**Diabetes Data Set**

NAME: IHECHILURU WINNER CHUKWUEMEKA

EMAIL: ihechiluruwinner@gmail.com

Linked-in:<https://www.linkedin.com/in/ihechiluru-winner/>



**PROBLEM STATEMENT:**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

**BRIEF**

In the course of this project, we’ll be using Machine Learning algorithms followed up as classification and regression to predict whether or not a person has diabetes. In classification, we have used KNN and Decision tree algorithms and in regression, we have used logistic regression.

These predictions are based on information about patients provided in the dataset such as the patient’s age, blood pressure, glucose levels, Body Mass Index and so on.

Since we are trying to predict whether or not a patient has diabetes, our output variable, “outcome” is divided into 0’s and 1’s signifying that the patient either has or does not have diabetes.

The dataset that was given to us for this project was collected by the “National Institute of Diabetes and Digestive and Kidney Diseases and all patients here are females at least 21 years of age of Pima Indian heritage.

**2. Algorithms Used:**

**KNN Algorithm:** K- Nearest Neighbour [Classification] K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. K-NN algorithm stores all





the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a good suite category by using K- NN algorithm. It comes under the supervised learning technique.

**Decision Tree Algorithm:** [Classification] Decision tree is a supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

**Logistic Regression:** This is a linear model for classification rather than regression. It is a supervised learning classification algorithm used to

predict the probability of a target variable. The nature of the target or

dependent variable is dichotomous, which means there would be only

two possible classes. Logistic regression predicts the output of a

categorical dependent variable.

**3. Tools Used:**

Tools used in the major project are:

a) Google Colab

b) Pandas Library

c) Seaborn Function to Plot Graph, and

d) Sklearn – Machine Learning website to import all the algorithms,

confusion matrix and accuracy score.

**DATA VISUALIZATION INSIGHTS:**

1. Pregnant Women have a higher risk of having diabetes. Further research shows that it might be due to the fact that Image result for does pregnancy cause diabetes all pregnant women have some insulin resistance during late pregnancy. Most pregnant women can produce enough insulin to overcome insulin resistance, but some cannot. These women develop gestational diabetes.
2. Patients with higher Glucose levels have higher risks of getting diabetes.
3. Abnormally High Insulin Levels could also result in diabetes. Further research shows that Hyperinsulinemia, which means the amount of insulin in your blood is higher than what's considered normal. Alone, it isn't diabetes. But hyperinsulinemia is often associated with type 2 diabetes. Insulin is a hormone that's normally produced by your pancreas, which helps regulate blood sugar.
4. To summarize the rest of the points, high BMI, blood pressure, Diabetes Pedigree Function and Older People also stand higher risks of having diabetes.

All information in this report is based on analysis and research and therefore is subject to further review. For more information on Diabetes Kindly contact your doctor or a verified medical practitioner.

**4. Conclusion:**

Keeping the random\_state=0 will never change the scores at multiple runtimes.

• KNN Algorithm:

o Training Score= 0.79

o Testing Score= 0.75

o Accuracy Score= 0.75

• Decision Tree Algorithm:

o Training Score= 1.00

o Testing Score= 0.77

o Accuracy Score= 0.77

• Logistic Regression:

2



o Training Score = 0.77

o Testing Score = 0.82

o Accuracy Score = 0.82

So naturally, the best model for precision is logistic regression, for identification of underlying problem creating factors k nearest neighbour is suggested as it gives us a clear cut idea of what is the root cause if analyzed properly.

Kindly access the link below for my google colab notebook

Google Colab

You can also use the link below to access the file if it doesn’t go through

https://colab.research.google.com/drive/1RzhAKYQvEEnq\_0Zb3dcEYhItvlIfFuK8?usp=sharing

Regards,

Ihechiluru Winner Chukwuemeka.

3