DIABETES PREDICTION MODEL

**ABSTRACT**

Diabetes is a deadly chronic disease which affects entire body system harmfully. Millions of people are affected by this disease and a considerable number of patients die every year because of its side effects. A diabetic patient suffers from a high level of blood sugar in the body. Undiagnosed diabetes may cause the nerve and kidney damage, heart and blood vessel disease, slow healing of wounds, hearing impairment and several skin diseases. Early detection of diabetes is very essential to have a healthy life. The recent development of Machine Learning approaches solves this kind of critical problems. The main objective of this study is to present a Machine Learning based solution (Artificial Neural Network) to solve the above problem. And also, the technologies and approaches used in previous researches to predict diabetes have been reviewed with their accuracy levels. All the previous studies have used "Pima Indian Diabetes Dataset" (PIDD) as the dataset but this research is based on a newly collected dataset. The overall development process can be categorized into four major development phases namely data collection and pre-processing, statistical analysis, development of machine learning model and development of front-end. Artificial Neural Network model has been developed and deployed while the model provides more than 92% accuracy on the sample testing dataset.

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **Content** | **Page No.** |
| Introduction | 2 |
| Existing Method | 2 |
| Proposed method with Architecture | 3 |
| Methodology | 6 |
| Implementation | 9 |
| Conclusion | 10 |

**INTRODUCTION**

Diabetes mellitus is characterized by abnormally high levels of sugar (glucose) in the blood. When the amount of glucose in the blood increases, e.g., after a meal, it triggers the release of the hormone insulin from the pancreas. Insulin stimulates muscle and fat cells to remove glucose from the blood and stimulates the liver to metabolize glucose, causing the blood sugar level to decrease to normal levels. In people with diabetes, blood sugar levels remain high. This may be because insulin is not being produced at all, is not made at sufficient levels, or is not as effective as it should be. The most common forms of diabetes are type 1 diabetes (5%), which is an autoimmune disorder, and type 2 diabetes (95%), which is associated with obesity. Gestational diabetes is a form of diabetes that occurs in pregnancy, and other forms of diabetes are very rare and are caused by a single gene mutation. For many years, scientists have been searching for clues in our genetic makeup that may explain why some people are more likely to get diabetes than others are. "The Genetic Landscape of Diabetes" introduces some of the genes that have been suggested to play a role in the development of diabetes.

