

Data Science Capstone Project

**Building Machine Learning for Diabetes data**

Prepared By:

Abdallah Shanaah

Feb/ 2021

# Abstract

This project addresses the use of a dataset that is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The aim of this project was to predict whether a person suffers from diabetes through his clinical diagnosis, such as insulin levels, blood pressure, glucose level, age, etc., and what may be the symptoms most related to diabetes from the symptoms, after processing the information, analyzing it and try to build more then

one classifier such as Logistic Regression, K-Nearest Neighbors and Random Forest, do some comparison between them to find the best classifier I reached that K-Nearest Neighbors is the best classifier with accuracy at the level of 74.8% was achieved.

Table of Contents

[Abstract 2](#_Toc62690947)

[Introduction 4](#_Toc62690948)

[Methodology 5](#_Toc62690949)

[Used Python Libraries 5](#_Toc62690950)

[Datasets and Input 5](#_Toc62690951)

[Preprocessing 7](#_Toc62690952)

[Data Cleaning 8](#_Toc62690953)

[Data Distribution 9](#_Toc62690954)

[Additional features 10](#_Toc62690955)

[Results 14](#_Toc62690956)

[Conclusion 15](#_Toc62690957)

[References 16](#_Toc62690958)

# Introduction

At the end of my study period, he started searching for specializations that I could specialize in in my field of study in those days. Building applications and web pages were the most popular specialties, so I began to learn in this field and I was able to progress well in the field of full stack developer, but I thought that I could be more creative. In other disciplines, I began to research until I got acquainted with the specialization of the data specialist and began to get to know more about this field and joined courses to enable me to learn better and faster

In this field, the information is the basis of the entire work. I found a set of data from the National Institute of Diabetes and Digestive and Kidney Diseases of clinical diagnoses such as insulin percentage, skin thickness, blood pressure and others for a number of patients trying to predict whether or not they have diabetes.This can help treat the disease at the beginning of its diagnosis

**Statement of the Problem**

The goal of this project is to predict, by analysis of patient data, whether the patient has diabetes

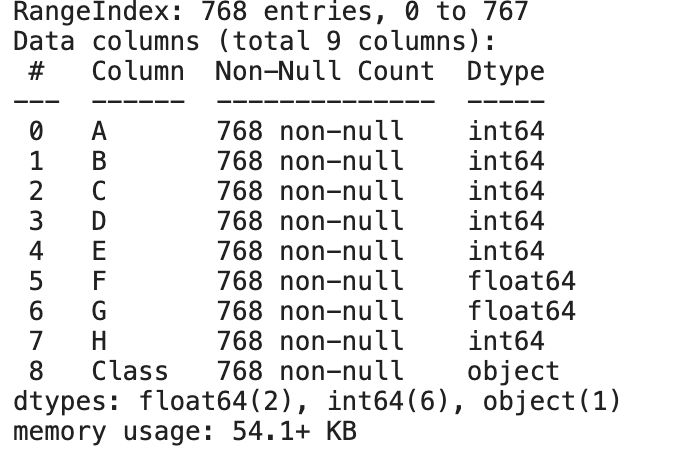
# Methodology

## Used Python Libraries

Data was pre-processed using *Pandas* and *NumPy* libraries and the learning/validating process was built with *scikit-learn and keras*. Plots were created using *Seaborn,* *matplotlib, mlxtend and missingno*.



## Datasets and Input

****

The input data CSV files:

* Dataset.csv

Which contains the following columns

**A-Pregnancies**: (Number of times pregnant)

**B- Glucose:** Plasma glucose concentration 2 hours in an oral glucose tolerance test

**C-Blood Pressure:** Diastolic blood pressure (mm Hg)

**D- Skin Thickness:** Triceps skin fold thickness (mm)

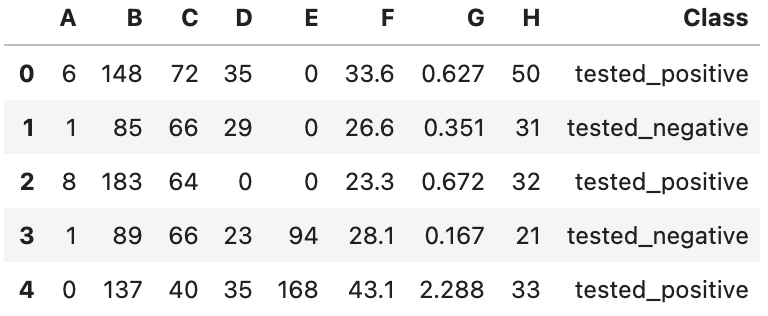
**E- Insulin:** 2-Hour serum insulin (mu U/ml)

