DIABETES OUTCOME PREDICTION

MACHINE LEARNING ENGINEER CAPSTONE PROJECT

OCTOBER 22, 2020

EMMANUEL KWASI FIAGBEDZI

DEFINITION

OVERVIEW

Diabetes mellitus is a chronic disease where the concentration of glucose in the blood of the patients is at an unusually high level (WHO, 2014). This disease could be caused by the failure of some cells in the body to respond to insulin, or the inability of the pancreas to produce insulin; a protein that stimulates blood glucose reduction (WHO, 2013). In the early stages of the condition, symptoms include frequent urination whilst at the end stage, is characterized by blindness, idiopathic foot ulcers, stroke and eventually death.

Approximately 8% of the world population (adults only) suffer from *diabetes* IDF (2019). According to the CDC (2020), *diabetes* ranks 7th among the diseases that cause the most deaths in the worldwide. Though several efforts have been directed towards the reduction of the incidence of *diabetes*, cases continue to rise (IDF, 2019). The risk of complications from *diabetes* can be prevented when measures aimed at controlling the blood glucose levels are applied at an early stage of the condition thus, making early detection and treatment of the condition very important. Also, *diabetes* tests in regions around the world that do not have ready access to healthcare takes days and sometimes months.

PROBLEM STATEMENT

Therefore, it is very important to have a machine learning model, that can predict the diabetic status faster than the traditional methods. This would ensure that interventions to reduce the risk of complications with regards to diabetes, are reduced. This problem was a binary classification task and characteristics of patients such as blood glucose levels, skin thickness and age were considered depending on the strength of correlation with diabetes outcome and would be used to predict whether a patient is diabetic or not.

To solve the problem, a support vector classifier model was trained to predict an individual's diabetes status based on the). The dataset used in this project is a Pima Indian Diabetes dataset

downloaded from Kaggle. Each row represents a patient and each column contains certain measured characteristic of the patient; Blood glucose Concentration, BMI, Skin Thickness, Insulin, Diabetes Pedigree Function and Outcome, which indicates whether the patient is diabetic or not. Portion of this dataset will be used to train a machine learning algorithm and the remaining portion of that dataset will be used for checking the out-sample performance of the algorithm. It was expected that this model will have better performance than a dummy classifier model used in the project proposal.

METRICS

The performance of each model will be checked by using precision and recall score. Precision tells how often the model is correct when it predicts a positive value whilst recall shows how sensitive the classifier is in detecting positive classes. Precision and Recall are used because they tell the performance of a model better when there is class imbalance than accuracy and roc_auc_score.

ANALYSIS

DATA EXPLORATION

The data was downloaded and loaded into a jupyter notebook. The figure below shows how the data looks like

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

There were 768 patients and nine measured characteristics. There were no missing values and duplicated entries in the data. Some summary statistics were also derived from the data

as shown below. It could be seen that some of the features such as Insulin, had a very wide range.

Out[26]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000

We also performed a Schapiro test on all the features to determine if they are normally distributed so that we can use the appropriate test. It was found out that all the features were normally distributed hence an independent t-test can be used for the statistical analysis.

:

	shapiro_test_statistic	p_value
Pregnancies	0.904278	1.608089e-21
Glucose	0.970104	1.986761e-11
BloodPressure	0.818921	1.584007e-28
SkinThickness	0.904627	1.751576e-21
Insulin	0.722021	7.915248e-34
BMI	0.949989	1.840562e-15
DiabetesPedigreeFunction	0.836519	2.477697e-27
Age	0.874766	2.401947e-24

EXPLORATORY VISUALIZATION

Exploratory visualizations and statistical analysis were done to see if there was correlation between the various features and the target variable ‘Outcome’. The independent t-test was

used as the statistical test with an alpha level of 0.05. From the figures below, it could be seen that all the comparisons showed statistical significance expect the Blood Pressure feature.

HO: There is no relationship between number of pregnancies and diabetes status.

H1: There is a relationship between number of pregnancies and diabetes status.

