****

**Diabetes Prediction using machine learning**

**Machine Learning**

**Report Project**

**Supervised by D. Adnan Salman**

**Summer 2020/2021**

**ABSTRACT**

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Blood glucose is your main source of energy and comes from the food you eat. [Insulin](https://www.niddk.nih.gov/Dictionary/I/insulin), a [hormone](https://www.niddk.nih.gov/Dictionary/H/hormone) made by the [pancreas](https://www.niddk.nih.gov/Dictionary/P/pancreas), helps glucose from food get into your cells to be used for energy. Sometimes your body doesn’t make enough—or any—insulin or doesn’t use insulin well. Glucose then stays in your blood and doesn’t reach your cells.

Over time, having too much glucose in your blood can cause [health problems](https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems). Although diabetes has no cure, you can take steps to [manage your diabetes](https://www.niddk.nih.gov/health-information/diabetes/overview/managing-diabetes) and stay healthy.

Sometimes people call diabetes “a touch of sugar” or “borderline diabetes.” These terms suggest that someone doesn’t really have diabetes or has a less serious case, but every case of diabetes is serious.

**INTRODUCTION**

Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Big data analytics is the process which analyses huge data sets and reveals hidden information, hidden patterns to discover knowledge from the given data. Considering the current scenario, in developing countries like India, Diabetic Mellitus (DM) has become a very severe disease. Diabetic Mellitus (DM) is classified as Non-Communicable Disease (NCB) and many people are

suffering from it. Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2-5 million patients every year lose their lives due to diabetes. It is said that by 2045 this will rise to 629 million.[1] Diabetes Mellitus (DM) is classified asType-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human’s body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly. Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person. A technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made. Predictive analytics can be done using machine learning and regression technique. Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing patient care, optimizing resources along with improving clinical outcomes.[1] Machine learning is considered to be one of the most important artificial intelligence features supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case. Machine learning is considered to be a dire need of today’s situation in order to eliminate human efforts by supporting automation with minimum flaws. Existing method for diabetes detection is uses lab tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming. This paper focuses on building predictive model using machine learning algorithms and data mining techniques for diabetes prediction. The paper is organized as follows Section II-gives literature review of the work done on diabetes prediction earlier and taxonomy of machine learning algorithms. Section III-presents motivation behind working on this topic. Section IV gives diabetes prediction proposed model is discussed. Section V gives results of experiment followed by Conclusion and References.

**MOTIVATION**

There has been drastic increase in rate of people suffering from diabetes since a decade. Current human lifestyle is the main reason behind growth in diabetes. In current medical diagnosis method, there can be three different types of errors1. The false-negative type in which a patient is already a diabetic patient but test results tell that the person is not having diabetes. 2. The false-positive type. In this type, patient is not a diabetic patient but test reports say that he/she is a diabetic patient. 3. The third type is unclassifiable type in which a system cannot diagnose a given case. This happens due to insufficient knowledge extraction from past data, a given patient may get predicted in an unclassified type. However, the patient must predict either to be in diabetic category or non-diabetic category. Such errors in diagnosis may lead to unnecessary treatments or no treatments at all when required. In order to avoid or reduce severity of such impact, there is a need to create a system using machine learning algorithm and data mining techniques which will provide accurate results and reduce human efforts.

**This model has five different modules. These modules include**

* **Dataset Collection**

This module includes data collection and understanding the data to study the patterns and trends which helps in prediction and evaluating the results. Dataset description is given below This Diabetes dataset contains 768 records and 9 attributes.

|  |  |  |
| --- | --- | --- |
| Attributes | Type |  |
| Pregnancies | **int64** | **X1** |
| Glucose | **int64** | **X2** |
| Blood Pressure | **int64** | **X3** |
| Skin Thickness | **int64** | **X4** |
| Insulin | **int64** | **X5** |
| BMI | **int64** | **X6** |
| DiabetesPedigreeFunction | **float64** | **X7** |
| Age | **int64** | **X8** |
| Outcome | **int64** | **Y** |

Table 1. Dataset Information

* **Data Pre-processing**

This phase of model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. So we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

* **Clustering**

In this phase, we have implemented K-means clustering on the dataset to classify each patient into either a diabetic or non-diabetic class. Before performing K-means clustering, highly correlated attributes were found which were, Glucose and Age. K-means clustering was performed on these two attributes. After implementation of this clustering we got class labels (0 or 1) for each of our record.

**The way kmeans algorithm works is as follows:**

* Specify number of clusters K.
* Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
* Keep iterating until there is no change to the centroids. i.e., assignment of data points to clusters isn’t changing.
* Compute the sum of the squared distance between data points and all centroids.
* Assign each data point to the closest cluster (centroid).
* Compute the centroids for the clusters by taking the average of the all-data points that belong to each cluster.
* **Model Building**

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, neural network.

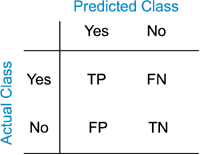
* **Evaluation**

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 Correct Prediction/Total Number of Prediction Made)**

Confusion Matrix- It gives us gives us a matrix as output and describes the complete performance of the model.



[This Photo](http://en.wikipedia.org/wiki/File:Binary_confusion_matrix.png) by Unknown Author is licensed under [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

TP: True Positive

FP: False Positive.