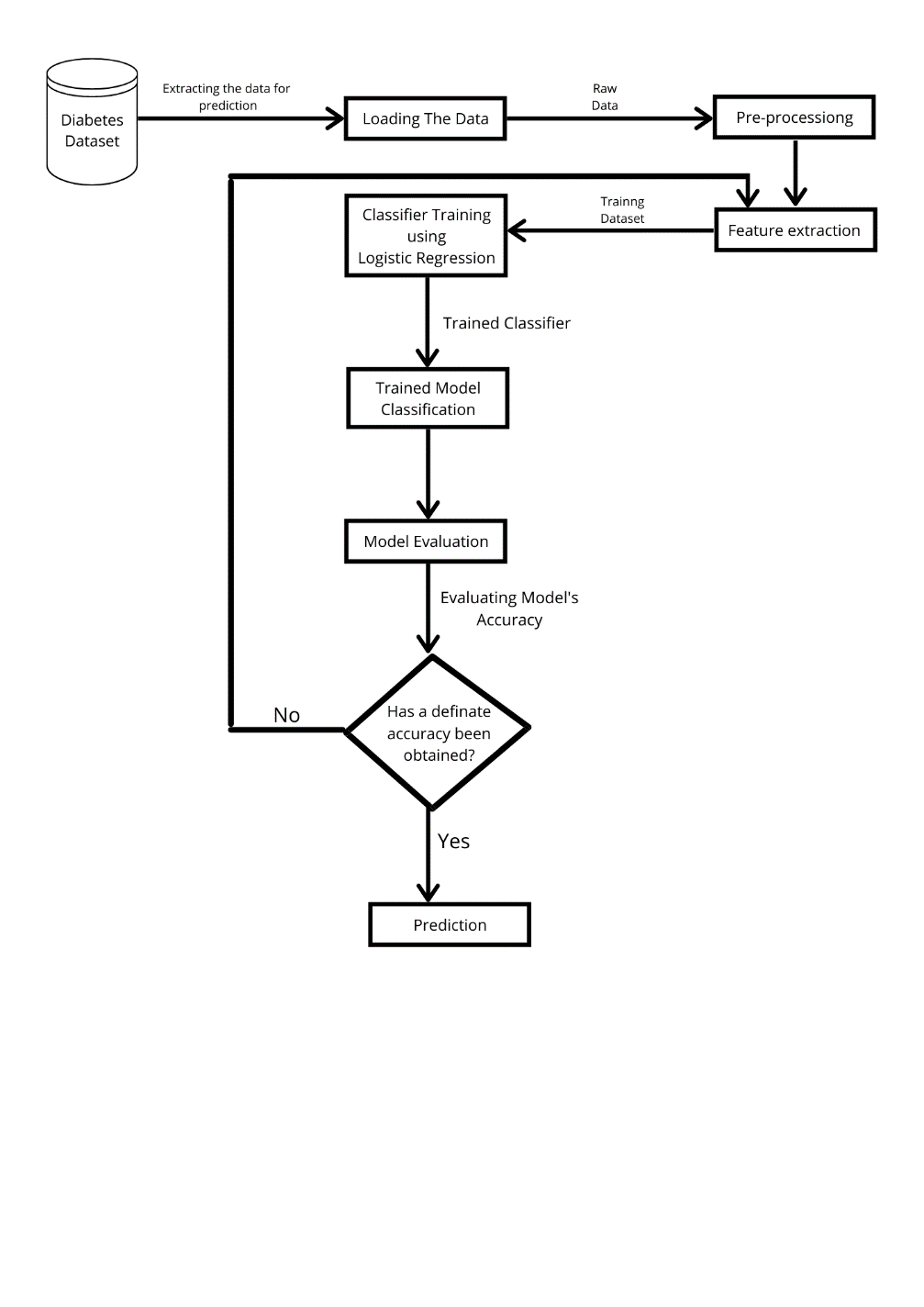
According to the research done by Katulanda and others (Katulanda et al., 2009), it is mentioned that by 2025 South East Asia will be the region with the highest number of diabetes patients in the world. Early identification of diabetes is important to have a healthy life. A healthy diet, regular exercises, maintaining average body weight, and eliminating alcohol and smoking are ways to prevent or delay the onset of type 2 diabetes. Diabetes can be controlled, and its consequences avoided or delayed with diet, physical exercises, medication and treatment for complications.

**EXISTING METHOD**

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 a 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.

**PROPOSED METHOD WITH ARCHITECTURE**

In view of the problem statement described in the introduction section, we propose a classification model with boosted accuracy to predict the diabetic patient. In this model, we have employed different classifiers like Decision Trees, KNN and Naïve Bayes. The major focus is to increase the accuracy by using resample technique on a benchmark well renowned diabetes dataset that was acquired from PIMA Indian Diabetes Dataset from UCI machine learning repository, which consists of eight attributes. The proposed framework is shown in Figure.



The framework is composed of the following important phases:

* Dataset Selection (PIMA Indian Diabetes Dataset)
* Data Pre-processing
* Feature extraction
* Learning by Classifier (Training) i.e., Logistic Regression
* Achieving trained model with highest accuracy
* Using trained model for prediction

The detail description of the components and the activities performed against each component is mentioned below

**Dataset Selection (Diabetes Dataset)**

In data mining and machine learning, the data selection is a process in which the most relevant data is selected from a specific domain to derive values that are informative and facilitate learning within that domain. In the study, we have used diabetes dataset having eight attributes that are used to predict the symptom of gestational diabetes in a female patient. This dataset was obtained from UCI repository and is a benchmark dataset. On the basis of historical information stored in the dataset such as age, body mass index, blood pressure and number of times pregnant the classifiers are trained for making decision whether diabetes test for an individual is positive or negative. The PIMA diabetes dataset only represents the Indian national females who are at least 21 years old. All of the attributes are of numeric-valued continuous data type. The attribute for class label is dichotomous variable (i.e., the binary response variable) within the PIMA dataset follows each tuple of the dataset. PIMA Indian Diabetes Dataset from UCI repository contains 768 instances. The PIMA dataset is converted from CSV to “. ARFF" format accepted by WEKA 3.6.13.