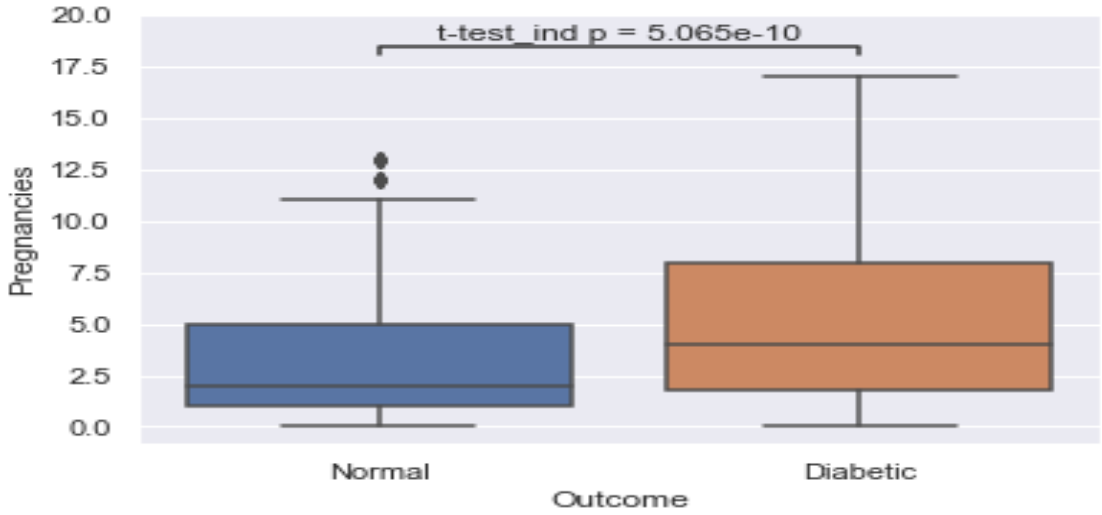

Fig 1: Relationship between **Number of Pregnancies** and **Diabetes**

As can be seen from Fig 1, the median number of pregnancies for the diabetes positive females were slightly higher than the negatives. Since the **p value <<< 0.5**, we can reject the null hypothesis and accept the alternative hypothesis that **there is an association between number of pregnancies and diabetes**.

HO: There is no relationship between age and diabetes status.

H1: There is a relationship between age and diabetes status.

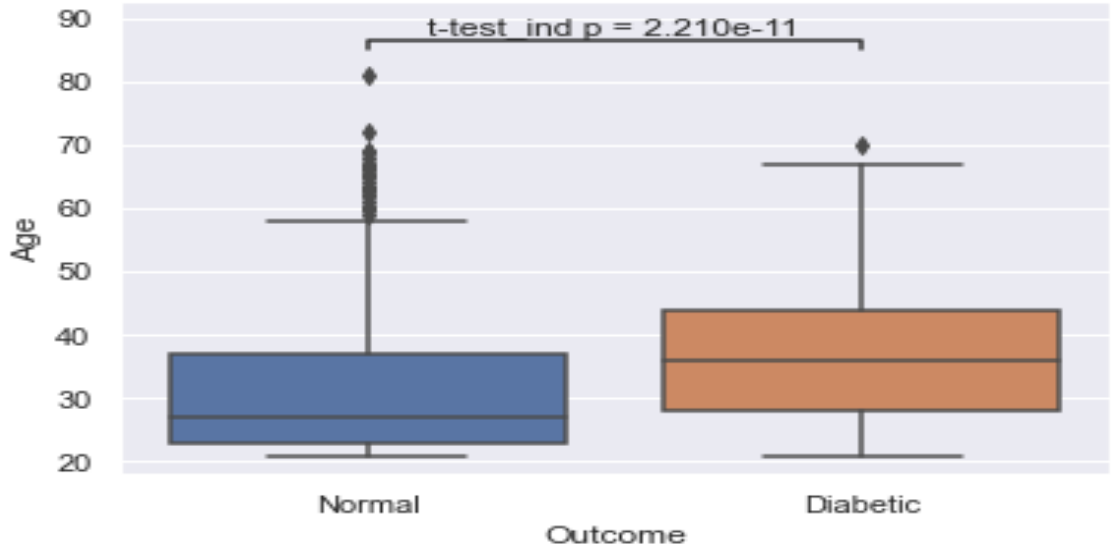

Fig 2: Relationship between **Age** and **Diabetes**

As can be seen from Fig 2, the median age for the diabetic females were slightly higher than the negatives. Since the **p value <<< 0.5**, we can reject the null hypothesis and accept the alternative hypothesis that **there is an association age and diabetes**.

HO: There is no relationship between Glucose and diabetes status.

H1: There is a relationship between Glucose and diabetes status.

