**Predictive Machine Learning Model For Diabetes In Suspects**

by

**Uwakwe Chikwado**

Sunmitted to:

**Dr. Onyema Osuagwu**

*A Project Submitted To Morgan State University, Maryland In Partial Fulfillment Of The Requirements For The Degree Of PhD Of Electrical And Computer Engineering.*

**7th December, 2021.**

TABLE OF CONTENTS:

1. Abstract-------------------------------------------------------------------------------3
2. Introduction--------------------------------------------------------------------------4
3. Importance---------------------------------------------------------------------------5
4. Methodology-------------------------------------------------------------------------6
5. Description of the dataset……………………………………………….6
6. Train/Test………………………………………………………………….7
7. Libraries…………………………………………………………………….8
8. Statistical representation …………………………………………………9
9. Correlation of the features in the dataset……………………………...10
10. Trustworthy of machine learning……………………………...…11
11. Conclusion------------------------------------------------------------------12
12. References------------------------------------------------------------------13

**ABSTRACT**

Diabetes is a disease which has ravaged the lives of many people across the globe. Any diabetic patient is prone to kidney disease, heart disease, stroke and eye problem etc therefore early diagnosis of diabetes and its treatment is necessary to avoid collapse of critical organs of the body. The primary objective of this project is to predict the diabetes stage in suspects, since it will enable the individual take some precautionary measures towards slowing down the process of diabetes development or outright preventive measures. The PIMA Indian diabetes dataset was downloaded from [www.kaggle.com](http://www.kaggle.com) and the Logistic Regression, KNN, SVM and Random Forest classifier were used to predict whether a suspect has diabetes or not. The SVM demonstrated strong and better test score accuracy with 83.12%.

**INTRODUCTION**

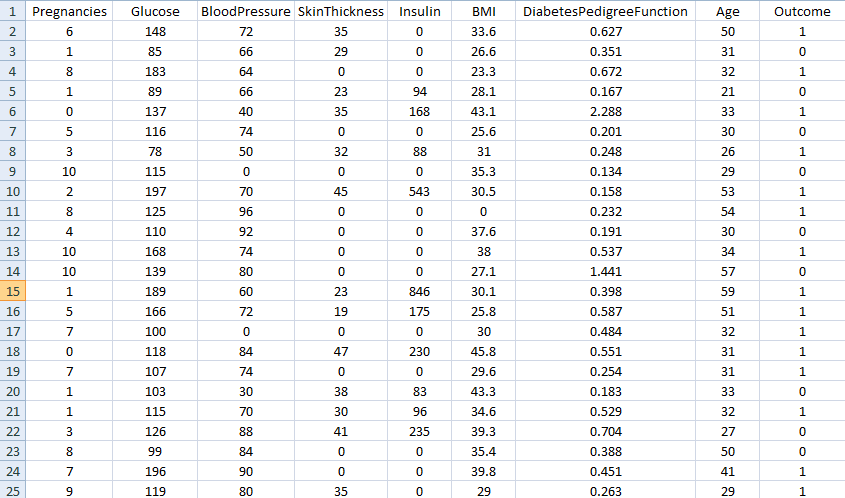
According to (Alaa Khaleel & Al-Bakry, 2021) Diabetes can be described as a metabolic disorder caused by a prolonged high level of glucose in the blood. Normal glucose level in the human body ranges from 70 to 99 mg per deciliter. If the glucose in the human body exceeds 126mg/dl, it shows diabetes, while glucose level between 100 – 125 mg/dl shows that such a person is pre-diabetes (Khanam & Foo, 2021). When the glucose in the blood is very high, it can give rise to complications such as heart disease, kidney failure, stroke and nerve damage etc (Tigga & Garg, 2020). However, (Mujumdar & Vaidehi, 2019) claims that approximately 2-5 million people die every year from diabetes complications, while over 400 million people live with this deadly disease in the world today and with the prospect to escalate to about 629 million patients by the year 2045. (Khanam & Foo, 2021) Assert that diabetes has no outright cure and therefore it is essential that it is detected early enough. It is important to diagnose diabetes whether at the early stage or later stage, when this is done, proper measures and treatment will be taken to manage the patient’s health to avoid total collapse or failure of critical organs of the patient (Derevitskii & Kovalchuk, 2020). But it is more important to identify or diagnose a suspect and determine whether the patient will develop the disease or not in the nearest future, say in the next one or two years. Early diagnoses will enable the medical personnel to guide and manage the suspect effectively to avoid developing the disease or probably prolong its development to a later period (Tigga & Garg, 2020). Hence, this project will focus on diagnosis and determining whether or not a suspect has diabetes.

