International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 07 | July - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Performance analysis of diabetic prediction by using machine learning algorithm

SUPRIYA P.M
P G Scholar
Department of Electronics and Communication
University B.D.T College of Engineering
Davanagere-577004, Karnataka, India

Dr. Shreedhar Murthy
Associate Professor
Department of Electronics and Communication
University B.D.T College of Engineering
Davanagere-577004, Karnataka, India

ABSTRACT

Diabetes is a common, chronic disease. Prediction of diabetes at an early stage can lead to improved treatment. Data mining techniques are widely used for prediction of disease at an early stage. In this research paper, diabetes is predicted using significant attributes, and the relationship of the differing attributes is also characterized. Various tools are used to determine significant attribute selection, and for clustering, prediction, and association rule mining for diabetes. Significant attributes selection was done via the principal component analysis method. I am finding to indicate a strong association of diabetes with body mass index (BMI), Blood pressure, Skin thickness, Insulin, diabetic prediction, and with glucose level, which is extracted via the A priori method. K-means clustering techniques were implemented for the prediction of diabetes. The KNN technique provided a best accuracy of 73%, and may be useful to assist medical professionals with treatment decisions.

***Keywords*: diabetes mellitus, KNN, machine learning, prediction.**

I. INTRODUCTION

The disease or condition whose effects are permanent is a chronic condition. These types of diseases affected in our of life, which is major adverse effect. Diabetes is one of the most acute diseases, and is present worldwide. The major reason of deaths in adults across the globe includes this chronic condition. Chronic conditions are also cost associated. A major portion of budget is spent on chronic diseases by governments and individuals. The worldwide statistics for diabetes in the year 2013 revealed around 382 million individuals had this ailment around the world. It was the fifth leading cause of death in women and eight leading cause of death for both sexes in 2012. Higher income countries have a high probability of diabetes. In 2017, approximately 451 million adults were treated with diabetes worldwide. It is projected that in 2045, almost 693 million patients with diabetes will exist around the globe and half of the population will be undiagnosed. In addition, 850 million USD were spent on patients with diabetes in 2017. Research on biological data is limited but with the passage of time enables computational and Statistical models to be used for analysis. A sufficient amount of data is also being gathered by healthcare organizations. New knowledge is gathered when models are developed to learn from the observed data using data mining techniques. Data mining is the process of extracting from data and can be utilized to create a decision making process with efficiency in the medical domain. Several data mining techniques have been used for disease prediction as well as for our knowledge discovery from biomedical data. Diagnosis of diabetes is considered a challenging problem for Quantitative research. Some parameters like A1c (The blood test that reflects The average blood glucose levels over the past 3 months of data), fructosamine white blood cell count, fibrinogen and hematological indices were shown to be ineffective due to some limitations. Different research Studies used these parameters for the diagnosis of diabetes .A few treatments have thought to raise A1C including chronic ingestion of liquor, salicylates and narcotics. Ingestion of vitamin C may elevate A1c.

When estimated by electrophoresis but levels may appear to diminish when estimated by chromatography. Most studies have suggested that a higher white blood cell count is due to chronic inflammation during hypertension. A family history of diabetes has not been associated with BMI and insulin However, an increased BMI is not always associated with abdominal obesity. A single parameter is not very effective to accurately diagnose diabetes and may be misleading in the decision making process. There is a need to combine different parameters to effectively predict diabetes at an early stage. Several existing techniques have not provided effective results. When different parameters were used for prediction of diabetes. In our study, diabetes is predicted with the assistance of significant attributes, and the association of the differing attributes. We examined the diagnosis of diabetes using ANN, RF and K-Means Clustering.