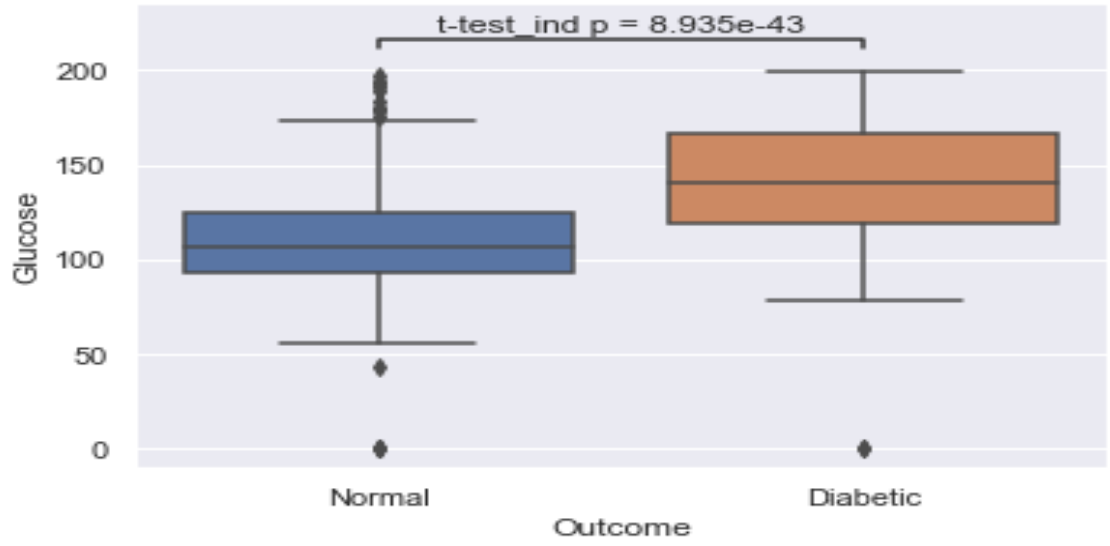

Fig 3: Relationship between **Glucose** and **Diabetes**

As can be seen from Fig 3, the median glucose levels for the diabetic females were slightly higher than the normal. Since the **p value < 0.5**, we can reject the null hypothesis and accept the alternative hypothesis that **there is an association glucose and diabetes**.

HO: There is no relationship between blood pressure and diabetes status.

H1: There is a relationship between blood pressure and diabetes status.

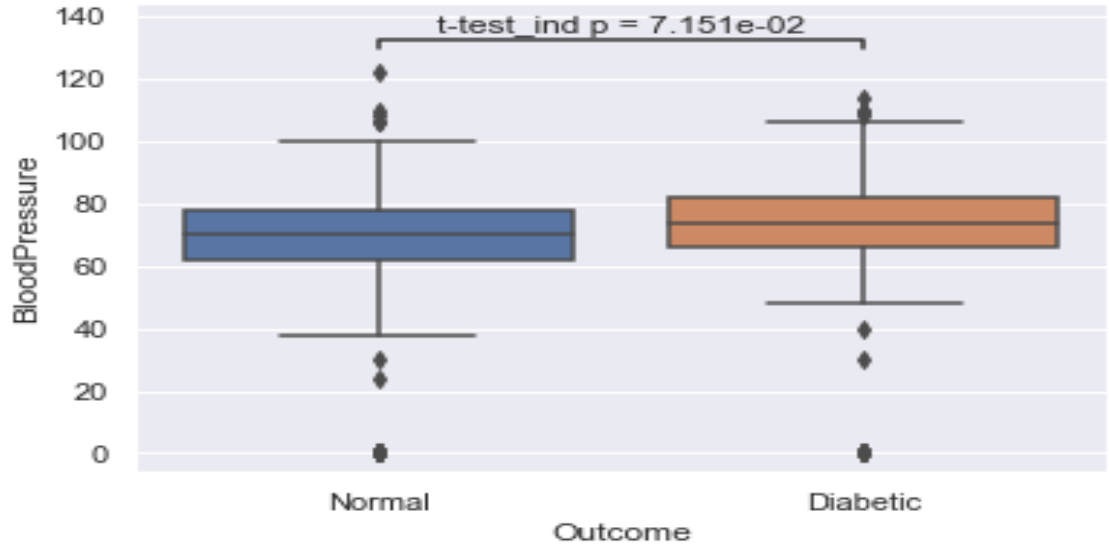

Fig 4: Relationship between **Blood pressure** and **Diabetes**

As can be seen from Fig 4, the median blood pressure levels for the diabetic females were slightly higher than the normal. Since the **p value > 0.5**, we cannot reject the null hypothesis and accept the alternative hypothesis that **there is an association Blood Pressure and diabetes**.

