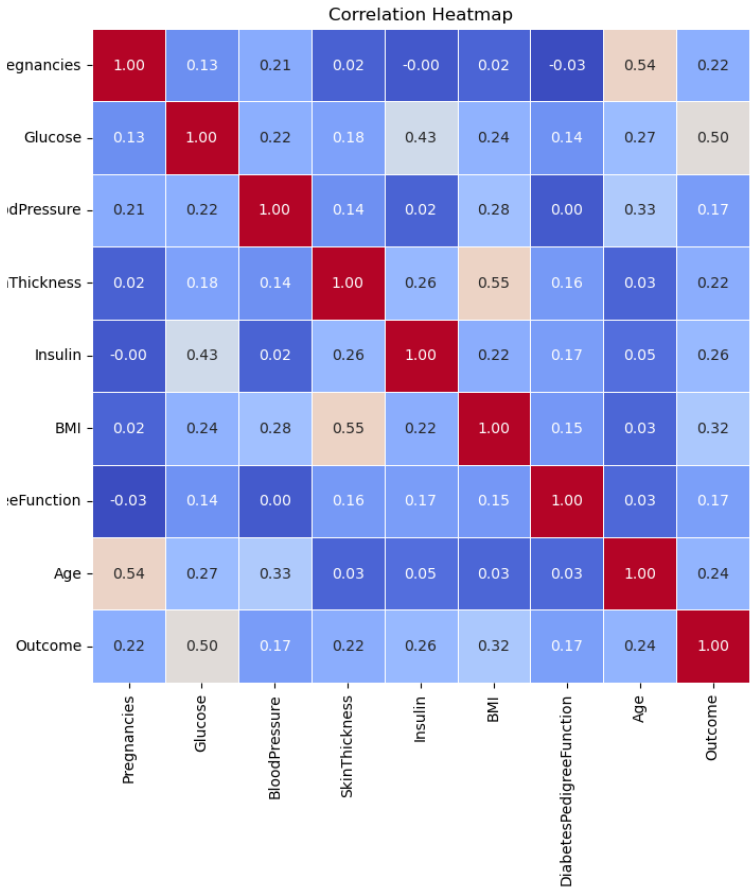
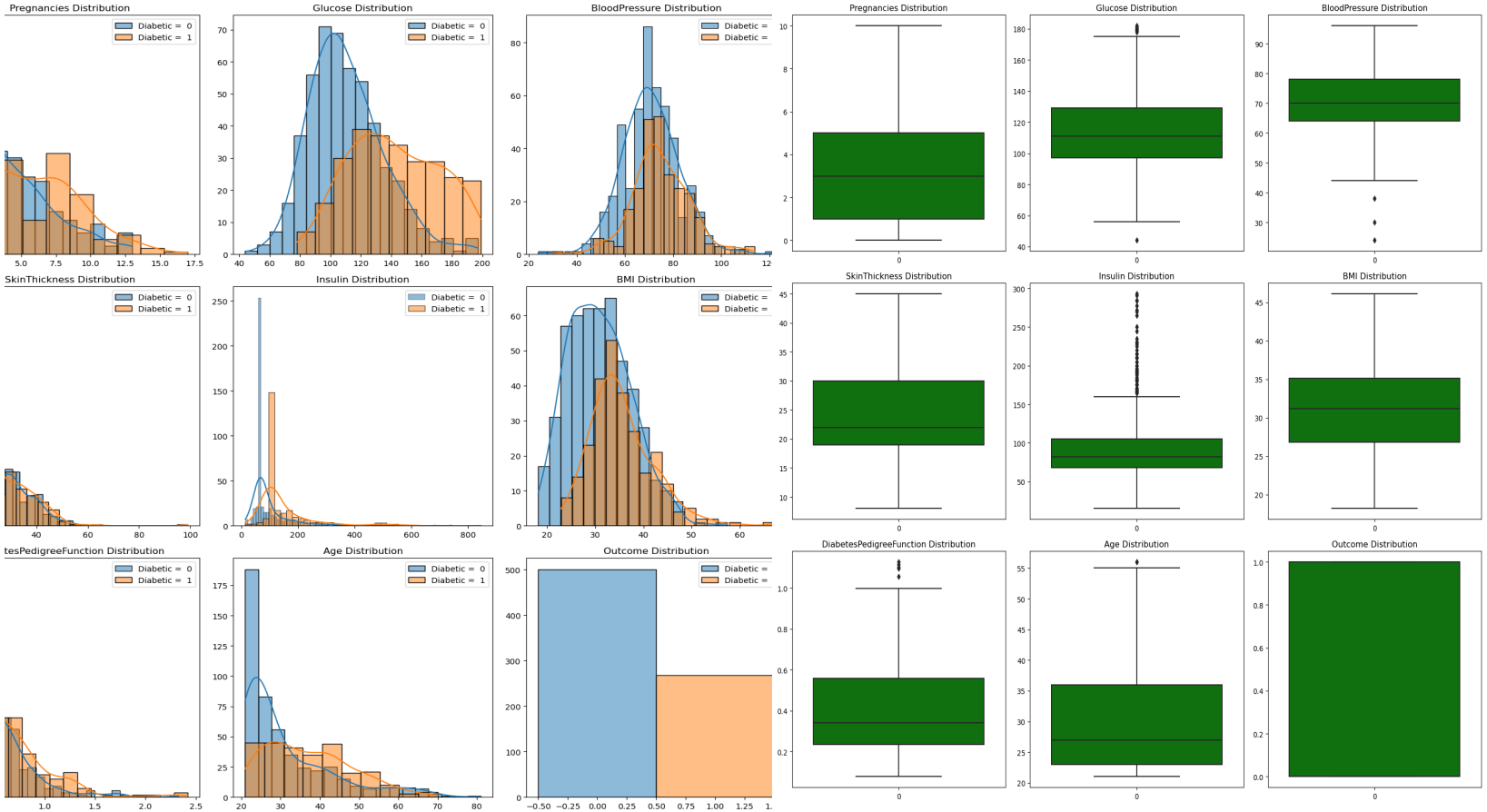
# Supervised Learning Project

