**DIABETTIC PREDICTION WITH FLASK**

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE & ENGINEERING**



**Submitted to:**

**Simarjit Singh Malhi**

**Submitted By:**

**Aditya Jain- 18BCS3365**

**Badal Jain- 18BCS3234**

**Shahe Faisal- 18BCS3016**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**Chandigarh University, Gharuan**

**INTRODUCTION:**

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. 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 its tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

**Methodology:**

Goal of the paper is to investigate for model to predict dia- betes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

1. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

Table 1: Dataset Description

|  |  |
| --- | --- |
| S No. | Attributes |
| 1 | Pregnancy |
| 2 | Glucose |
| 3 | Blood Pressure |
| 4 | Skin thickness |
| 5 | Insulin |
| 6 | BMI(Body Mass Index) |
| 7 | Diabetes Pedigree Function |
| 8 | Age |

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.

Figure 1: Ratio of Diabetic and Non Diabetic Patient

1. **Data Preprocessing**- Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively the process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre processing in two steps.
   1. **Missing Values removal**- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimentiosnality of data and help to work faster.
   2. **Splitting of data**- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.
2. **Apply Machine Learning**- When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction. The Techniques are follows-
3. **Logistic Regression**- Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories.

It classify the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes.

Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

Sigmoid function P = 1/1+e – (a+bx) Here P = probability, a and b = parameter of Model.

Ensembling - Ensembling is a machine learning technique Ensemble means using multiple learning algorithms together for some task. It provides better prediction than any other individual model thats why it is used. The main cause of error is noise bias and variance, ensemble methods help to reduce or minimize these errors. There are two popular ensemble methods such as Bagging, Boosting, ada-boosting, Gradient boosting, voting, averaging etc. Here In these work we have used Bagging (Random forest) and Gradient boosting ensemble methods for predicting diabetes.

**Model Building**

**Step1:** Import required libraries, Import diabetes dataset.

**Step2:** Pre-process data to remove missing data.

**Step3:** Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

**Step4:** Select the machine learning algorithm i.e, Logistic regression, Random Forest and Gradient boosting algorithm.

**Step5:** Build the classifier model for the mentioned machine learning algorithm based on training set.

**Step6:** Test the Classifier model for the mentioned machine learning algorithm based on test set.

**Step7:** Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

**Step8:** After analyzing based on various measures conclude the best performing algorithm.

**Step9**: Integrating ML model in website with flask API.

**Software and Hardware requirement:**

**Hardware**

● 64 Windows / Linux / MAC laptops for development

**Software**

● Python 3.8 with machine learning libraries (e.g. scikit-learn, numpy, pandas)

● Jupiter/ Spyder

**Platforms**

● Anaconda