**F- BMI:** Body mass index (weight in kg/ (height in m) ^2)

**G-Diabetes Pedigree Function:**

**H- Age:** (years)

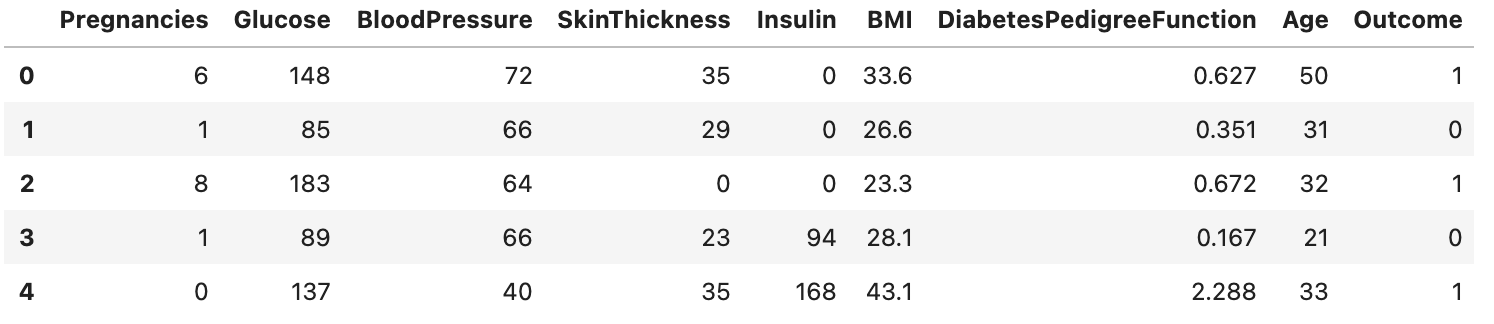
**CLASS -** **Outcome**: Class (tested\_negative, tested\_positive)



Here, the’ Class or Outcome’ is the target class, given the other column, ‘Class or Outcome’ defines whether the given patient is having diabetes(positive) or not(negative)

