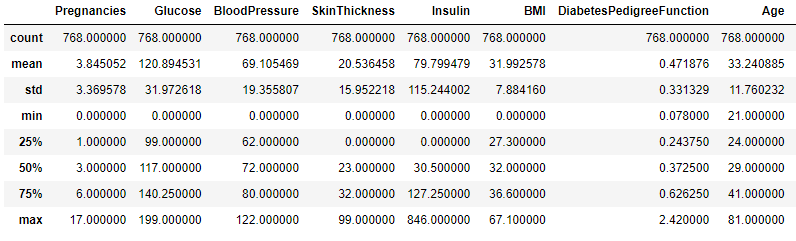
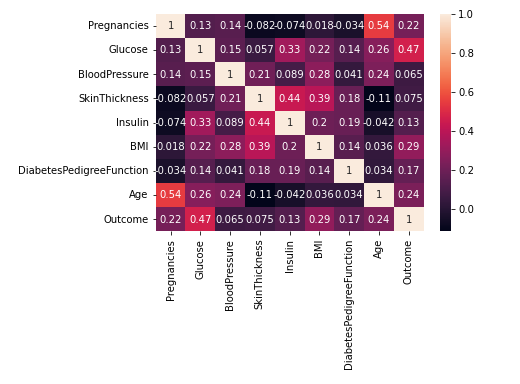
**Binary Classification**

**Part 1: Diabetes\_dataset\_1**

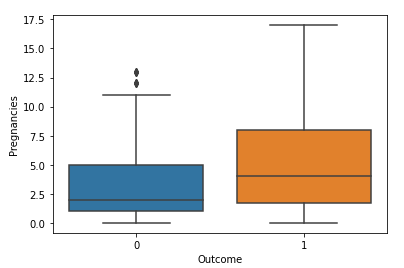
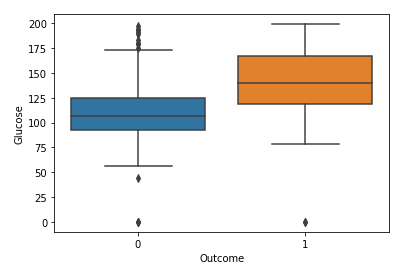
The dataset from Diabetes\_dataset\_1.csv has 768 rows and 9 columns. The column ‘Outcome’ is the label class. There is no missing data in the dataset.

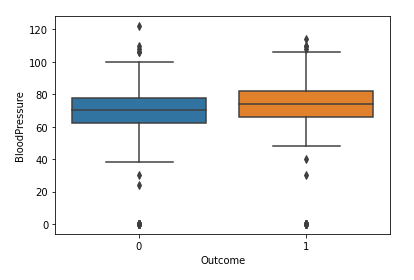
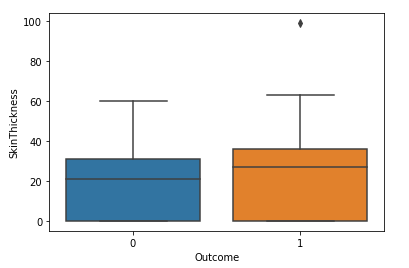


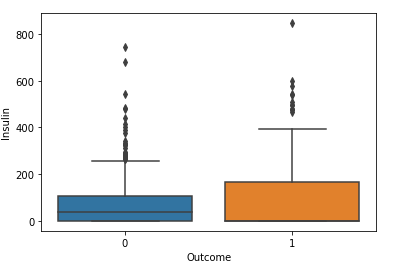
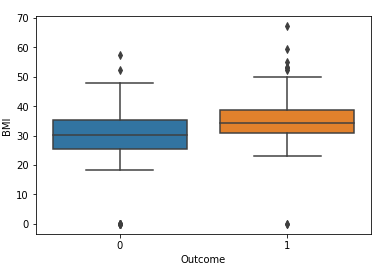
Above I have included data ranges and a few statistical values of the different features and below is a heatmap of correlations of the different columns with each other.

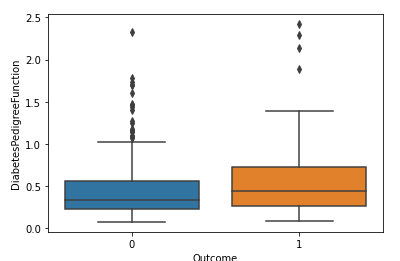
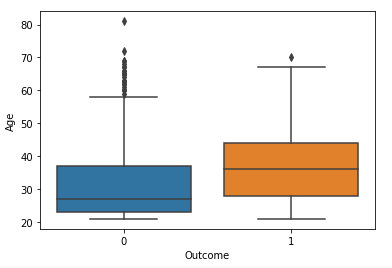


To see how each feature varies with different outcomes, I made box plots of each.