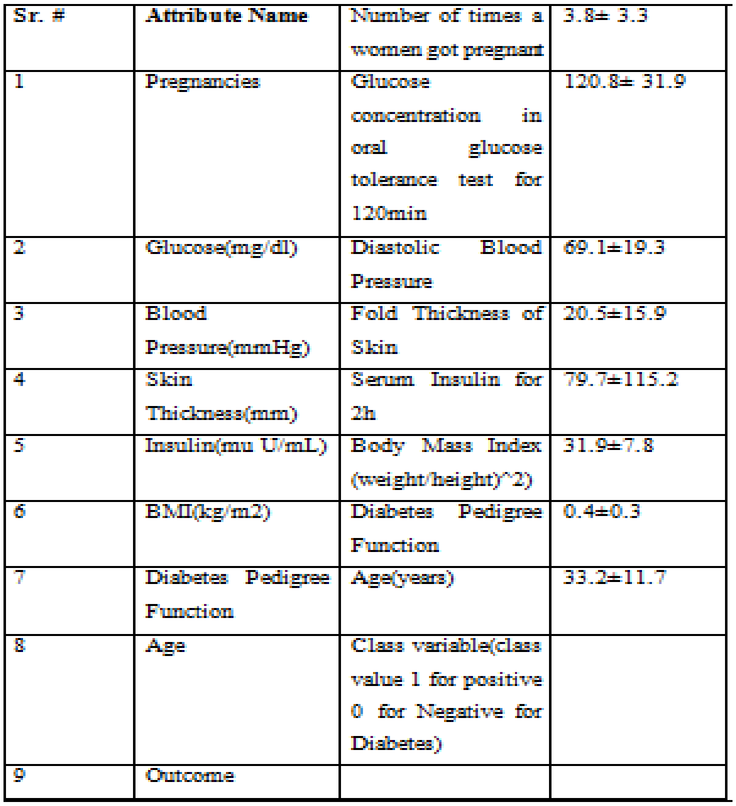
II. RELATED WORK

Shetty et al. used KNN and the Naïve Bayes technique has been used for the prediction of diabetes. Their technique was implemented as an expert software program, where users provide input in terms of patient records and the finding that either the patient is diabetic or not. Singh et al. applied different algorithms on datasets of different types. They used the KNN, random forest and Naïve Bayesian algorithms. The K-fold cross-validation technique was used for evaluation. Ahmed utilized patient information and plan of treatment dimensions for the classification of diabetes. Three algorithms were applied which were Naïve Bayes, logistic, and J48 algorithms. Anton yet al. utilized medical data for diabetes prediction. Naïve Bayes, function-based multilayer perceptron (MLP), and decision tree-based random forests (RF) algorithms were applied after pre-processing of the data. A correlation based feature selection method was employed to remove extra features. A learning model then predicted whether the patient was diabetic or not. Using a pre-processing technique, results were improved when employing Naïve Bayes as compared with other machine learning algorithms. Amine et al. compared different data mining algorithms by using the PID dataset for early prediction of diabetes. Sellappan Palaniappan et al. proposed a heart disease prediction system by using the Naïve Bayes, ANN and decision tree algorithms. Delen et al. used logistic regression, ANN, and decision trees to predict breast cancer using a large dataset. Shadab Adam Pattekari and Asma Praveen developed a web based application for prediction of myocardial infarction using Naive Bayes. Anuja Kumari and R. Chitra used the SVM model to diagnose diabetes using a high-dimensional medical dataset.

III. METHOD AND MATERIAL.

1. Dataset

The dataset used in this study, is originally taken from the National Institute of Diabetes and Digestive and Kidney Diseases (publicly available at: UCI ML Repository). The main Objective of using this dataset was to predict through diagnosis whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Many limitations were faced during the selection of the occurrences from the bigger dataset. The type of dataset and problem is a classic supervised binary classification. The Pima Indian Diabetes (PID) dataset having: 9 = 8 + 1 (Class Attribute) attributes, 768 records describing female patients (of which there were 500 negative instances (65.1%) and 268 positive instances (34.9%)). The detailed description of all attributes is given in Table 1.