## Preprocessing

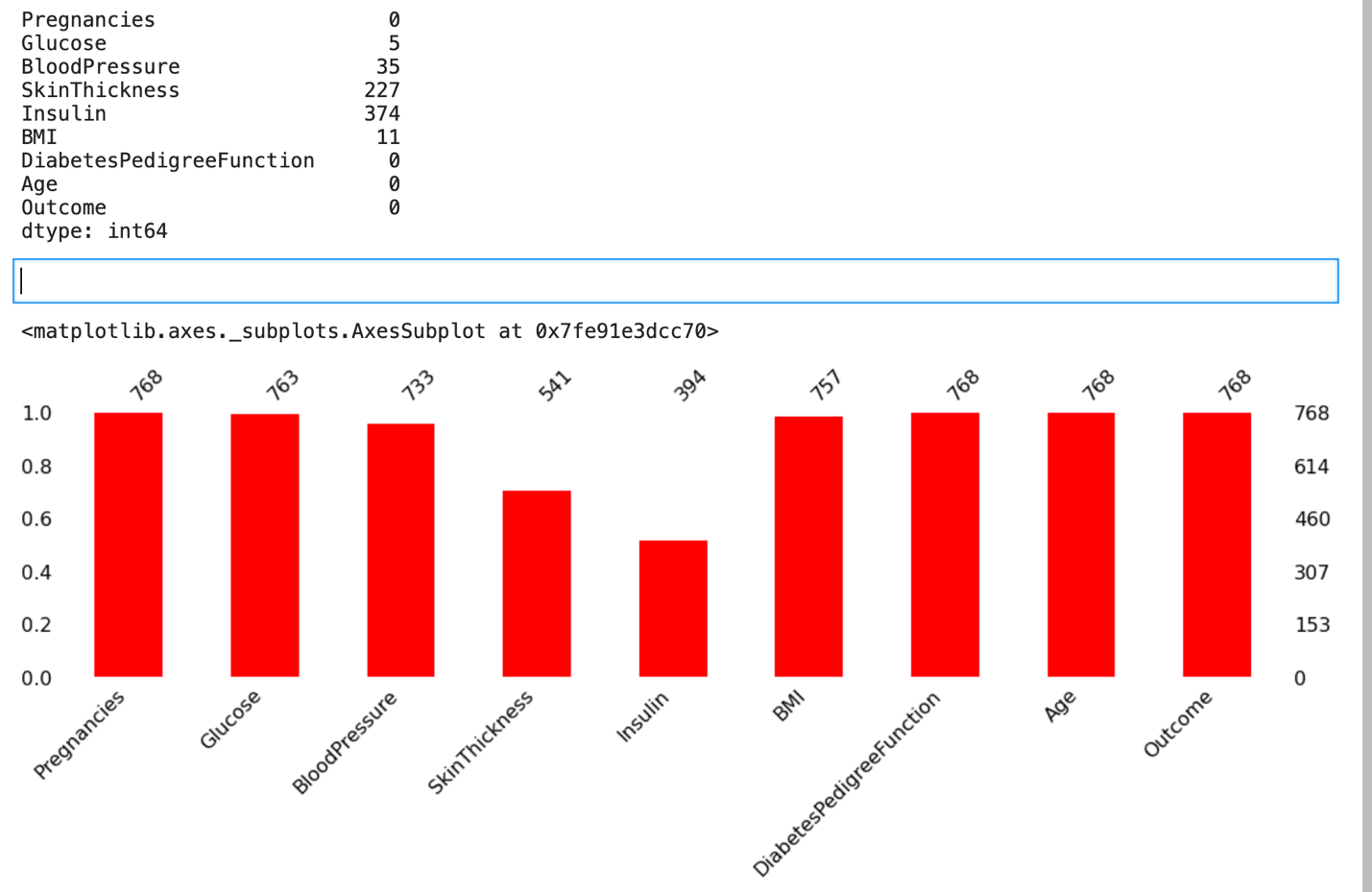
Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is aproven method of resolving such issues. Data preprocessing prepares raw data for further processing.

After reading the data from the csv file initially, we have to change the name of the columns to be used correctly Second, we have to change the values of the category or score columns to (0,1) in order for us to use it

## Data Cleaning

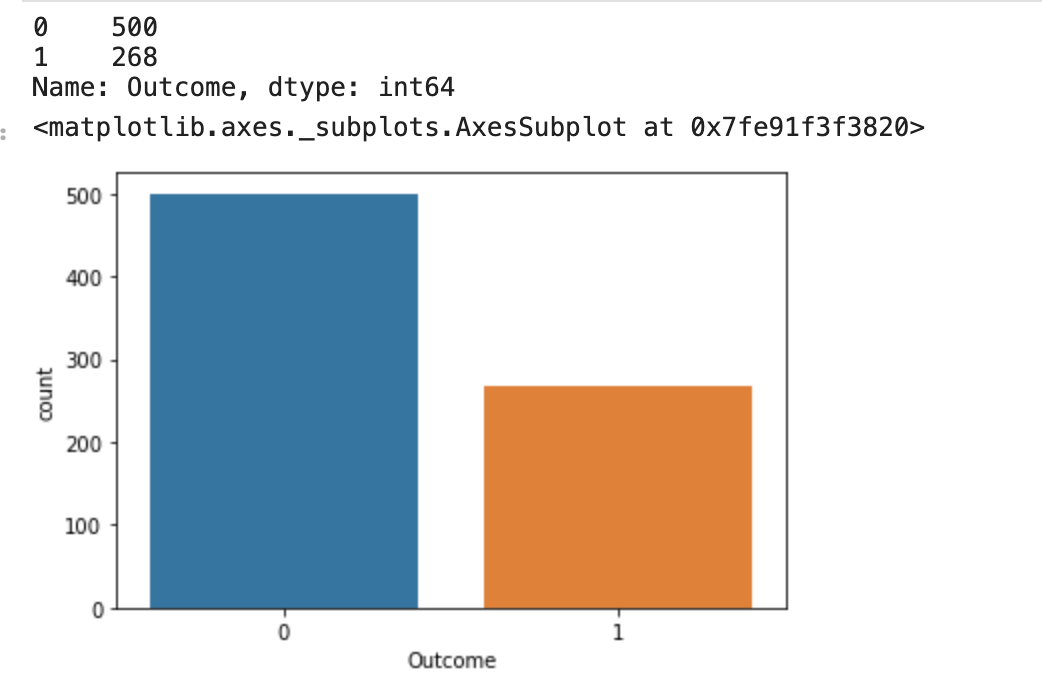
After Change the names of the columns and the Outcome value we have to check if there is any Null value because this could affect the classifier result.

