EAST WEST UNIVERSITY

Project Report

Project title: Diabetes Prediction

Course: CSE475

Title: Machine Learning

Section: 02

Semester: Spring 2021

Submitted by:

2016-1-68-016	Khandokar Zaeem Hasan
2016-2-60-145	Swarup Mondal
2016-3-60-018	Zakir Hossain
2017-2-68-001	Shimanto Krishna Chakroborty
2018-1-60-075	Md. Moniruzzaman Shanto

Date of Submission: 26/05/2021

1. **Introduction:** In this machine learning project we are going to predict diabetes using some populate machine learning algorithms. Nowadays so many people are dealing with diabetes. Diabetes prediction using machine learning can help doctors and normal people to recognize diabetes properly in very short time. Moreover, Doctors and people related with health sector can use it to save their time. Now a days every health sector has large number of databases with full of health information. That's why we can use this data to predict diabetes and we can save time and money both.

2. **Methodology:**
    1. Random Over sampling used to balance the data set. Produced equal number of yes and no rows to balance and shape the data set.
    2. Handled the missing data by using statistical method. Used mean value to fill the missing data.
    3. Used naïve bayes, decision tree and random forest to predict initially.
    4. Cross validation has used to make sure the algorithms are actually performing well.
    5. After that we used Voting Classifier and in it, we used naïve bayes, decision tree and random forest as estimators.
    6. Repeated K-Fold has used to know the average accuracy Voting Classifier is giving.

3. **Implementation:**

3.1. Data collection: We collected our data for diabetes predict from Kaggle. The name of the data set is Pima Indians Diabetes database. dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

3.2. Data Processing: Random Over sampling and statistical method has used for data processing.

3.3. Model Development: We used "Jupyter Notebook" and python language to implement this project.

3.4. Results:

a) Accuracy using Naive Bayes: 74 %. average Accuracy using cross validation: 71 %
b) Accuracy using Decision Tree: 78 %. average Accuracy using cross validation: 75 %
c) Accuracy using Random Forest Classifier: 88 %. Average Accuracy using Random Forest Classifier: 81 %.
d) Accuracy using Voting Classifier: 82 %. Average Accuracy for Repeated K-Fold: 76 %.

4. **Conclusion:**

4.1. Challenges: The main challenges were to find right data set. After that Handling the missing data and balancing the data set was a good challenge.

4.2. Limitation: This project cannot predict for all kind of data set. We have developed the model which specially works with the data set we selected. Moreover, we have developed this project by giving priority to time, that's we did not used any algorithm that takes time and gives more correct accuracy.

4.3. Future Directions: Hospitals, Health care center and other health organizations can use the project or other developed versions of the project to predict diabetes.