**IMPORTANCE OF PREDICTING DIABETES**

In line with (Chajewski & Nichols, 2009) as the cost of healthcare increases, rather than curative measures, it is important to consider disease prevention. In between normal blood glucose and hyperglycemic plasma levels there is a stage called the pre-diabetes, people who fall under this category either have impaired fasting glucose or impaired glucose tolerance or both (Derevitskii & Kovalchuk, 2020). These people stand the risk of moving from pre-diabetes to diabetes. However, identifying this class of patients early and initiating the corresponding precautionary measures by way of change of lifestyle or use of preventive medicine can retard or prevent the development of diabetes in such a person (Tigga & Garg, 2020).

**METHODOLOGY**

Going forward in this work, the PIMA Indian diabetes dataset has been downloaded from kaggle data repository ([www.kaggle.com](http://www.kaggle.com)). Analysis has been done on Pima Indian diabetes dataset (PIDD) before putting it to work. Age, BMI (Body Mass Index) and blood glucose concentration are some of the essential factors in diagnosis of diabetes and the **Random Forest** machine, Support Vector Machine, Logistic Regression and K Nearest Neighbor learning algorithm were employed here to do the prediction.



**DESCRITION OF THE DATASET**

The dataset has 8 independent variables (features) and one outcome (Label). The independent variables are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and the target variable (Outcome).

The independent variables and their meaning:

Pregnancies This is the number of times pregnant

Glucose plasma glucose concentration a 2 hour in an oral glucose tolerance

Blood Pressure Diastolic blood pressure (mmHg)

Skin Thickness Triceps skin fold thickness (mm)

Insulin 2 hour serum insulin (mu U/ml)

BMI Body Mass Index (weight in kg/ height in m^2)

Diabetes Pedigree Function if anybody has or had diabetes in the family

Age Age (years)

Outcome The class variable (0 or 1)

**TRAIN / TEST**

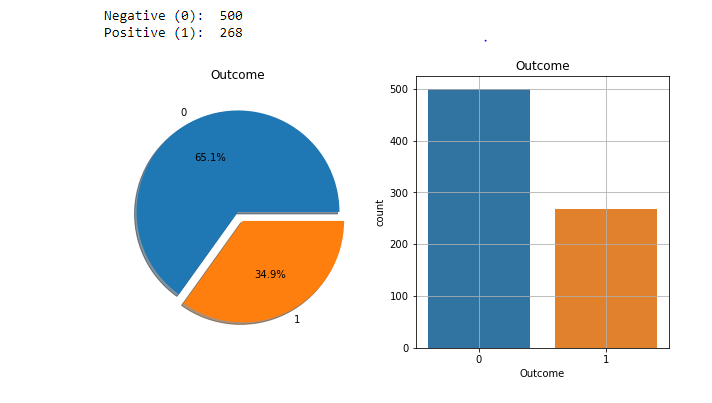
The 80% of the dataset was used to train the algorithms to predict whether or not a patient has diabetes and the 20% of the remaining dataset was used for validation.

**LIBRARIES**

The basic libraries imported for this project include: Numpy, Pandas, matplotlib, seaborn, scikitlearn (sklearn).

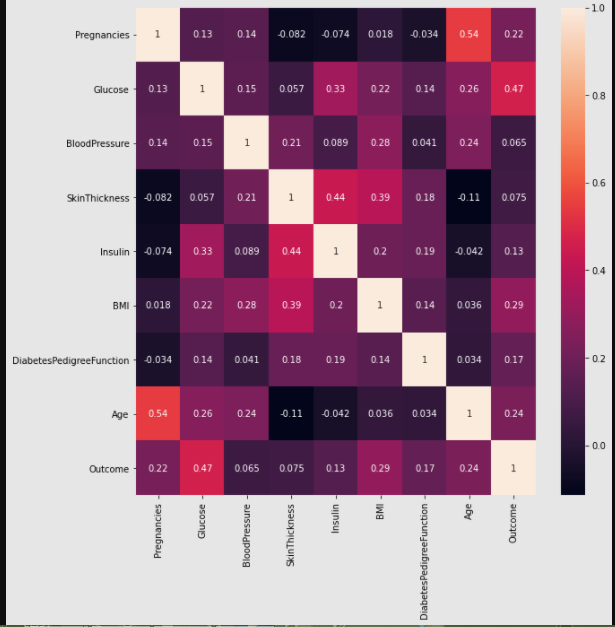
After importing the libraries, the dataset was loaded. Here, the dataset was explored for a better understanding. To understand the variables, the head of the dataset was looked at and by default; python will display the first top 5 of the dataset, also python will display the last 5 records of the dataset as the tail. Python can also display randomly any number of the dataset if needed. The shape of the dataset is to know the number of rows and columns, which in this case there are 768 rows and 9 columns. Also, to know the list of types of the columns, whether it is integers, floating numbers etc, and the command “df.dtypes” was used. Here, to check for null values in the dataset, the command “df.info ()” was used and in this case there was no null values. For statistical summary of the data which include the mean, minimum value, maximum value, standard deviation and percentiles etc we used the command “df.describe ()”. The dataset was cleaned of duplicates and null values. The null values were checked using the “df.isnull ().sum ()” command and there are no null values in the dataset. But the Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age returned minimum values and the number of zero values were checked and replaced with the corresponding mean values.