Sr. #	Attribute Name	Number of times a women got pregnant	3.8± 3.3
1	Pregnancies	Glucose concentration in oral glucose tolerance test for 120min	120.8± 31.9
2	Glucose(mg/dl)	Diastolic Blood Pressure	69.1±19.3
3	Blood Pressure(mmHg)	Fold Thickness of Skin	20.5±15.9
4	Skin Thickness(mm)	Serum Insulin for 2h	79.7±115.2
5	Insulin(mu U/mL)	Body Mass Index (weight/height)^2)	31.9±7.8
6	BMI(kg/m2)	Diabetes Pedigree Function	0.4±0.3
7	Diabetes Pedigree Function	Age(years)	33.2±11.7
8	Age	Class variable(class value 1 for positive 0 for Negative for Diabetes)	
9	Outcome		

2. Data preprocessing

In real-world data there can be missing values and/or noisy and inconsistent data. If data quality is low then no quality results may be found. It is necessary to preprocess the data to achieve quality results.

Cleaning, integration, transformation, reduction, and discretization of data are applied to preprocess the data. It is important to make the data more appropriate for data mining and analysis with respect to time, cost, and quality.

2.1. Data cleaning

Data cleaning consists of filling the missing values and removing noisy data. Noisy data contains outliers which are removed to resolve inconsistencies. In our dataset, glucose, blood Pressure, skin thickness, insulin, and BMI have some zero (0) values. Thus, all the zero values were replaced with the median value of that attribute.

2.2. Data reduction

Data reduction obtains a reduced representation of the dataset that is much smaller in volume yet produces the same (or almost the same) result. Dimensionally reduction has been used to reduce the number of attributes in a dataset. The principal component analysis method was used to extract significant attributes from a complete dataset. Glucose, BMI, diastolic blood pressure and age were significant attributes in the dataset.

2.3 Data transformation

Data transformation consists of smoothing, normalization, and aggregation of data. For the smoothing of data, the binning method has used. The attribute of age has been useful to classify in five categories as shown in Table 2.

Table 2
Binning of age

Age(Years)	Age Bins
≤30	Youngest
31–40	Younger
41–50	Middle aged
51–60	Older
≥61	Oldest

Blood glucose concentration in patients who do not have diabetes is different from patients with diabetes. Glucose values have been divided into 5 categories as shown in Table 3.

Table 3.
Binning of glucose

Glucose	Glucose Bins
≤60	Very Low
61–80	Low
81–140	Normal
141–180	Early Diabetes
≥181	Diabetes

A strong association has been found between healthy and diabetic patients regarding their blood pressure levels. Blood pressure has been divided into five different categories as shown in Table 4.

Table 5.
Binning of Blood Pressure.

Blood Pressure	Diastolic Blood Pressure Bins
<61	Very low
61–75	Low
75–90	Normal
91–100	High
>100	Hypertension

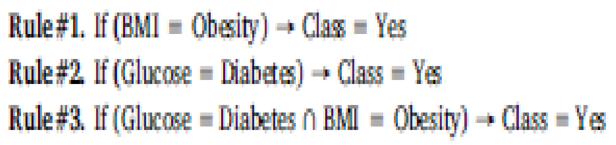
The relationship between BMI and diabetes prevalence is consistent. The prevalence of diabetes and obesity is increasing concurrently worldwide. Furthermore, previous studies have shown that BMI is the most important risk factor for type 2 diabetes BMI values have been categorized into five classes in Table 5.

Table 5
Binning of BMI

BMI	BMI Bins
<19	Starvation
19–24	Normal
25–30	Overweight
31–40	Obese
>40	Very Obese

3. Association rule mining

The association of blood glucose, blood pressure, age, and BMI with diabetes also depended on socio economic, geographic, and clinical factors are shown in Table 6.