The next figure will show the number of each of columns



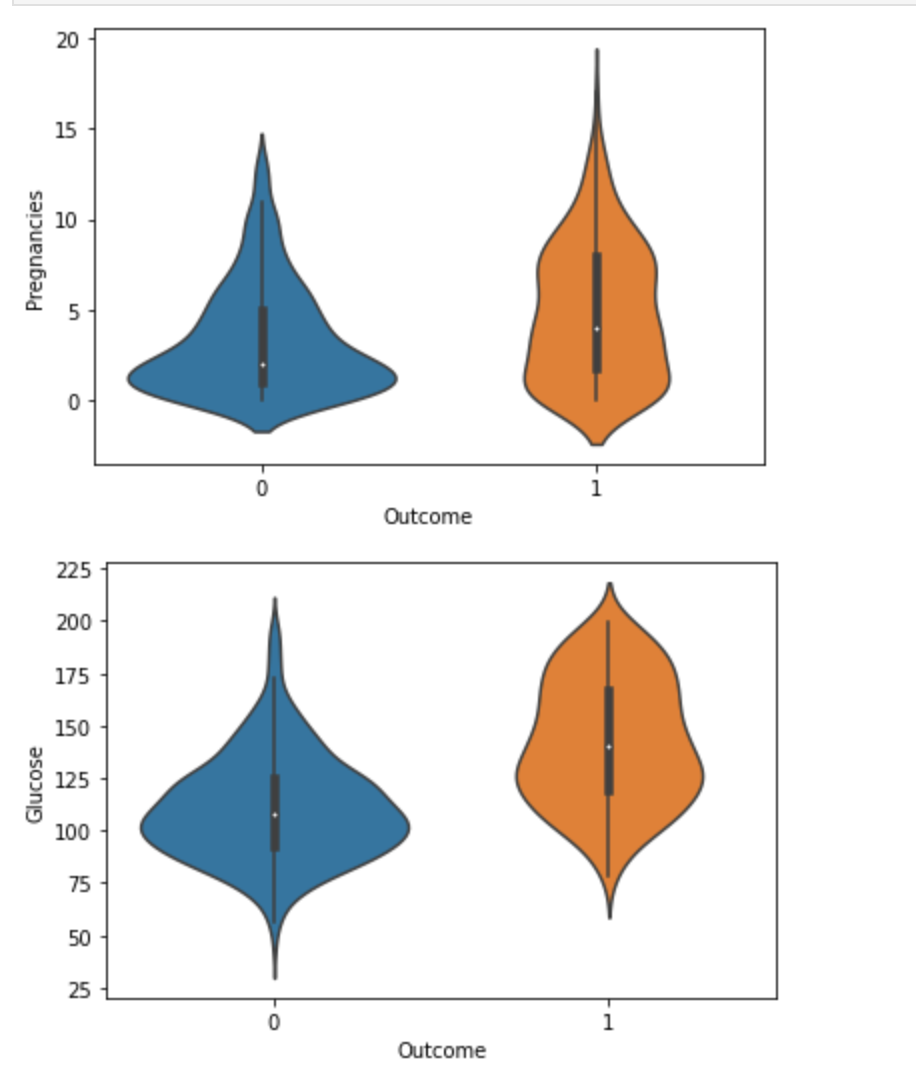
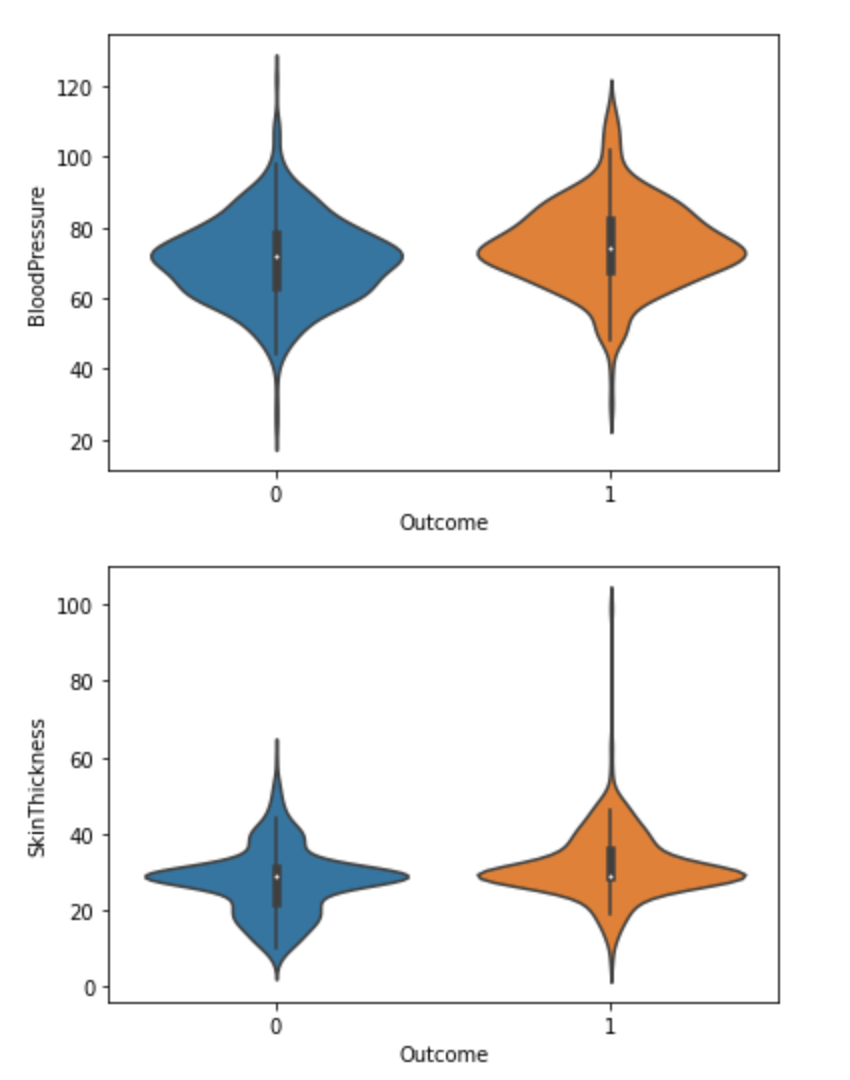
## Data Distribution

