7 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 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 the means of splitting it into training and testing sets. It serves as a more accurate method of prediction of diabetics.

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 even when it comes to large datasets like PIMA dataset

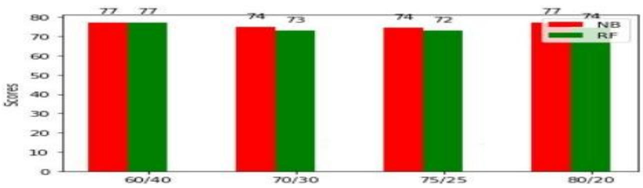

Fig. 7. Comparison of Test results for various splits

Fig. 8. Prediction using Random Forest

Train/Test	Train result	Test result
60/40	97.61	77.27
70/30	98.70	73.16
75/25	98.61	72.92
80/20	98.37	74.37

III. CONCLUSION

Identifying diabetics or predicting the upcoming of a diabetic life can be propelled by using various machine learning techniques like Naïve Bayesian Network, Random Forest, etc. In this paper, we can conclude that the best method of prediction of diabetics 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 PIMAdataset, the aim of splitting the dataset is to find the highest/best accuracy of the algorithms and how they would respond if the data split 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. Preprocessing 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 the 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. Therefore,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 diabetics.

FUTURE SCOPE

Healthcare professions found it hard to find healthcare data and perform analysis on them due to lack of tools, resources. But using ML, we can overcome this and can perform analysis on real-time data leading to better modeling, predictions. This enhances and improves overall healthcare services. Now, IoTs being integrated with ML in order to make smart healthcare devices that sense if there is any change in the person's body, health data when he uses the device (Pacemaker, Stethoscope, etc.) and this will notify the person regarding this through an app. This helps in easy monitoring, advanced prediction and analysis thereby reducing errors, saving time and life of people.

REFERENCES

1. Priyanka Indoria, Yogesh Kumar Rathore. A survey: Detection and Prediction of diabetics using machine learning techniques. IJERT, 2018.
2. Khaleel, M.A., Pradhan, S.K., G.N Dash. A Survey of Data Mining Techniques on Medical Data for finding frequent diseases.IJARCSSE, 2013.
3. K. Vembandasamy, R. Sasipriya, E. Deepa. Heart Diseases Detection using Naïve Bayes Algorithm. IJISET, 2015.
4. Tawfik Saeed Zekia, Mohammad V. Malakootib, Yousef Ataeipoorc, S. Talayeh Tabibid. An Expert System for Diabetics Diagnosis. American Academic & Scholarly Research JournalSpecialIssueVol.4,No.5, September 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 Diabetics Research", Jan 8, 2017.
7. Eka Miranda, EdyIrwansyah, AlowisiusY. Amelga, Marco M. Maribondang, MulyadiSalim. Detection of cardiovascular Disease Risk's Level for Adults using naïve Bayes Classifier, The Korean Society of Medical informatics (KOSMI), 2016.
8. ZhengT, XieW, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. A Machine Learning-Based Framework to identify Type 2 Diabetics through Electronic Health Records, International Journal of medical informatics (IJMI),2017, Vol9, pages120-127.
9. Francesco Mercaldo, Vittoria Nardone, Antonella Santone (2017). Diabetics Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques, ProcediaComputerScience112, 2017,2519-2528.
10. Rahul Joshi, Minyechil Alehegen . Analysis and Prediction of Diabetics Disease using Machine Learning Algorithm: Ensemble Approach, International Research Journal of Engineering and Technology(IRJET),2017Volume 04Issue10, e-ISSN:2395-0056.
11. Jimmy Singla, Dinesh Grover.The Diagnosis of Diabetic Nephropathy using Neuro-Fuzzy expert system, Indian Journal of Science and Technology (IJST), 2017, Vol10(28) ISSN (online)0974-5645.
12. Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, Leila Shahmoradi.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.HMV: A Medical Decision Support Framework using Multi-layer Classifiers for disease prediction, Journal of Computational Science 13, 2016,10-25.
14. Mekruksavanich, S. Medical Expert System Based Ontology for Diabetics Disease Diagnosis. In Software Engineering and Service Science (ICSESS), 7th IEEE International Conference Pages 383-389,IEEE,2016.
15. Rajeswara Rao, D., Vidyullata Pellakuri, SathishTallam, RamyaHarika, T. International Journal of Computer Science and Information Technologies (IJCSIT),2015, Vol. 6 (2),1103-1106.