Rule#1. If (BMI = Obesity) → Class = Yes
Rule#2. If (Glucose = Diabetes) → Class = Yes
Rule#3. If (Glucose = Diabetes ∩ BMI = Obesity) → Class = Yes

IV. METHODOLOGY

1. Modeling

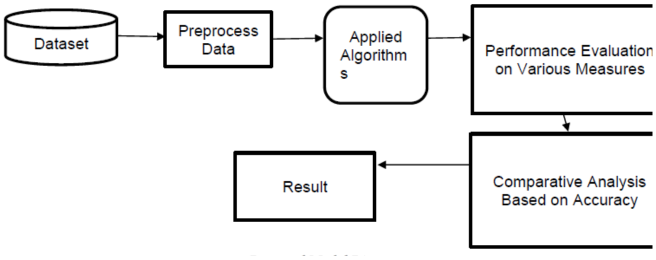
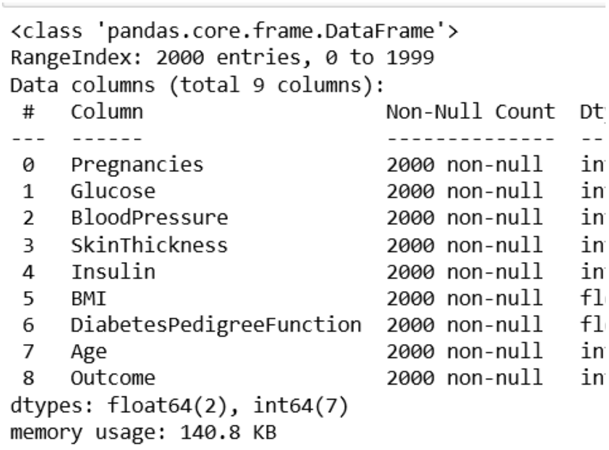

Dataset
Preprocess Data
Applied Algorithm s
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Result

Figure 1: Proposed and model diagram.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dt
---  ------                    --------------  --
 0   Pregnancies               2000 non-null   in
 1   Glucose                   2000 non-null   in
 2   BloodPressure             2000 non-null   in
 3   SkinThickness             2000 non-null   in
 4   Insulin                   2000 non-null   in
 5   BMI                       2000 non-null   fl
 6   DiabetesPedigreeFunction  2000 non-null   fl
 7   Age                       2000 non-null   in
 8   Outcome                   2000 non-null   in
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

There is no null value dataset.

2. Apply machine learning algorithm:

Models were used for early prediction of diabetes, following K-Nearest Neighbor Algorithm: K-nearest neighbor.

The simple classification and regression algorithm that used non parametric method proposed by Aha et. al. The algorithm records all valid attributes and classifies new attributes based on their resemblance measure. To determine the distance from point of interest to points in training data set it uses tree like data structure. The attribute is classified by its neighbors. In a classification technique, the value of k is always a positive integer of nearest neighbor. The nearest neighbors are chosen from a set of class or object property value.

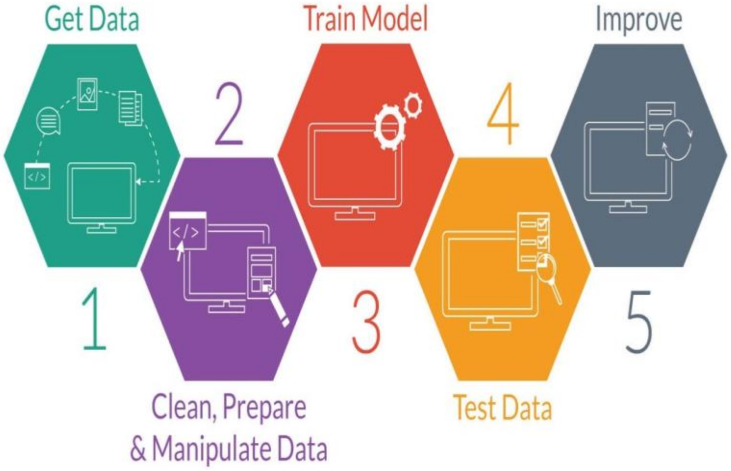
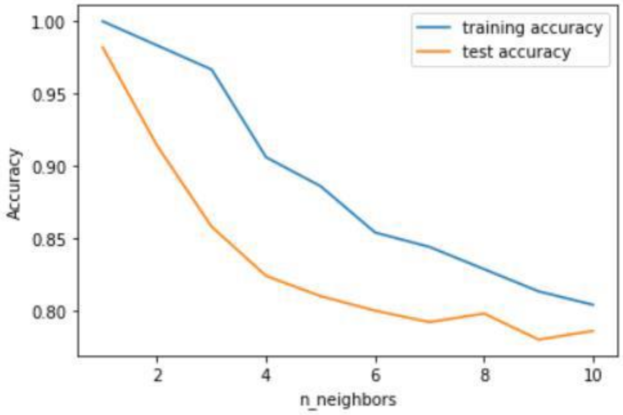