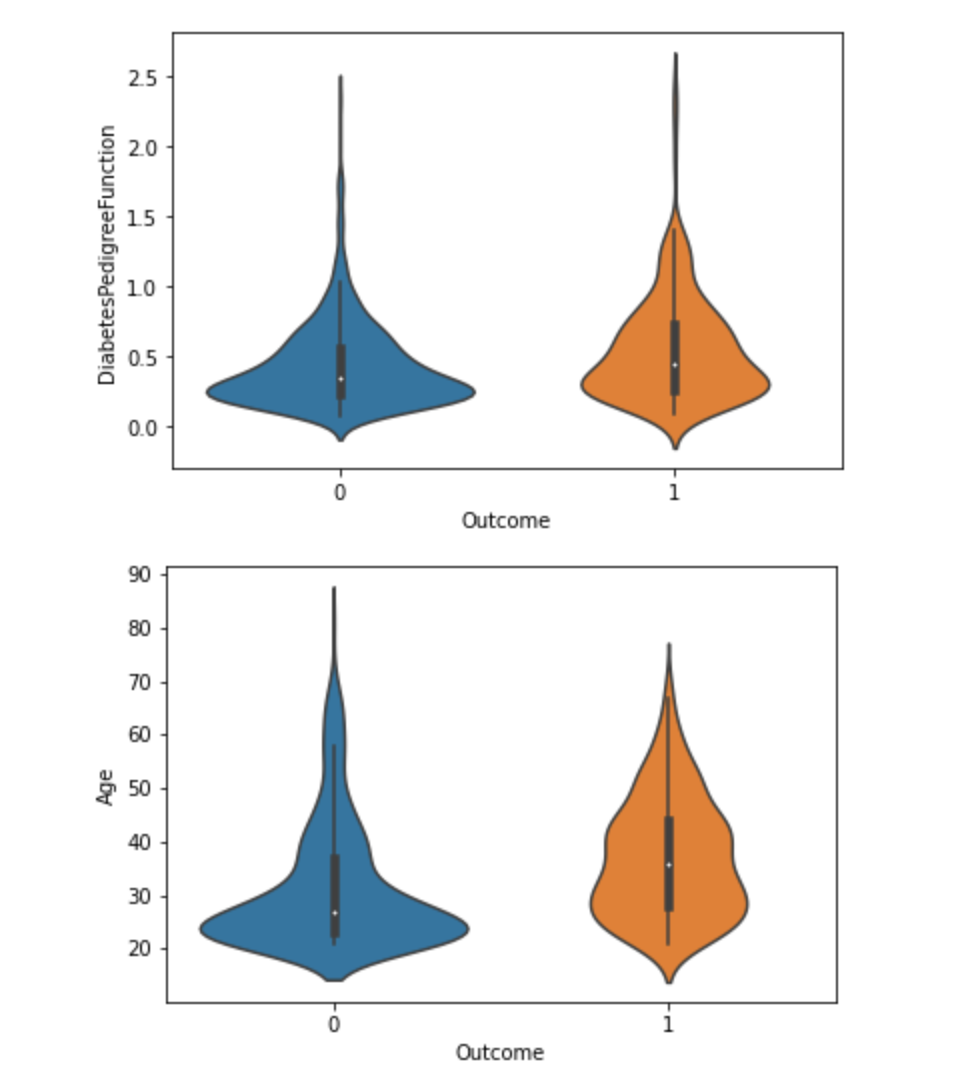
The below figure shows how the target class is distributed

