CAPSTONE PROJECT

**Topic:** Predicting diabetes using Machine Learning

**Dataset used:** Pima Indians Diabetes dataset from kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database)

Introduction

In this project I have used the Pima Indians diabetes dataset in kaggle, to perform a descriptive analysis on the dataset and develop a machine learning model to predict whether or not a patient has diabetes.

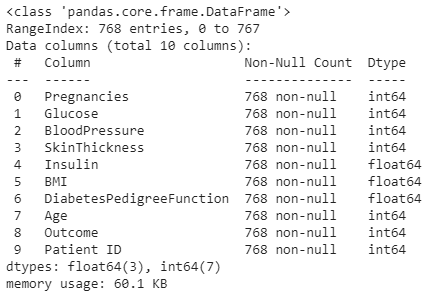
Background information of the dataset:

This data set was originally from the National institute of Diabetes and digestive and kidney diseases. In particular all patients in the dataset are females at least 21 years old of Pima Indian heritage.

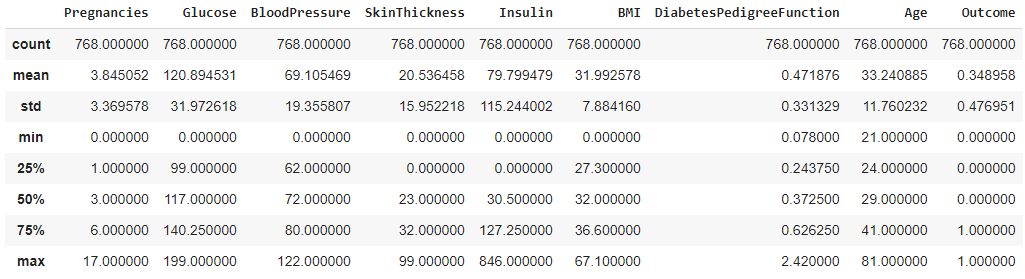
Descriptive statistic:

This dataset has total 768 records and 9 features/attributes. All variables are numerical variables.

Details of features as follows,



Descriptive statistics of features



“Outcome” is the dependent variable(Y) which gives the indication if the patient has diabetes or not.

0🡪 healthy

1 🡪 Diabetes

Rest of the variables are all independent variables (X).

Methology

1. **Pre-processing data**

First of all it is very important that you study your data and take necessary steps to transform raw data into much cleansed and to an understandable format, if not you will be training your model with incorrect data.

I followed following steps for data preprocessing.

* First I checked if the dataset had any null records. This particular dataset didn’t have any null values.
* Checked if any of the patient records are duplicated within the dataset.
* When reviewing the dataset, I found out that some of the features (especially Insulin & Skinthickness) had zero vales, assume it’s due to not having those particular information about the patient, in such cases I have treated these values by replacing them with the median value. (Main reason why we replaced it with median instead of Mean is because, if there are outliers mean value could be impacted).
* After visualizing data plots, it was clear that some of the features had outliers.
* To address the above issue, I selected data in the IQR (Each feature picked all data between the lower & the upper limit range).

*IQR= Q3 (75% percentile)-Q1 (25% percentile)*

*Lower limit = Q1 – 1.5\*IQR*

*Upper limit = Q3 + 1.5\*IQR*

1. **Selecting features**

* I used the “corr” function in python to understand the correlation between the dependent & independent variables.
* When compared, two variables (“Insulin” &”Blood pressure”) had a relatively low correlation with dependent variable (Y=Outcome). Accordingly I decided to remove these features when training the model as it’s not having a strong impact on the final prediction.
* I have used the seaborn heatmaps to visualize the correlation.

1. **Train test split**

* I used train test split function to split the full data set for training & testing/Validation.
* Used 80% of the data to train the model & 20% of the data to test the model prediction.

1. **Building the model**

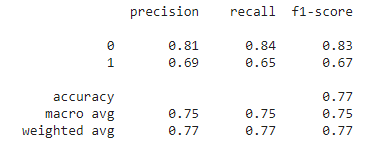
* I used supervised machine learning approach to build the model.
* Since this is a binary classification problem, I build two models (logistic regression & random forest classifier)
* Based on the results, will proceed with the model with highest accuracy.

1. **Model evaluation**

* I evaluated the model performance by comparing the test actuals against the test model predictions.
* I used the confusion matrix, Accuracy, Precision, recall, F1 functions to evaluate the model performance.

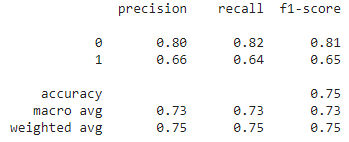
1st model performance (logistic regression)-:

Accuracy -: 0.772727



2nd model performance (random forest classifier)-:

Accuracy -: 0.75324



Conclusion

From the ML foundation course, I learned lot about machine learning concepts and how to build machine learning models. I used this knowledge and additional information I gathered from external sources to complete this project.

For my capstone project, I gathered data from a public data source (kaggle) and build a machine learning model to predict if a patient has diabetes by simply considering the features ex-: glucose, BMI, age..etc.

Despite Time & resource limitations, I was able to fine tune the model to bring the accuracy levels of the predicted data up to 75%-77%, however believe this model can be improved furthermore to predict more accurately.