Analyzing the diabetes dataset using classification algorithms

G.Aswathi[1], T.k.Ramalakshmi[2], P.Monica[3], V.Aarthi[4], S.Jamuna[5]
[1,2,3,4,5]*Department of Computer Application, Thiagarajar College of Engineering (Affiliated to Anna University) Madurai, India*

***Abstract* - Diabetes is one of the deadliest diseases in the world. It is not only a disease but also creator of different kinds of diseases like heart attack, blindness etc. The objective of this analysis is to identify whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.**

***Index Terms* – Classification, Numpy, Pandas , Sklearn Seaborn.**

INTRODUCTION

This dataset is originally from the National Institute of Diabetes and Digestive .The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database.

In particular, all patients here are females at least 21 years old of Pima Indian heritage. In this data set, there are 9 baseline variables Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome for each of 768 diabetes patients.

METHODOLOGY

1) Examining the dataset: The dataset has 768 rows and 9 columns. Dataset is small and well labeled. The BMI column and Diabetes Pedigree Function column has float values and all the other columns has integer values. Our target variable is outcome column and the rest of them are independent variable. We find that some features like Glucose, Blood pressure, Skin Thickness, Insulin, BMI have unrealistic value zero which represent missing data. Data Cleaning will take place as data has got lot of missing values. Handling missing values can be done either by replacing null values with mode or mean or replacing the null value with a random variable.

2) Visualization: .Data visualization is the graphical representation of information and data. The inference from the visualization are, the count plot tells us that the dataset is imbalanced, as the number of patients who don't have diabetes is more than those who do. From the correlation heatmap, we can see that there is a high correlation between Outcome and [Glucose, BMI, Age, Insulin] ,We can select these features to accept input from the user and predict the outcome. We used the Matplotlib library to plot the data to understand what it is conveying.

3) Data pre-processing: Data preprocessing is the process of transforming raw data into an understandable format. The activities done in the Data Preprocessing are replacing zero values with NaN (in this dataset Glucose, Blood pressure , Skin Thickness, Insulin, BMI have zero values) ,Replacing Nan with mean Values, Statistical summary, Feature scaling using MinMaxScaler and Checking Dimensions.

4) Training and testing split: Training data and test data sets are two different but important parts in machine learning. Training set is the one on which we train and fit our model basically to fit the parameters whereas test data is used only to assess performance of a model. There are different methods for splitting the dataset, the most common following ratio is 80:20 or sometimes 70:30. We will use an 80:20 ratio.

5) Performing classification: Prepare the model using the X_train and y_train (training data) using the following algorithms:

- Logistic regression
- K Nearest neighbors
- Support Vector Machine

- Random Forest Classification
- Naive Bayes
- Decision Tree

With the different algorithms we have got different values while applying accuracy score metric. While applying confusion matrix for Gradient boosting, we observed that the algorithm has correctly detected 90 diabetes true positives, 21 diabetes true negatives, 30 diabetes false positives and 33 false negatives. We have also generated a classification report for Gradient boosting.

Table

S. No	Classification		
	Classification technique	Accuracy value	Comments
1	Logistic Regression	72.07792207792207	Logistic regression has the same accuracy as Gradient boosting
2	K Nearest neighbors	78.57142857142857	This algorithm has the highest accuracy
3	Support Vector Classifier	73.37662337662337	This algorithm has the medium accuracy
4	Naive Bayes	71.42857142857143	This algorithm has second least accuracy
5	Decision tree	68.18181818181817	Decision tree algorithm has given least accuracy for this dataset
6	Random Forest	75.97402597402598	This algorithm has the second highest accuracy
7	Gradient boosting	72.07792207792207	same accuracy as Logistic regression

CONCLUSION

Early detection of diabetes is one of the significant challenges in the healthcare industry. Data mining methods are very helpful for detecting it at an early stage. This work presents machine learning based diabetes prediction using classifier methods. After comparing the accuracy of various algorithms, we find that K-nearest neighbor algorithm has higher accuracy of .78%. We studied the reports of all the algorithms author used and we performed the Gradient Boosting.

REFERENCES

[1] https://www.kaggle.com/uciml/pima-indians-diabetes-database

[2] https://analyticsindiamag.com/7-types-classification -algorithms

[3] https://www.mygreatlearning.com/blog/seaborn-tutorial/

[4] https://www.nexsoftsys.com/articles/data-analysis-usi ng numpy-and-pandas.html

[5] https://www.dataschool.io/simple-guide-to-confusion-matrix terminology/

[6] https://www.webmd.com/diabetes/diabetes-causes