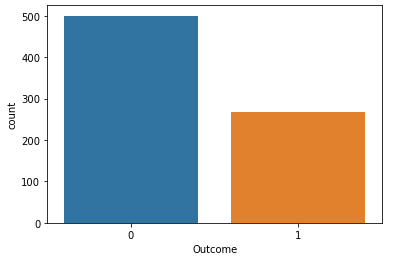
 

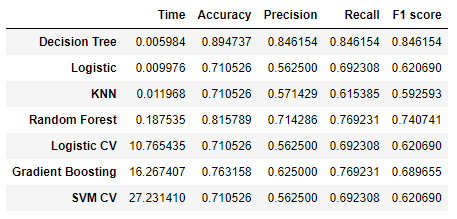
 

From this we can say that Glucose is the feature where we can clearly see that a higher glucose level is more likely to have diabetes. The other features are less distinct to say clearly.

We have 500 examples of outcome 0 and 268 examples of outcome 1. We can say that the classes are balanced.



After training various classification models on the data and comparing them with a validation set, we get.



Here we can see that the decision tree model gives us the most accuracy and takes the least time so we should use this model.

Checking feature importance using the decision tree we get:

Pregnancies: 0.05219329 Glucose: 0.28543296

BloodPressure: 0.10981675 SkinThickness: 0.02578222

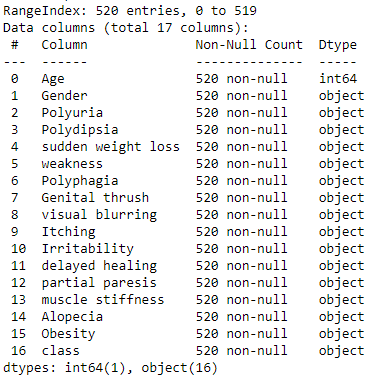
Insulin: 0.05054088 BMI: 0.17901464

DiabetesPedigreeFunction: 0.15148789 Age: 0.14573137

From this we can see that Glucose is the most important feature, followed by BMI, DiabetesPedigreeFunction and Age.

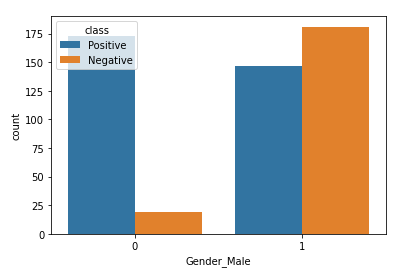
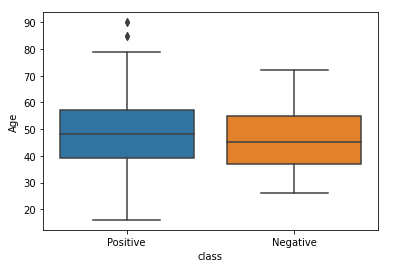
**Part 2: Diabetes\_dataset\_2**

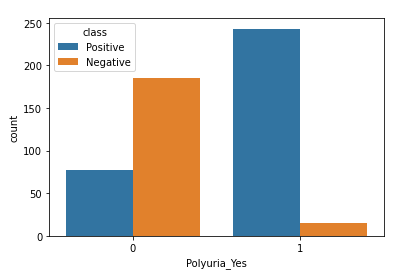
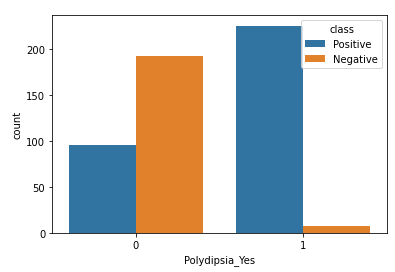
The dataset from Diabetes\_dataset\_2.csv has 520 rows and 17 columns. The column ‘class’ is the label class. There is no missing data in the dataset.

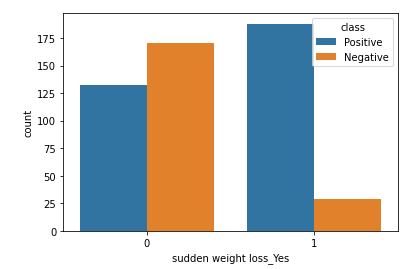
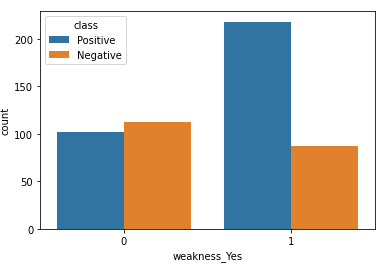


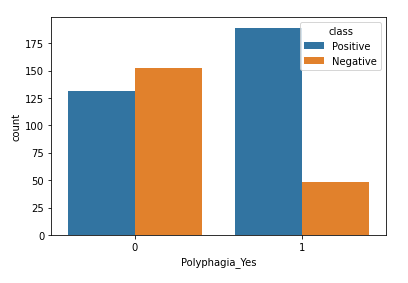
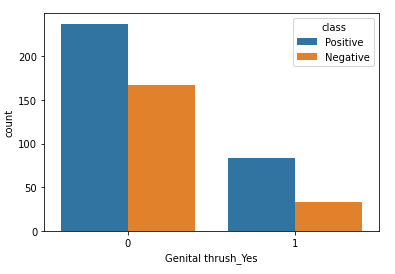
Above I have included information of the dataset. We can see that most features are non-numeric but the machine learning models require numeric data. To fix this I use dummy variables.