HO: There is no relationship between skin thickness and diabetes status.

H1: There is a relationship between skin thickness and diabetes status.

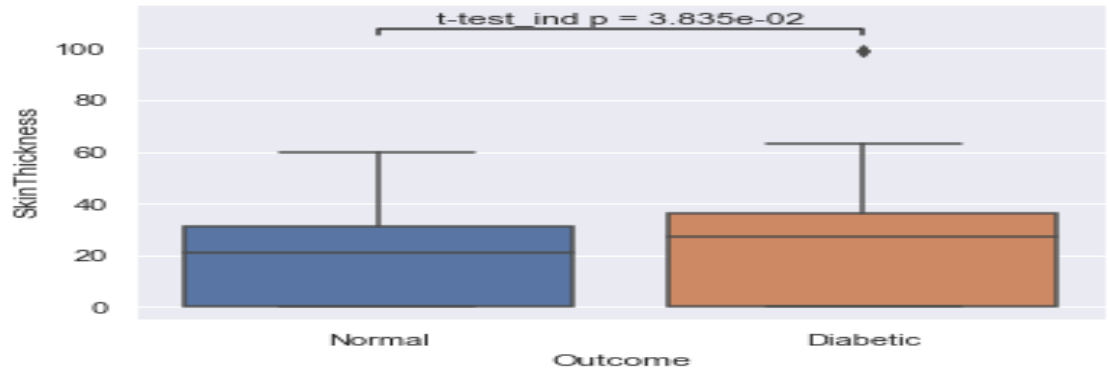

Fig 5: Relationship between **Skin thickness** and **Diabetes**

As can be seen from the Fig 5, the median skin thickness for the diabetic females were slightly higher than the normal. Since the **p value < 0.5**, we can reject the null hypothesis and accept the alternative hypothesis that **there is an association Skin thickness and diabetes**.

HO: There is no relationship between Insulin and diabetes status.

H1: There is a relationship between Insulin and diabetes status.

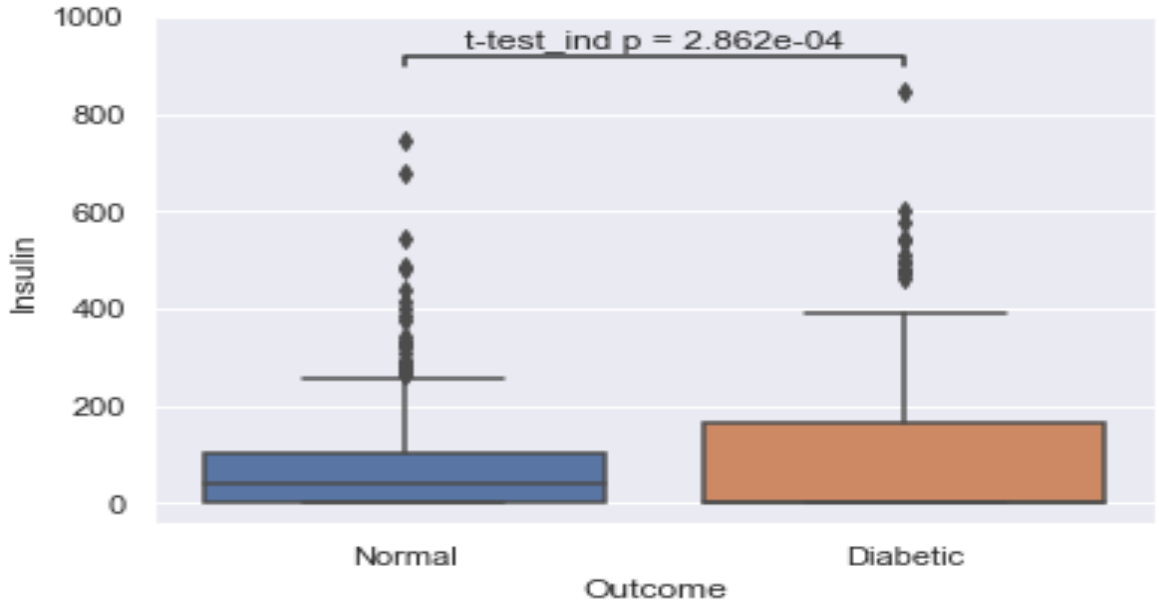

Fig 6: Relationship between **Insulin** and **Diabetes**

As can be seen Fig 6, the median Insulin for the diabetic females were slightly lower than the normal. Since the **p value < 0.5**, we can accept the alternative hypothesis and reject the null hypothesis that **there is no association between Insulin and diabetes**.

