**Assignment 4: Decision Trees**

Tina Jin and Virginia Weston

**Exploratory Data Analysis**

For our Exploratory Data Analysis, we initially ran a scatter plot matrix to explore relationships from all of the variables excluding the outcome. All of the variables were quantitative, making it easier to decipher relationships between them. Moreover, by using a heatmap, we were able to conceptualize any linear relationships between the data. The strongest linear relationship among the data was between age and pregnancy, with a correlation coefficient of 0.54. Although not considered a strong linear relationship, the two variables can be considered correlated to one another.

**Diagram

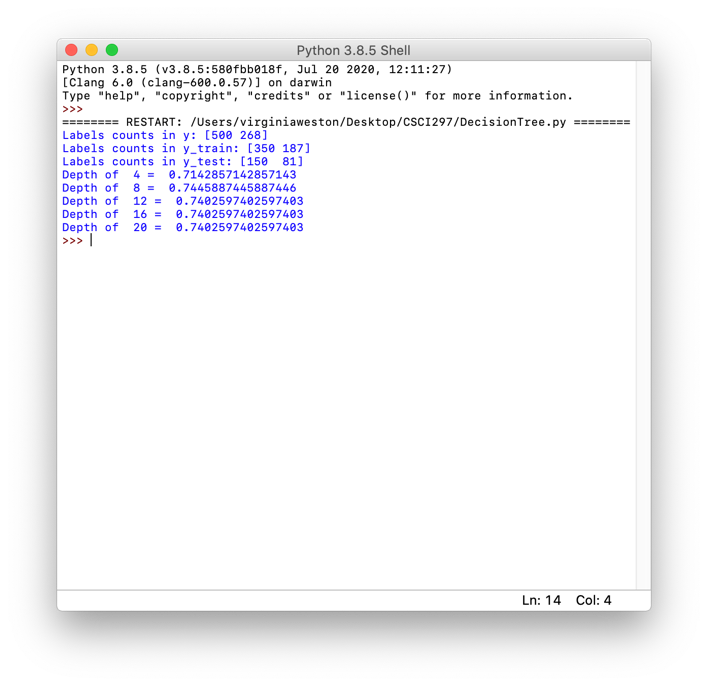
Description automatically generatedGraphical user interface, application

Description automatically generated**

**Decision Tree**

The most important part in the design aspect of our Decision Tree model was adjusting hyperparameters. We chose to adjust how we standardized the data, depth, as well as criterion. We chose to eventually exclude standardizing the data from our model because it ultimately made no difference in the accuracy. The model originally produced the accuracy for depths of 4, 8, 12, 16, and 20 in order for us to compare whether standardizing the data or switching the criterion would make a difference in accuracy. As a result of this testing of depths and other hyperparameters, we concluded that a depth of eight results in the best accuracy for the Decision Tree model.

**Left is using Gini as a criterion check, right uses entropy**



The highest accuracy rate produced from this model was found by using a depth of eight and using the Gini impurity. We also calculated the precision of our decision tree model by called the skit classification report. Below are the results that the model prints. The accuracy is 0.74, and the precision is 0.79.

A picture containing text

Description automatically generated

**Random Forest**

A picture containing text

Description automatically generatedA picture containing text

Description automatically generatedImplementing the random forest algorithm greatly improved our model because it was able to average our original decision tree with a depth of eight. The accuracy increased from 0.74 to 0.76 by implementing a Random Forest algorithm on our decision tree model. Using the RandomForestClassifier API call, we altered our hyperparameters to maximize accuracy. This consisted of setting the number of estimators to 200 and changing the boostrap to False in order to build each tree. Essentially, this model greatly improved upon our previous Decision Tree model because it used a “deep” Decision Tree model to create a more accurate, shallow tree in order to classify the data.