**Data preprocessing**

Data preprocessing is a technique of machine learning that comprises of converting raw data into an logical or comprehensible format. The real-world data is mostly incomplete, inconsistent, unreliable, redundant and having missing values etc. Data preprocessing is a conventional technique of eliminating such problems which are also known as noise. Preprocessing involves certain activities like data cleaning, integrating the data, transformation of data, data reduction, data discretization and data cleaning. Here the dataset is checked for duplicate values, missing values and type miss-matches etc. All these inconsistencies are eliminated from this dataset, in the phase called data preprocessing phase. It is important to clean the dataset before training it on a classifier in order to better learn the hidden patterns in the dataset. The set of pertinent feature vector fed to the classifier help it learn more accurately in a shorter span of time.

**Feature Extraction**

After setting the classification objectives, we apply principal component analysis (PCA) on the dataset to determine the most suitable set of attributes that can help achieve better classification. The set of attributes suggested by the PCA are termed as feature vector in this study. Feature reduction or dimensionality reduction will be benefitted us by reducing the computation and space complexity. Simple and more robust models should be developed, which are easier to understand and also saves the cost. Therefore, we applied PCA on the entire PIMA dataset within the WEKA tool. A threshold value of 0.21 is selected and all the attributes having range of greater than and equal to 0.21 is selected for further experimentation.

**Learning by Classifier (Training) i.e., Logistic Regression**

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either yes or no, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification

**Achieving trained model with highest accuracy**

This is the final step of prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score. Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples. It is given as

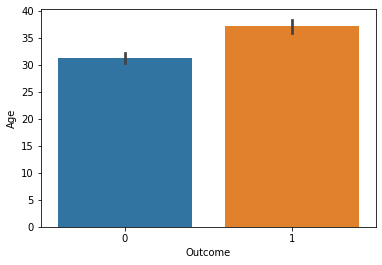
**Accuracy= Number of Correct Predictions / Total number of predictions Made**

**Using trained model for prediction**

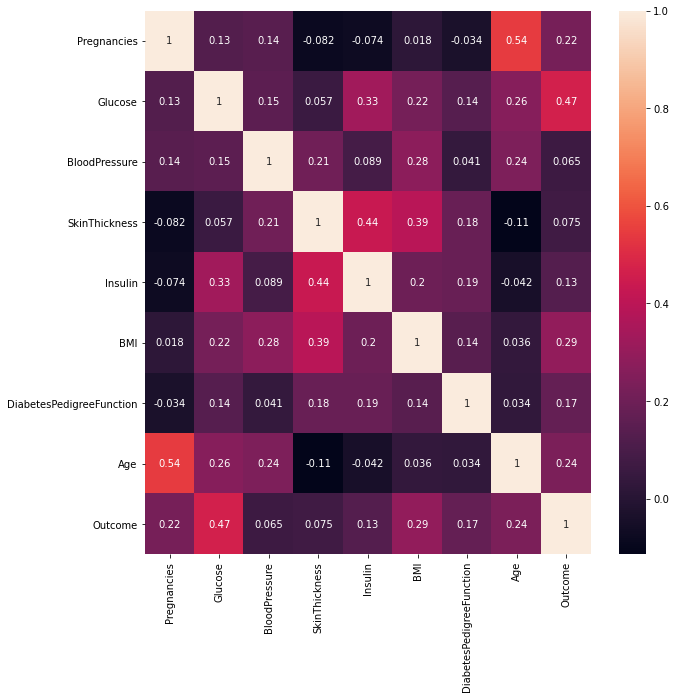
By using the python library named “Django” as well as various different libraries used for the prediction like “pandas”, “NumPy”, “sklearn” and “seaborn”; a web app has been set up which can be useful for the mass to predict the whether they are suffering from diabetes or not.