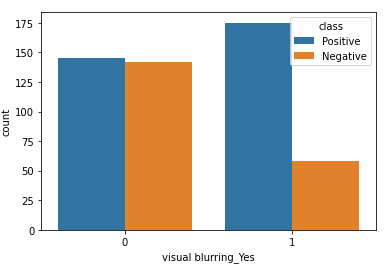
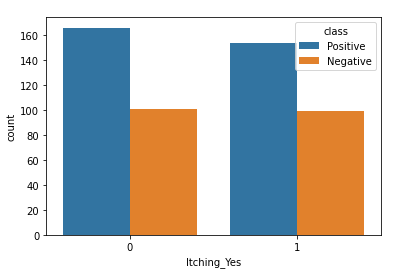
To see how each feature varies with different classes, I made a box plot for age and count plots for the rest

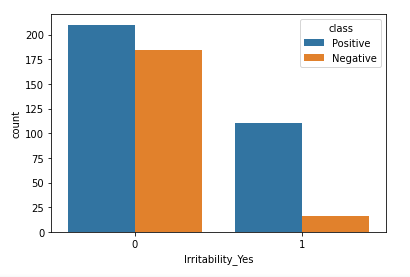
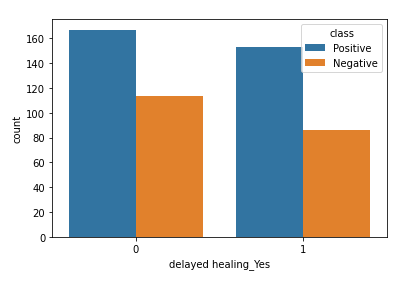


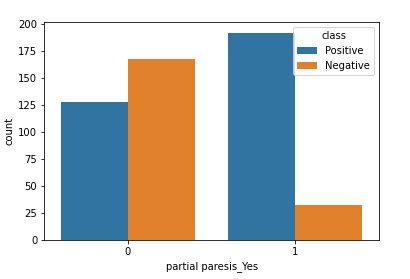
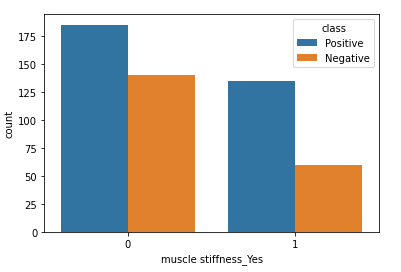
 

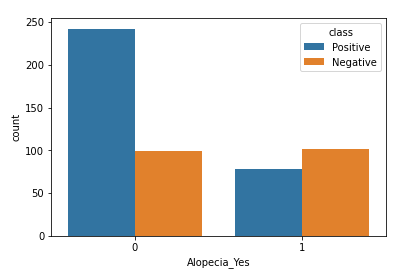
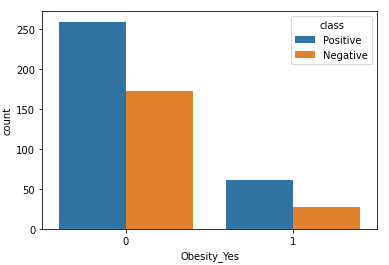
 

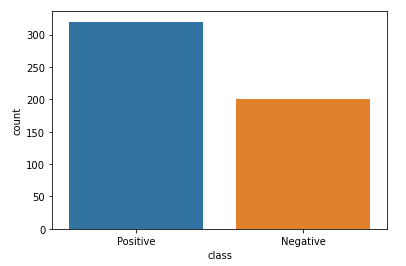
 

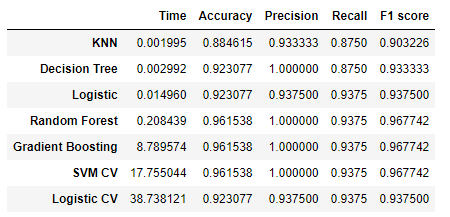
 

From this we can say that Polyuria\_Yes is the feature where we can clearly see that having Polyuria makes one more likely to have diabetes. The other features are less distinct to say clearly.

We have 320 examples of positive class and 200 examples of negative class. We can say that the classes are balanced.



After training various classification models on the data and comparing them with a validation set, we get.



Here we can see that the random forest model gives us the most accuracy and doesn’t take a lot of time so we should use this model.

Checking feature importance using the decision tree we get:

Age: 0.094 Gender\_Male: 0.114

Polyuria\_Yes: 0.457 Polydipsia\_Yes: 0.116

sudden weight loss\_Yes: 0.027 weakness\_Yes: 0.001

Polyphagia\_Yes: 0 Genital thrush\_Yes: 0.022

visual blurring\_Yes: 0.027 Itching\_Yes: 0.002

Irritability\_Yes: 0.020 delayed healing\_Yes: 0.026

partial paresis\_Yes: 0.010 muscle stiffness\_Yes: 0.011

Alopecia\_Yes: 0.046 Obesity\_Yes: 0.027

From this we can see that having Polyuria is the most important feature, followed by having Polydipsia and having male gender.