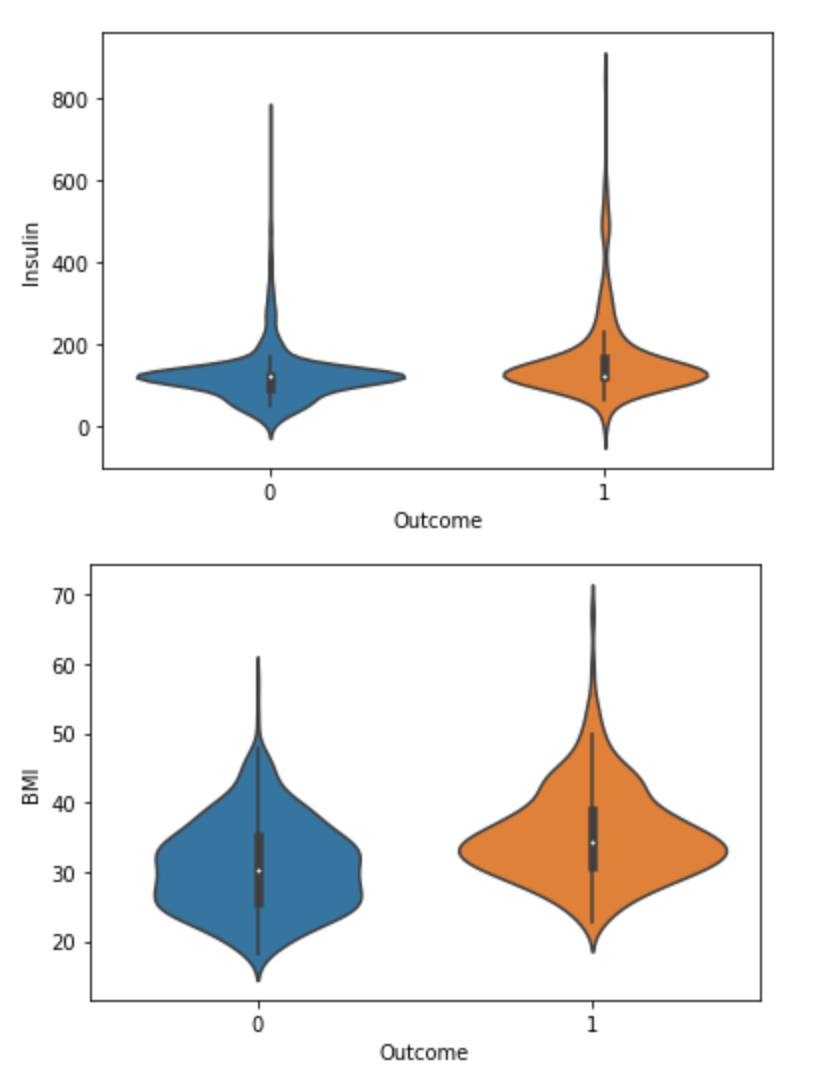


# Additional features

The relationship between the target class and each of the other columns

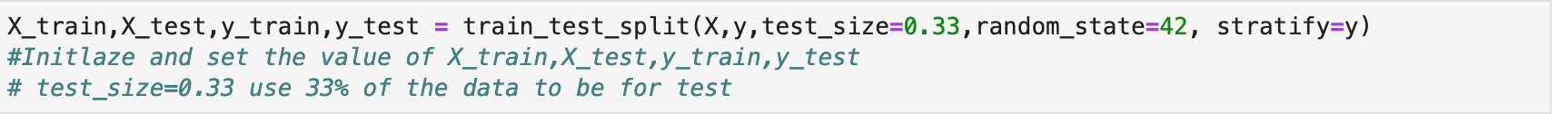