**METHODOLOGY**

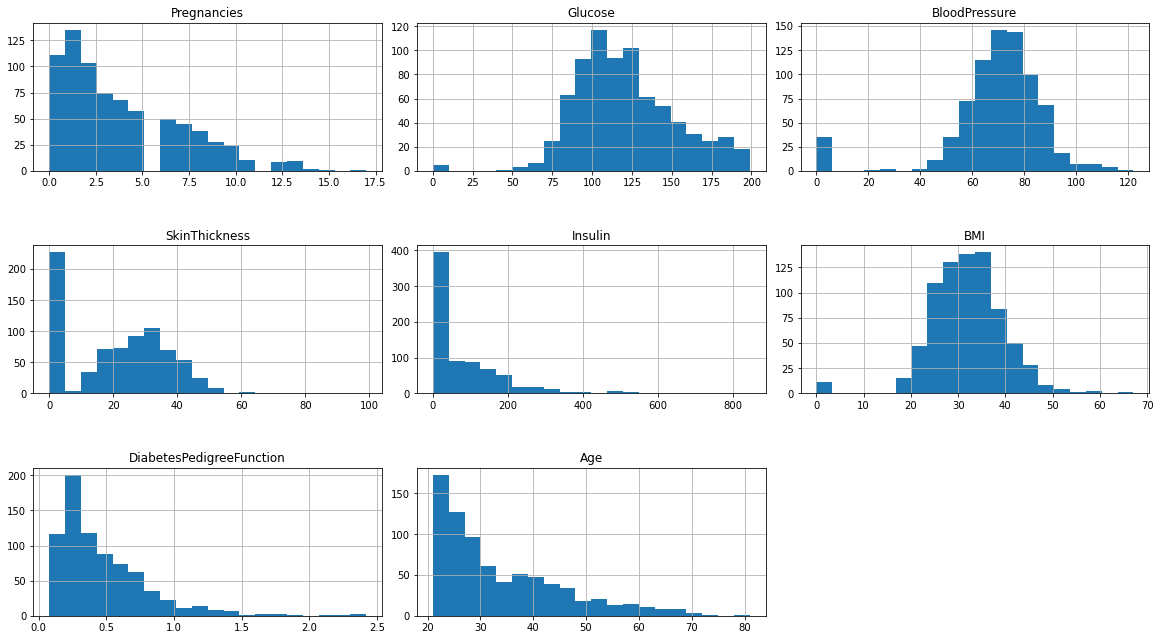
Firstly, a data has been fetched from the system named “diabetes.csv” and then number of data points as well as its features are looked out. The dataset consists of 9 features - 'Pregnancies', 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'. The diabetes data set consists of 768 data points, with 9 features each. “Outcome” is the feature we are going to use for prediction, 0 means No diabetes, 1 means diabetes. The figure is a plot of the mean age for each of the output classes. We can see that the mean age of people having diabetes is higher.



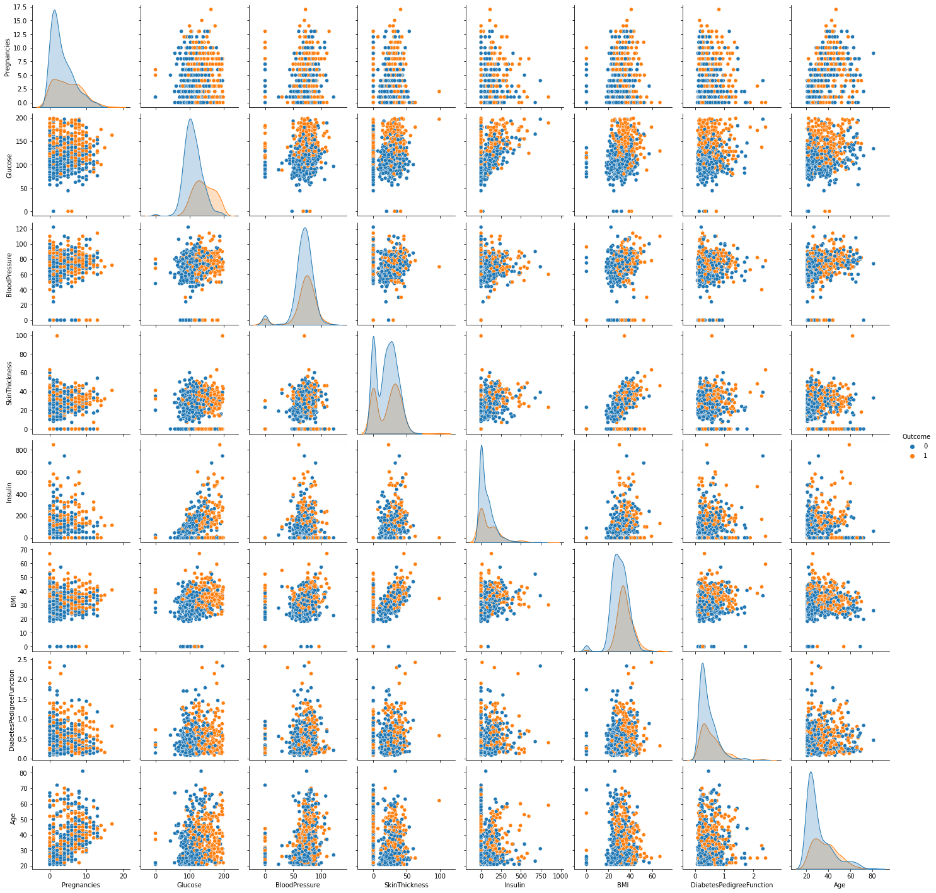
Also, it has been verified thoroughly that no void values are present within the dataset so that it will provide as possible as accurate results for the given values.



