SUPERVISED LEARNING PROJECT

BRETT FORSEN

EXPLORATORY DATA ANALYSIS

- No null values were found in the data, however values of 0 existed in places that they should not. For example, Skin Thickness.

- From a glance at the visualizations, it seems as if the number of pregnancies has a fair impact on the likelihood of having diabetes. Glucose levels are probably the most significant contributor, however.

- There existed some extreme values which would need to be taken care of in order to perform an accurate analysis. For example, a BMI value of 67.1, or Skin Thickness greater than 75.

CLEANING & FEATURE ENGINEERING

- My first step was to eliminate values that I considered to be far too extreme and likely due to error. This included the BMI, Blood Pressure, and Glucose features.

- Skin Thickness values of less than 1 or greater than 75 were converted to the mean of the column.

- The values were scaled using SkLearn's StandardScaler. This was especially important for this dataset. For instance, number of Pregnancies were typically in the low single digits, whereas Glucose levels were often over 100.

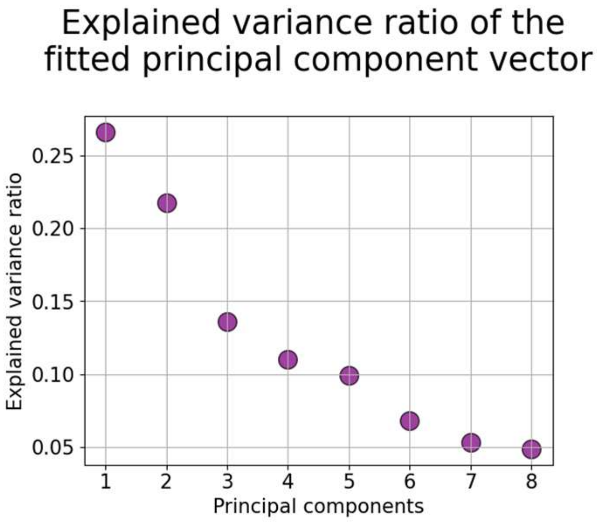
MODEL SELECTION – COMPARING FEATURE SELECTION METHODS (1/2)

- For this dataset I chose to compare the results of two separate feature selection models, PCA and Random Forest.

- The PCA results can be seen on the right graph.

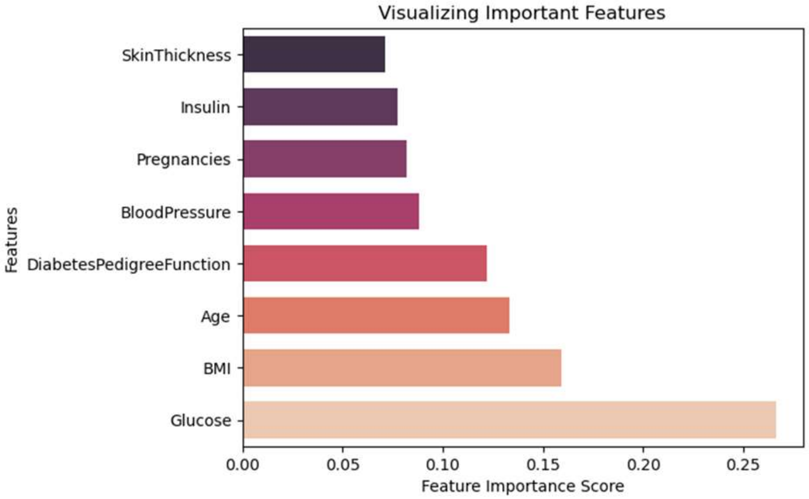
- The order of the principal components mentioned is as follows:

Pregnancies, Glucose Levels, Blood Pressure, Skin Thickness, Insulin, BMI, Genetics (DPF), and Age.



MODEL SELECTION – COMPARING FEATURE SELECTION METHODS (2/2)

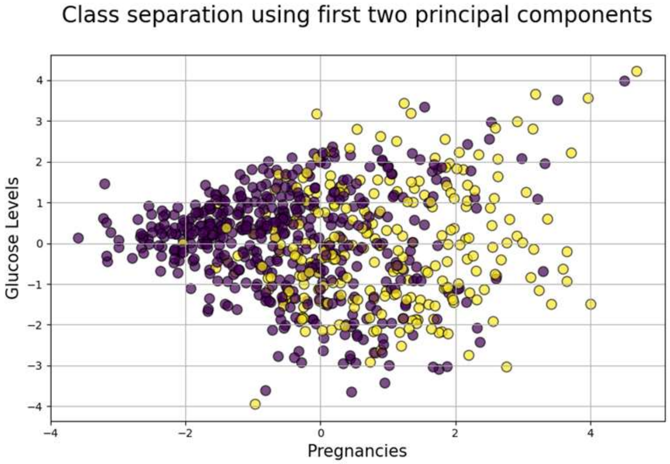
- The Random Forest was run twice, leaving out the least important features from the first iteration.
- The first Random Forest results can be seen on the right:
- The second iteration of the Random Forest appeared to be less accurate, implying that all features were at least of moderate importance.



MODEL COMPARISON

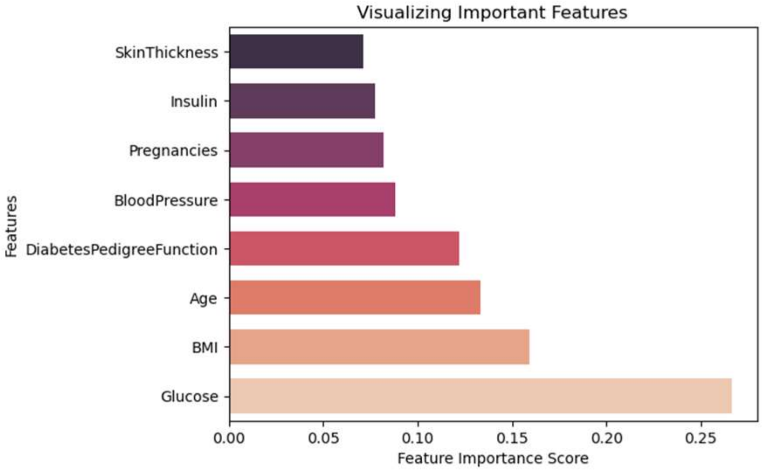
PCA

Number of pregnancies was considered the most important feature by a fair margin. Glucose levels were not far behind.



Random Forest

With Random Forest, pregnancies were considered to be dramatically less significant. BMI and Age were noticeably more so than in the PCA.



CONCLUSION:

- Glucose levels undoubtedly have a strong positive correlation with the appearance of diabetes. This makes sense considering how diabetes is caused, and how it affects the body.

- PCA rated certain components, Age, Genetics (DPF), and BMI as having only a slight indication as to whether or not an individual has diabetes.

- Obviously there is some fundamental difference in the way in which Random Forest and PCA operate that causes these dramatic differences. Further testing would be needed, although this could be partially explained by using both Supervised (Random Forest) and Unsupervised (PCA) machine learning methods.