HO: There is no relationship between BMI and diabetes status.

H1: There is a relationship between BMI and diabetes status.

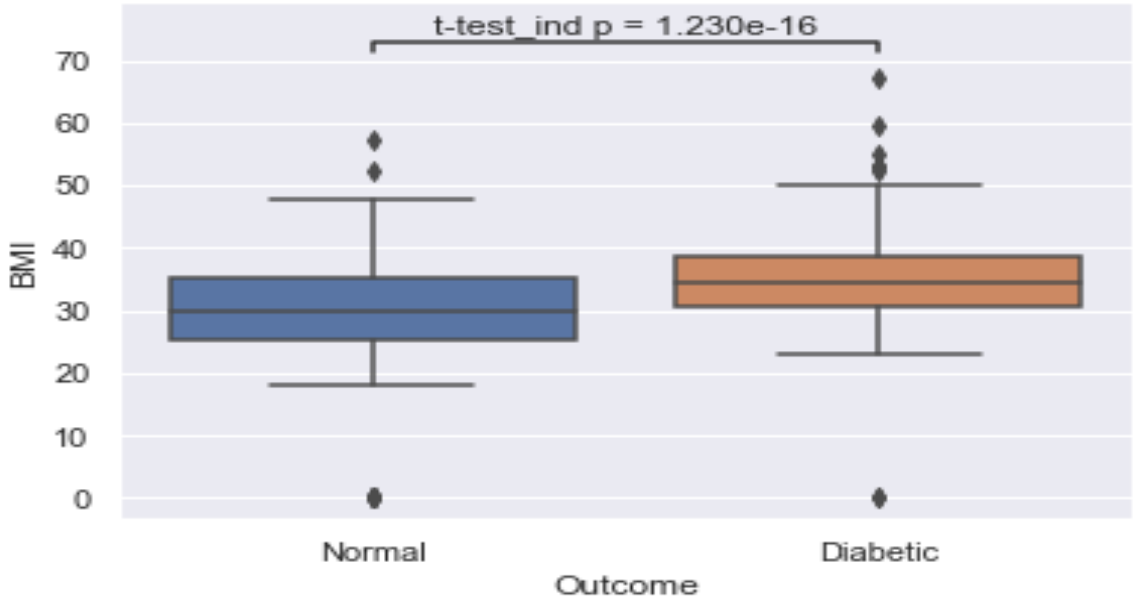

Fig 7: Relationship between **BMI** and **Diabetes**

As can be seen from Fig 7, the median BMI for the diabetic females were slightly higher than the normal. Since the **p value < 0.5**, we can reject the null hypothesis and accept the alternative hypothesis that **there is an association BMI and diabetes**.

HO: There is no relationship between Diabetes Pedigree Function and diabetes status.

H1: There is a relationship between Diabetes Pedigree Function and diabetes status.

