**Diabetic Prediction with Flask Framework**

Submitted in partial fulfillment of the requirements for the award of degree of

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE & ENGINEERING**



**Submitted By:**

**Badal Jain Shahe Faisal Aditya Jain**

**18bcs3234 18BCS3016 18BCS3365**

**Mentor Signature**

**Prof. Simarjit Singh Malhi**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**Chandigarh University, Gharuan**

**June202****1**

**Introduction**

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

**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.

|  |  |
| --- | --- |
| 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.

1. **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-

**PHASE -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

**Setting up the Model:-**

**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.

**REQUIREMENTS SPECIFICATIONS**

The requirements depends upon the various factors which are taken into consideration, for this project the requirements are based on algorithm used, programming language used and the data set on which the project is based on. Product Perspective The final application is coded on Python 3.0 so until the newer version of this application is developed which will be independent of the required environment, the product needs Python3.0 installed on the system. The user can use any IDE such as Spyder or IDLE or can use the open source distribution Anaconda, the application will run smoothly with all these features installed. 3.2 User characteristics The users of this system will be mostly doctors and the people who may encounter with some symptom of the disease which they are unaware off. It will be helpful for the doctors as the doctor can be reminded and cross check with the possible diseases (to overcome human errors like diligence, versatile, tiredness). This can also be helpful for the patients to find out the diseases, when no other help is possible and also can go to specialist of that particular disease instead of going to the general doctor and getting referred which results in time and money wastage and causes human effort. The user needs to input the list of symptoms which they are experiencing. The corresponding output will be the probable disease with the maximum likelihood.