In the above heatmap, brighter colors indicate more correlation. As we can see from the table and the heatmap, glucose levels, age, BMI and number of pregnancies all have significant correlation with the outcome variable. Also notice the correlation between pairs of features, like age and pregnancies, or insulin and skin thickness.



None of the above feature is showing the symbol of consistent data as all are non-uniform. But if observed thoroughly we can find that pregnancy is bimodal while blood pressure and bmi are slightly uniform as we can overserve outliers. Infect, there are outliers present in each feature.



By the observing the paiplots respectively with each feature versus outcome, we can say that having diabetes is dependent on some features like age, insulin, glucose and BMI.

After this the whole dataset is divided for training and testing the model.

In this case, we are using 80% of the data for training purpose while the remaining 20% for testing purpose. Also, as the whole prediction is based on five main features like Age, Insulin, Glucose, BMI and Outcome, I have used two variables X and Y, where X is containing data from Insulin, Glucose, BMI and Age while Y consists of Outcome only. The model has been implemented using Logistic Regression which provides 80% accuracy in the model.

**IMPLEMENTATION**

**CODE:**

data = data[['Glucose','Insulin','BMI','Age','Outcome']]

X = data.drop("Outcome", axis=1)

Y = data['Outcome']

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2)

X\_train

### **Training the model**

model = LogisticRegression()

model.fit(X\_train, Y\_train)

### **Making Predictions**

predictions = model.predict(X\_test)

predictions

**OUTPUT**:

array([1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1,

0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1,

0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1,

0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0,

1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,

1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0,

1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1],

dtype=int64)

### **Evaluation**

accuracy = accuracy\_score(predictions, Y\_test)

accuracy

**OUTPUT:**

0.8051948051948053

**CONCLUSION**

Data mining plays an important role in various fields such as artificial intelligence (AI) and machine learning (ML), statistics and database systems. The core objective of this study is to enhance the accuracy of predictive model. The accuracy can be increase by improving the performance of the data, the algorithms or even by algorithm tuning. We enhance the accuracy by improving the data in preprocessing phase that really works well. Applying bootstrapping resampling technique on this PIMA dataset will increases the accuracy of almost all classifiers but the decision trees lead over others. It is also concluded that the accuracy of a model is highly dependent on the dataset. So, this technique works very well on PIMA diabetic dataset but may not guaranteed the same results on a different dataset. In future work includes it is plan to use further more advanced classifiers such as artificial neural networks (ANN), genetic algorithm (GA) and evolutionary algorithm (EA). The diabetes dataset considered in this study might not consider some other important factors that are related to gestational diabetes, like metabolic syndrome, family history, habit of smoking, lazy routines, some dietary patterns etc. The appropriate prediction model would want additional relevant data to make it more accurate. This would be accomplished by gathering diabetic patient’s datasets from various sources, to generate a better relevant prototype. This is a limitation of this research.