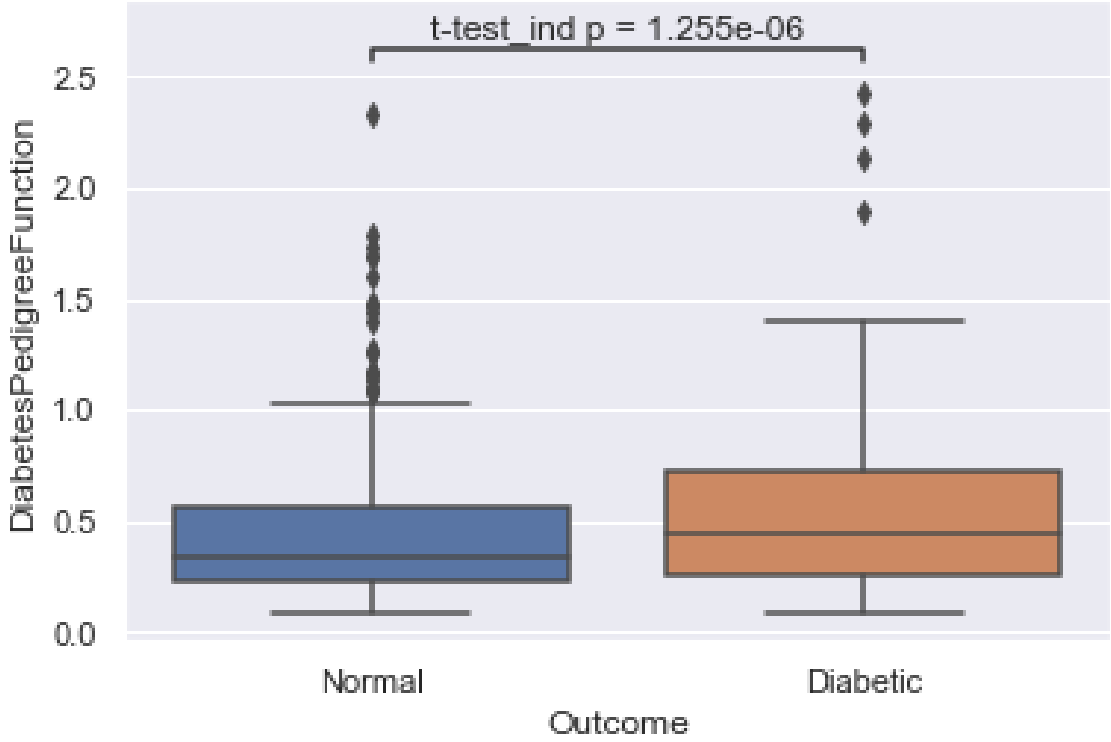

Fig 8: Relationship between **Diabetes Pedigree Function** and **Diabetes**

As can be seen from Fig 8, the median Diabetes Pedigree Function for the diabetic females were slightly higher than the normal. Since the **p value < 0.5**, we can reject the null hypothesis and accept the alternative hypothesis that **there is an association Diabetes Pedigree Function and diabetes**.

ALGORITHMS AND TECHNIQUES

Since this was an imbalanced classification problem, an algorithm which works well with imbalanced classes was used. This was support vector classifier (SVC). It was used alongside a robust scaler which would normalize the data to reduce the effect of outliers and a principal component analysis (pca) transformer which selected the top 5 principal components from the features to be fed into the SVC.

A voting classifier consisting of pipelines just like the svc pipeline described above but with the classifiers logistic regression and Bernoulli naïve bayes was used.

BENCHMARK

A dummy classifier was used as the baseline model. It had a recall score of 49 % and a precision score of 32 %. Any model that gets a higher score than this would be considered a successful model. Fig 9 shows the confusion matrix and precision recall curve of this baseline model.

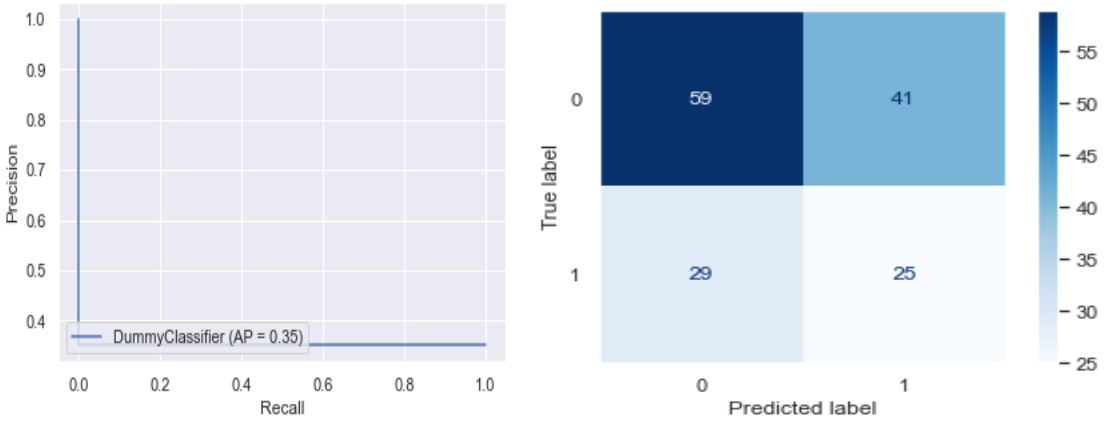

Fig 9: A precision recall curve (left) and a confusion matrix (right) for the benchmark model

As shown in Fig 9, the precision recall curve hovers around the bottom left corner which is not good. Also, there were 29 false positives which doesn't meet our objective. Furthermore, the Fischer's exact test statistic was 1.2 with a p value of 0.6 which is far greater than the alpha value of 0.05. Any model that obtains a significant p value would be a better model.