**STATISTICAL REPRESENTATION**



The value of the target variable of the dataset which is represented by the ‘Outcome’ is in ‘0’ and ‘1’ which makes it a binary classifier. Here, there are two sets of people, those who have diabetes which is represented by ‘1’ and those who do not have diabetes which is also represented by ‘0’. The histogram above shows that out of 768 cases, 500 persons do not have diabetes while 268 people have diabetes. The pie chart demonstrates the record in percentages where the 65.1% represents the 500 people who do not have diabetes and the 34.9% represent the 268 people who have diabetes.

**CORRELATIONS OF THE FEATURES IN THE DATASET**



From the correlation heatmap, the high correlation between the Outcome and (pregnancies, glucose, BMI and age) is obvious. Hence, these 4 feature variables can be accepted as input from the user and the Outcome will be predicted.

**TRUSTWORTHINESS OF MACHINE LEARNING**

* Generalizability: For the algorithm to be trustworthy, the model must predict accurately, although trustworthiness is not limited to accuracy.
* Robustness: this is to be sure that the model can withstand noisy and adversarial data. That is, in case there is some noise in the training data, we expect that the behavior of the model won’t be affected.
* Privacy: since the machine learning algorithms are trained on sensitive data, can we trust the models to have access to those data? And do we trust that the model will not leak the information during the time is being used.
* Interpretability: no matter the complexity of the model, we should be able to interpret the decisions made by the model. This can help us establish trust for the algorithm.
* Fairness: the objective here is to train models that understand fairness and also train models which are fair especially in some sensitive issues such as gender and race etc.

**CONCLUSION**

There are many people in the world today living with diabetes and many other people are at the risk of pre-diabetes. This project is to predict whether or not an individual has diabetes based on their daily lifestyles and family history. There are four algorithms that were employed; they include the random forest classifier, support vector machine, logistic regression and the K Nearest Neighbor. These algorithms were used on PIMA Indian diabetes dataset, to predict whether a suspect has diabetes or not. The results of the prediction, shows that the support vector machine has the best prediction of 83.12%, followed by the Random Forest classifier which has 79.2% accuracy, then Logistic Regression 77.3% and finally the K Nearest Neighbor which predicted with 74.67% accuracy. However, identifying the diabetic patients early and initiating the corresponding precautionary measures by way of change of lifestyle or use of preventive medicine can retard or prevent the development of diabetes in such a person.

# BIBLIOGRAPHY

Alaa Khaleel, F., & Al-Bakry, A. M. (2021). Diagnosis of diabetes using machine learning algorithms. *Materials Today: Proceedings* .

Chajewski, O. S., & Nichols, J. H. (2009). The importance of testing for pre-diabetes-Using the right tool. *US Endocrinology* .

Chaki, J., Thillai Ganesh, S., Cidham, S. K., & Ananda Theertan, S. (2020). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University - Computer and Information Sciences* .

Derevitskii, I. V., & Kovalchuk, S. V. (2020). Machine Learning-Based Predictive Modeling of Complications of Chronic Diabetes. *Procedia Computer Science* .

Jashwanth Reddy, D., Mounika, B., Sindhu, S., Pranayteja Reddy, T., Sagar Reddy, N., Jyothsna Sri, G., et al. (2020). Predictive machine learning model for early detection and analysis of diabetes. *Materials Today: Proceedings* .

Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express* .

Lama, L., Wilhelmsson, O., Norlander, E., Gustafsson, L., Lager, A., Tynelius, P., et al. (2021). Machine learning for prediction of diabetes risk in middle-aged Swedish people. *Heliyon* .

Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science* .

Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. *Procedia Computer Science* .

Xiao, M.–X. X., Lu, C.–H. H., Ta, N., Wei, H.–C. C., Yang, C.–C. C., & Wu, H.–T. T. (2022). Toe PPG sample extension for supervised machine learning approaches to simultaneously predict type 2 diabetes and peripheral neuropathy. *Biomedical Signal Processing and Control* .