****

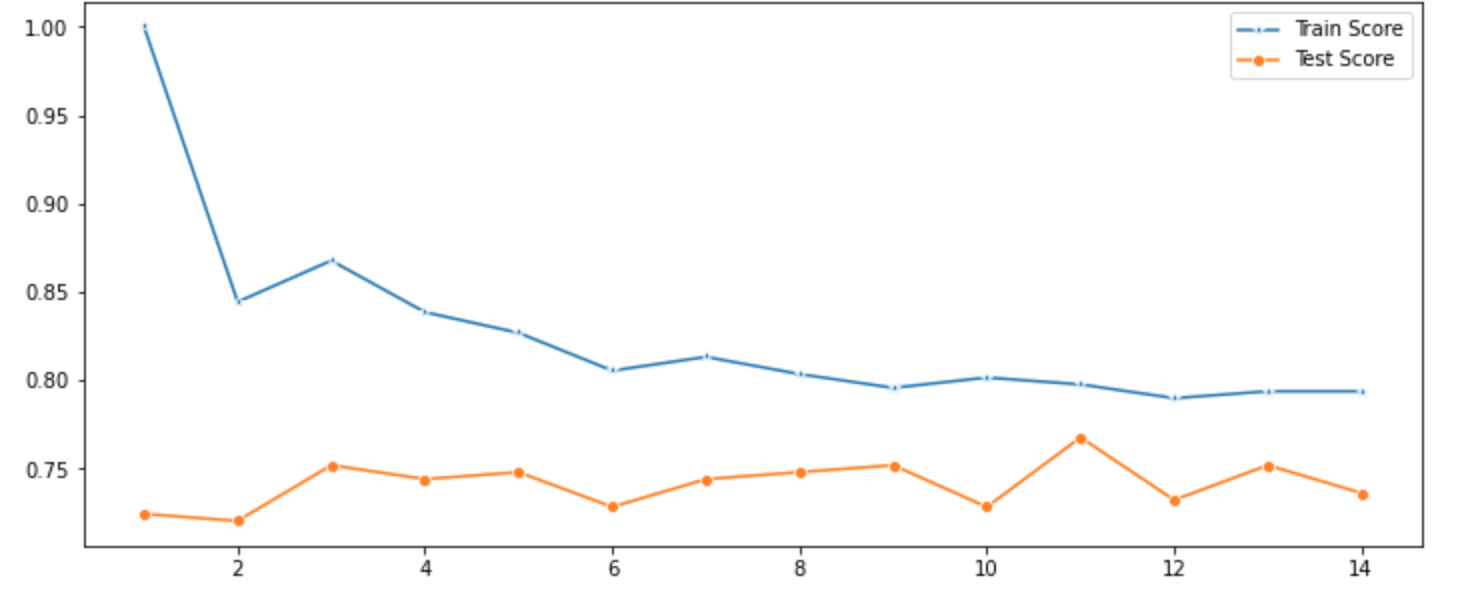


# Results

After scaling the data, set the X, Y we have to split the data in to train data and test data and set 33% of the data to de for testing:

