Assignment -3

Predicting Diabetes based on the health information using KNN

Name : Hari teja Narayanabhatla

Panther ID: 002566679

Group : MSA

**Business Question:** Most of the people in society are unaware of how dangerous diabetes is

in the long run of their life. There is no proper indication of how this starts but there are

some noticeable features in the body that can indicate whether a person is diabetic or non-

diabetic. Most of the organizations are developing a model to diagnostically predict

whether a patient has diabetes, based on certain diagnostic measurements.

**Dataset Overview:**

The datasets consist of several medical predictor variables and one target variable, Outcome.

Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level,

age, and so on.

The only categorical variable in the dataset is the target variable , the Outcome column. In order to

maintain the data as normal as possible, no man or women of age less than 21 years are kept in this

dataset. There are no null values in the data.

**Data Visualization:**

First tried to check if the data is balanced properly or not. Hence using pandas and

matplotlib tried to visualize the data. Chart, bar chart

Description automatically generated

Fig 1: Shows the distribution of the classes in the given dataset.

There exists no null value or unknown value for any column in the given dataset. Hence, it’s

a clear dataset and we went ahead with further processing.

**Data Correlation**

I tried to find if there is any correlation between any of the columns and if yes, we can

remove them the so that model can give good results.

**Calendar

Description automatically generated**

Fig 2: It gives us information that there are some columns that are correlated to each other.

**Effective Neighbors Selection:**

In order to select a number of feasible neighbors for the given problem, we have taken a

range of neighbors from which we will visualize and select which one is giving us better

accuracy.

The major work of KNN is to find the distances between a query and all the examples in the

data, selecting the specific number neighbors(K) closest to the query.

Chart, line chart

Description automatically generated

Fig 3: We have trained the model on different ranges of neighbors and found that the best accuracy is given by 7 neighbors.

**Conclusion:**

Using KNN and identifying the best number of neighbors we had predicted which person is

having diabetes and which are not. I got an accuracy of 74% based on the training data with

the best neighbors being 7 for this data.