METHODOLOGY

PREPROCESSING

The feature matrix X contained all the columns in the data except the 'Outcome' column. Outliers were detected using the Tukey method. However, since there were no rows with more than two features having outliers, none were dropped.

They were then passed a robust scaler that normalizes all the features in the feature matrix with the parameters all left at their default value. A PCA transformer was then used to extract the first five principal components from all the 8 features. The response vector 'Outcome'

was converted into an integer data type using label binarizer in some instances, as well as the pandas series method, astype().

IMPLEMENTATION

The robust scaler, pca as well as a number of classifiers such as (LogisiticRegression, SVC, BernoulliNB, KNeighborsClassifier, RandomForestClassifier and GaussianNB) were all combined into pipelines. These pipelines were trained and evaluated using ten-fold cross-validation with 'recall' and 'precision' as the scoring metric. The top three pipelines i.e. those with (LogisiticRegression, SVC, and BernoulliNB) in terms of highest cross validated precision and recall were passed to the next step.

REFINEMENT

The top three pipelines were taken through rigorous five-fold grid search cross validation with precision and recall as the metric. The best estimators from these three models were used to build a voting classifier. A 10-fold cross validated recall score was performed for the voting classifier as well as for the top pipeline of the top three best estimators (SVC pipeline). Confusion matrices as well as precision recall curves were also plotted for these models. The final model, which happened to be the SVC pipeline was taken through a Fischer exact test and a p value significance number was calculated at a significance level of 0.05

RESULTS

MODEL EVALUATION AND VALIDATION

A grid search cross validation was performed on the svc model and the best parameters selected. This model had a best recall score of 80 % and an average precision of 63 %. These values were far higher than that of the dummy classifier. The confusion matrix below also showed that the model had only 6 % false positives out of 154 test cases which fits our objective of a model that can easily detect positive cases of diabetes.

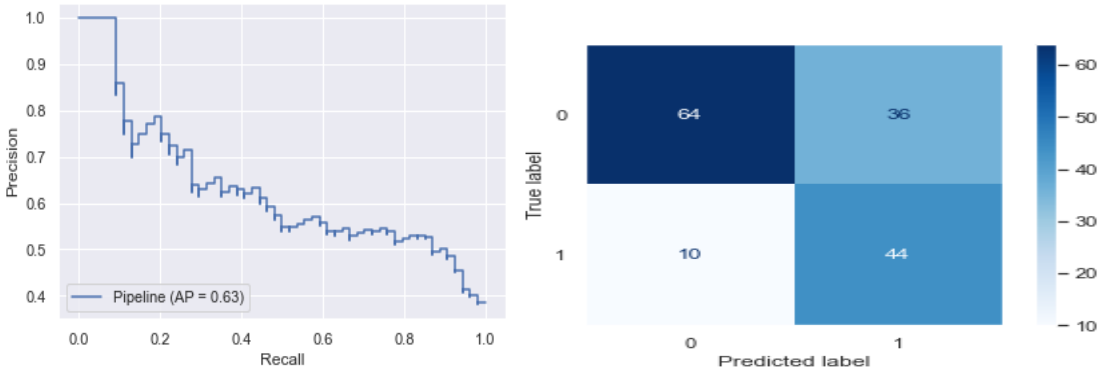

Fig 10: A precision recall curve (left) and a confusion matrix (right) for the final model

As shown in fig 10, the precision recall curve has pushed towards the right-hand corner compared to that of the dummy classifier which shows an improvement. Also, there were far less false positives (10) than that of the dummy classifier (29).

JUSTIFICATION

A Fischer's exact test was used to test whether the values in the confusion matrix were obtained by chance. The test showed significance with a fischers_test_value of 7.82 and p_value of 4.6 X 10-8 which is lower than the alpha level of 0.05 and thus significant. Since the benchmark did not show any significance, we can conclude, that the final model is indeed a better classifier than the benchmark model