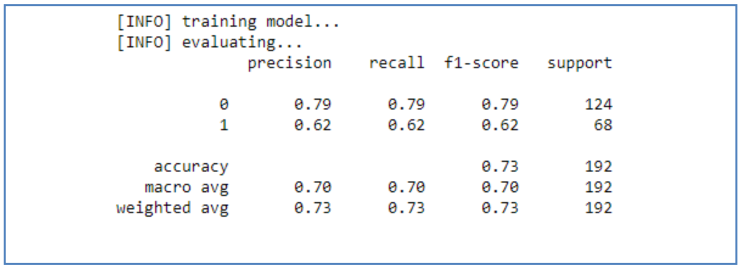
Figure 2: An overview of overall process.

V. RESULTS AND DISCUSSION

The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set, its “nearest neighbors.”



The above plot shows the training and test set accuracy on the y-axis against the setting of neighbor’s on the x-axis. Considering if we choose one single nearest neighbor, the prediction on the training set is perfect. But when more neighbors are considered, the training accuracy drops, indicating that using the single nearest neighbor leads to a model that is too complex. The best performance is somewhere around 9 neighbors.



```
[INFO] training model...
[INFO] evaluating...
              precision    recall  f1-score   support

           0       0.79      0.79      0.79       124
           1       0.62      0.62      0.62        68

    accuracy                           0.73       192
   macro avg       0.70      0.70      0.70       192
weighted avg       0.73      0.73      0.73       192
```

The algorithm has resulted into 73 % accuracy

Machine learning and data mining techniques are valuable in disease diagnosis. The capability to predict diabetes early, assumes a vital.

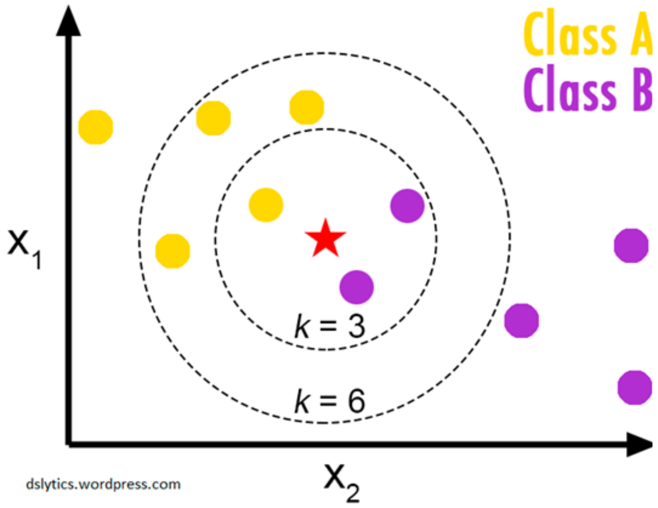


Figure 3: Correct and in correct clustering instances.

VI. CONCLUSION

The proposed project is an attempt to used Machine learning concept to analyze the dataset of samples being collected from kaggle repository, it has Indian women dataset with various attributes, we can conclude that BP, weight and age plays key role of becoming diabetic

REFERENCES

1. Dataset and Data Information: https://www.kaggle.com/uciml/pima-indians-diabetes-database
2. Pandas Profiling: https://towardsdatascience.com/speed-up-your-exploratory-data-analysis-with-pandas-profiling-88b33dc53625
3. Violin Plots: https://seaborn.pydata.org/generated/seaborn.violinplot.html
4. Shapiro-Wilk Test for Normality: https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/
5. Scaling/Outlier Treatment: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
6. Outlier Treatment: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba
7. SMOTE: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
8. Precision and ***Recall: https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c***