FN: False Negative.

TN: True Negative.

Accuracy for the matrix can be calculated by taking average of the values lying across the main diagonal. It is given as-

**Accuracy = (TP+FN)/N**

N: Total number of samples**.**

**F1 score**-It is used to measure a test’s accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-

**F1 = 2 \* (1/ (1/precision) + (1/recall))**

F1 Score tries to find the balance between precision and recall.

**Precision**: It is the number of correct positive results divided by the number of positive results predicted by the classifier. It is expressed as**:**

**Precision = TP/ (TP+FP)**

**Recall**: It is the number of correct positive results divided by the number of all relevant samples. In mathematical form it is given as-

**Recall = TP/ (TP+FN)**

* **RESULTS**

After applying various Machine Learning Algorithms on dataset, we got accuracies as mentioned below. Logistic Regression gives highest accuracy of 96%.

|  |  |
| --- | --- |
| Algorithms | Accuracy |
| Logistic Regression | **73.37%** |
| Random Forest | **75.97%** |
| KNN | **72.72%** |
| SVC | **73.37%** |
| Gaussian Naïve Bayes | **70.77%** |
| Decision Tree | **70.12%** |
| Neural network with 150 epoch | **78.9%** |
| Neural network with 1000 epoch | **99.87 %** |

Table 2. Accuracy Table

**Confusion Matrix for Logistic Regression is given below:**

|  |  |  |
| --- | --- | --- |
|  | Diabetic | Non-Diabetic |
| Diabetic | **87** | **13** |
| Non-Diabetic | **28** | **26** |

Table 3. Confusion Matrix for **Logistic Regression**

**Confusion Matrix for SVM is given below:**

|  |  |  |
| --- | --- | --- |
|  | Diabetic | Non-Diabetic |
| Diabetic | **88** | **12** |
| Non-Diabetic | **29** | **25** |

Table 3. Confusion Matrix for **SVM**

**Confusion Matrix for random forest is given below:**

|  |  |  |
| --- | --- | --- |
|  | Diabetic | Non-Diabetic |
| Diabetic | **84** | **16** |
| Non-Diabetic | **21** | **33** |

Table 3. Confusion Matrix for **random forest**

**The different performance measures that are being compared are Accuracy, F1-Score, Precision and Recall. The Confusion matrix for the algorithm with highest accuracy is mentioned in Table 3.  
Visualization of these accuracies helps us to understand variations among them clearly.**

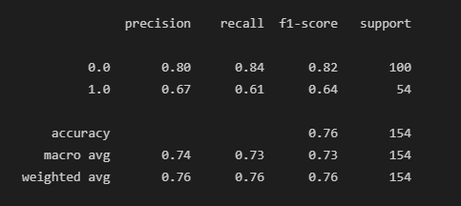
**Try to fix or develop the data in order to improve the accuracy :**

Attempting to improve the data in order to improve the Accuracy by replacing all zero values ​​to the values ​​of the mean for every features.

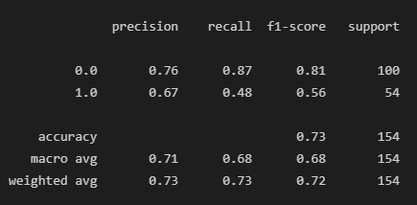
|  |  |  |
| --- | --- | --- |
| Algorithms | Accuracy with normal Dataset | Accuracy with develop Dataset |
| Logistic Regression | **73.37%** | **80.51%** |
| Random Forest | **75.97%** | **79.22%** |
| SVC | **73.37%** | **80.51%** |

Table 4. before and after improving the data

**Classification report before improving the data**

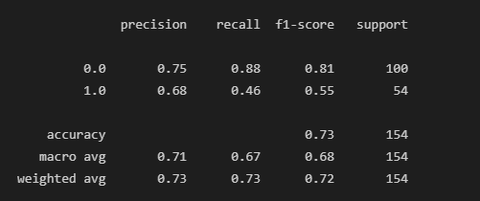
****

**Random Forest**

****

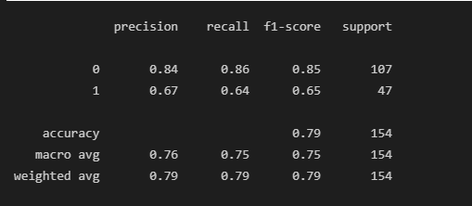
**Logistic**

**Regression**

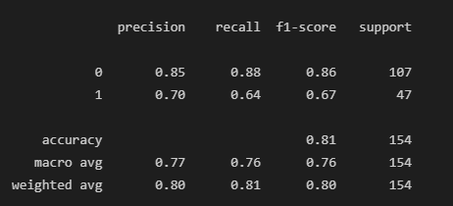
****

**SVC**

**Classification report after improving the data**

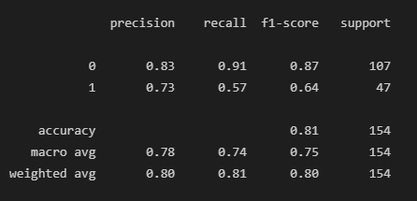
****

**Random Forest**



**Logistic**

**Regression**



**SVC**

prepared by:

* Hussein A Barham
* Roaa Arabasi
* Abdelrahman marea