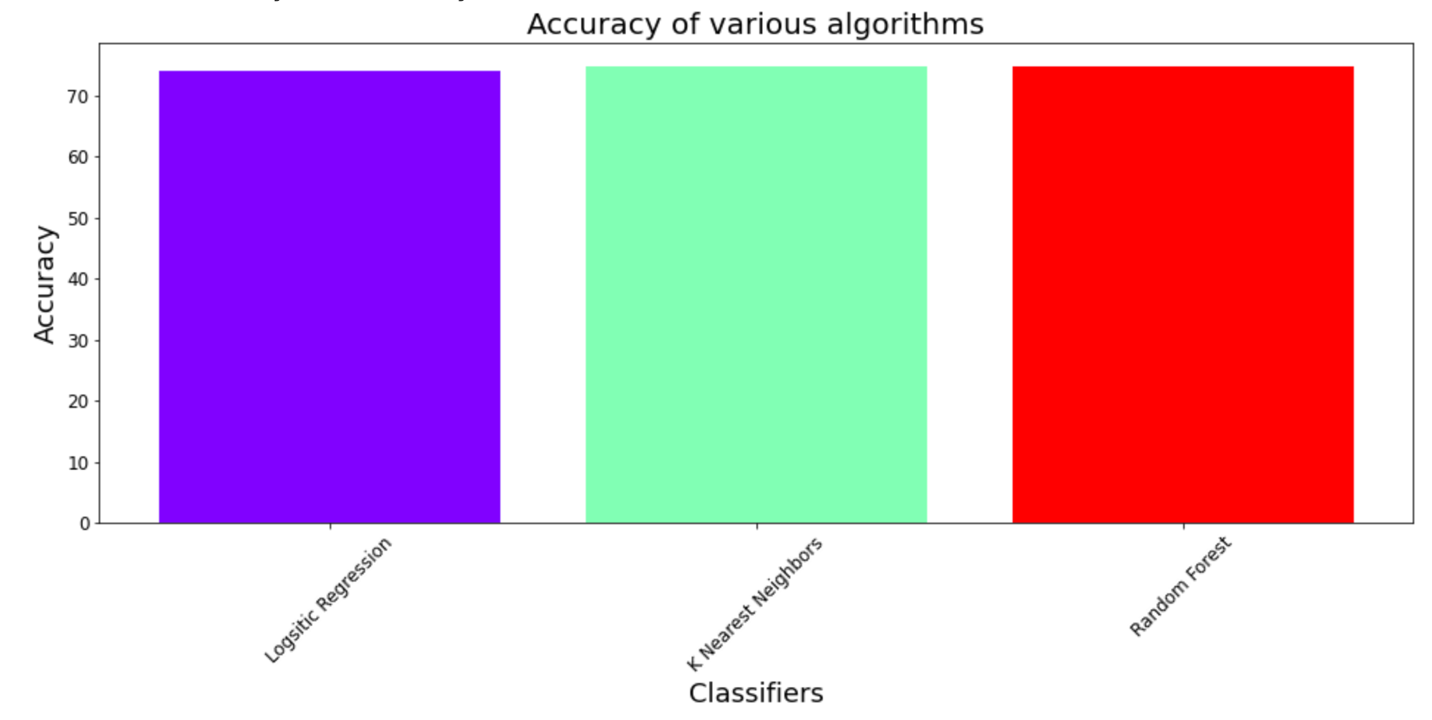


We have also to select the best K for the KNeighbors Classifier to find the Best which is k=11



Finally, and After set the predictions, fitting and build the Logistic Regression classifier, Nearest Neighbors classifier and Random Forest classifier

And do Comparison between the accuracy from all of the three classifiers. I find that Nearest Neighbors classifier is the best with 74.8%



# 

# Conclusion

In the end, I believe that better results can be achieved by first finding more information that can greatly help in improving the workbook, secondly trying to find the right way to fill in and analyzing the missing information in data, and finally choosing the right classifier by trying more than one classifier as well as changing its properties.

To help solve this growing crisis in the world and ease its treatment in time

# References

[1] <https://www.kaggle.com/uciml/pima-indians-diabetes-database>

[2] <https://scikit-learn.org/stable/index.html>

[3] <https://matplotlib.org/>

[4] <https://www.udemy.com/course/python-for-data-science-and-machine-learning-bootcamp/>

# 