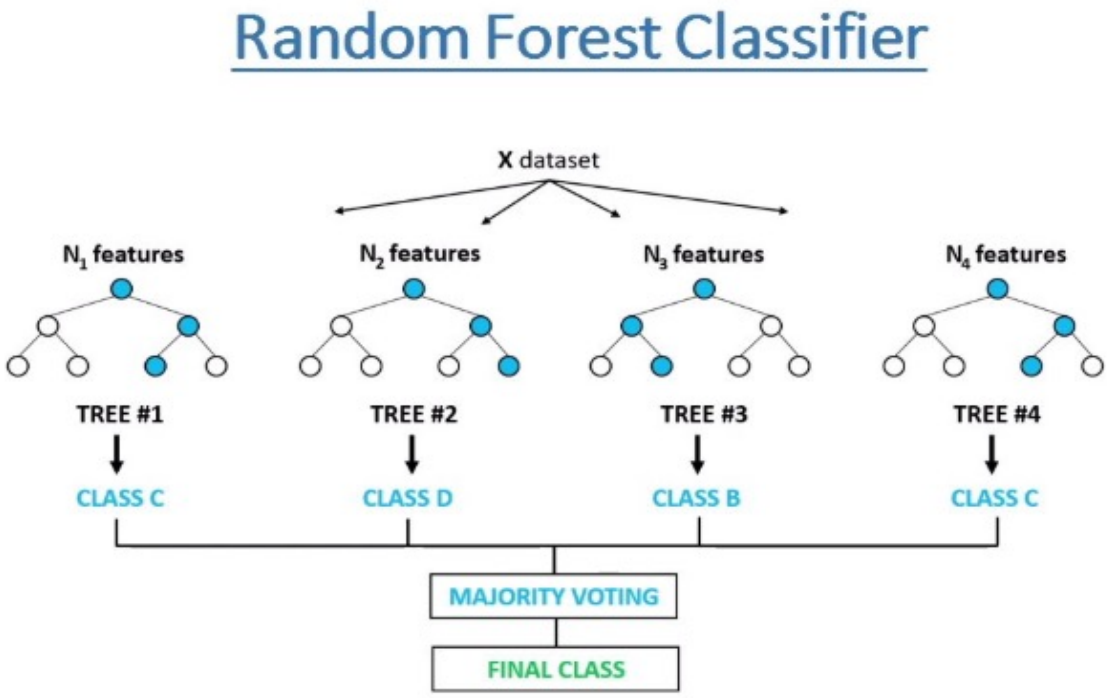
By Skyler Wilson

# Exploratory Data Analysis



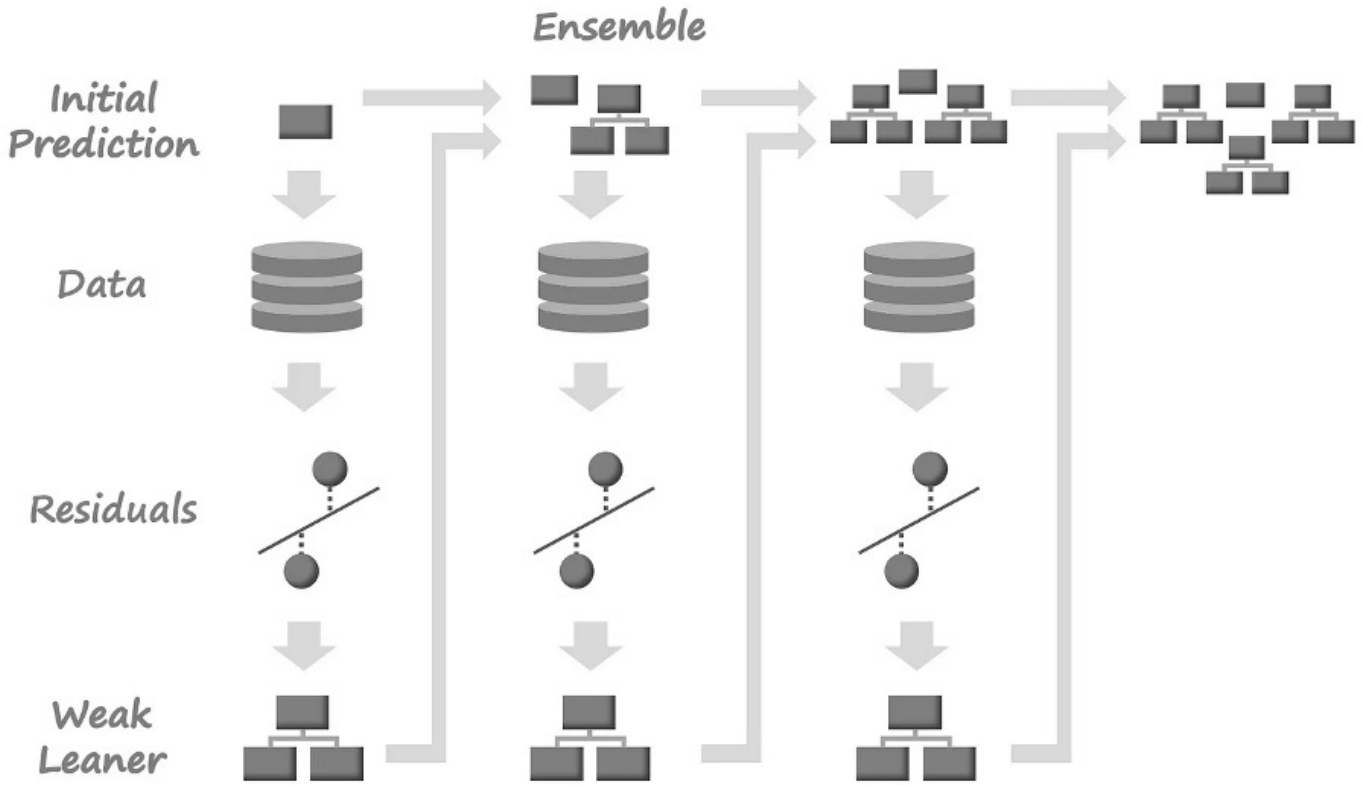



# Random Forest Classifier




A random forest classifier is an ensemble method that aggregates several decision trees together to yield a more accurate machine learning model that is less biased and smaller variation

# XGBoost Classifier




XGBoost classification is an ensemble method that aggregates multiple ‘weak models’ to form a more robust and accurate final model. XGBoost uses extreme gradient boosting to iteratively improve the models fit

# Key Takeaways

The model accuracy was slightly better for the RandomForest than XGBoost at predicting whether an individual has diabetes based on the feature variables provided

The three most important features for prediction according to the RandomForestClassifier were Insulin, Glucose, and SkinThickness whereas in the XGBoostClassifier the three most important features for prediction were Insulin, Age, and BMI

While the RandomForestClassiffier has a higher accuracy score and precision , XGBoost has a better recall score, and the F1 scores are very close in value. The reason the recall value difference is important is because it means that there are fewer false negatives, this is important because a false negative classification for an individual with diabetes is life threatening, since they will presumably not receive adequate treatment.

According to the heatmap of correlation the most highly correlated feature with the outcome was the level of Glucose, this is interesting because it was not the most important feature in either of the models