Decision Tree & Ensemble – Assignment 7
ID: eo9232
Name: Md Reza
IE7860 – Winter 2022

Project Report

Background and Methods:

Diabetes Prediction: This study proposes machine learning approaches to build decision trees and ensemble models that include Decision Tree, Random Forest, Bagging, and XGBoost to predict patients with diabetes and compare models' performance.

Pre-processing:

Diabetes Dataset	**1.** Check for missing values **2.** Impute missing values with mean **3.** Scaled the dataset

Feature Importance:

Model Name	Grid Search with Cross-Validation	Justification	Best Parameters
Random Forest	GridSearchCV (From Scikit-learn)	To find the best activation function, optimizer, & other parameters.	`Best parameters: {'bootstrap': False, 'max_depth': 7, 'max_features': 'auto', 'max_leaf_nodes': 600, 'min_samples_split': 14, 'n_estimators': 100}`
XGBoost	GridSearchCV (From Scikit-learn)	To find the best activation function, optimizer, & other parameters.	`Best parameters: {'colsample_bytree': 0.9, 'gamma': 0.2, 'max_depth': 5, 'reg_alpha': 0.1, 'subsample': 0.8}`

MLP Model Evaluation:

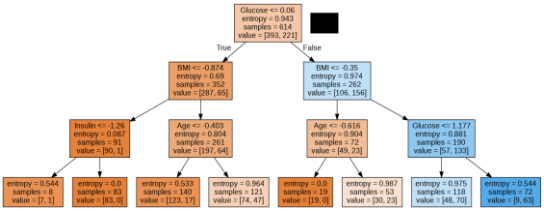
One way to measure models' performance is to compare the accuracy score and confusion matrix derived from models' predictions based on inputs, targets, and other required parameters. To evaluate model performance, the accuracy of each model is compared, and predictions are plotted with a confusion matrix. For details, please refer to the model-based observations plot below. Please refer to the **Model Summary Table** in the appendix below for more information on model performance.

Summary:

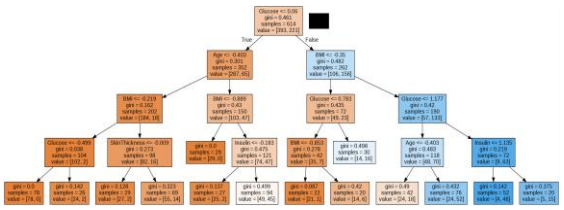
After comparing models' output, it turned out that accuracy scores and confusion matrix play a crucial role in evaluating the performance. The goals were to build the **Decision Tree & Ensemble** models to demonstrate the power of machine learning in terms of classification & prediction. The observations-based models' outputs are as follows:

Decision Tree: The decision tree with entropy has 76% accuracy, but the accuracy slightly improved to about 78% with pruning. The comparisons between the decision tree and pruned decision tree are demonstrated in the figures below:

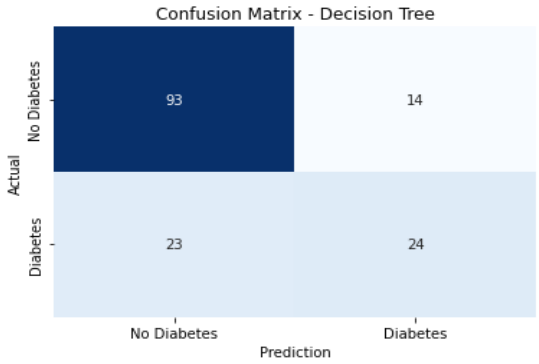
Single Tree



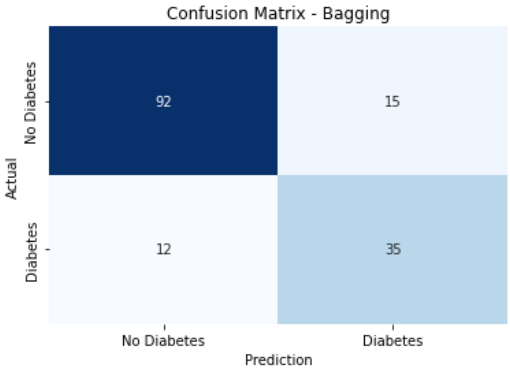
Pruned Tree



However, with a 76% accuracy, the Confusion Matrix could detect many possible diabetes cases.

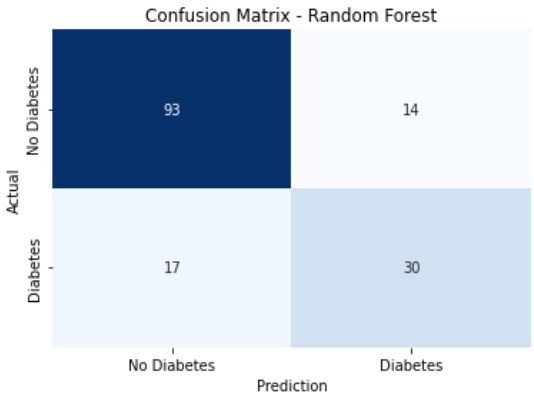


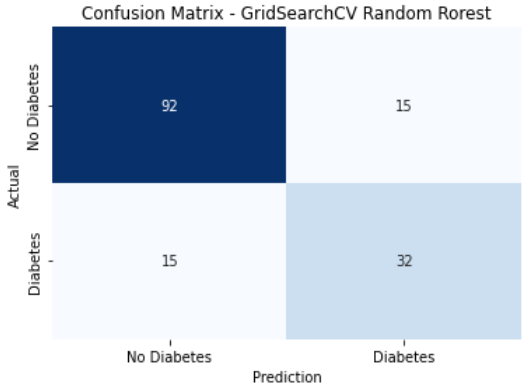
Bagging: Bagging (Bootstrap Aggregation) creates several subsets of training data and randomly replaces them that helping reduces the variance of a decision tree and improve performance.



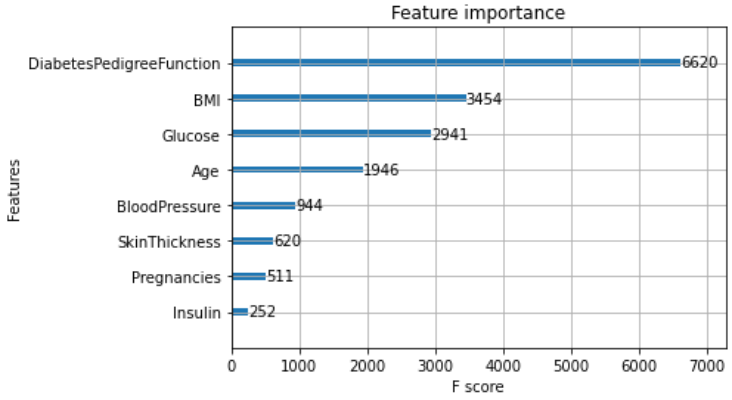
It turns out that Bagging improved the accuracy, and with 82% accuracy, the prediction on the confusion matrix has slightly improved.

Random Forest: The random forest has 80% accuracy, but the accuracy slightly improved to about 81% with important features from grid search cross-validation. The comparisons between the random forest and hyper tuned random forest are demonstrated in the figures below:





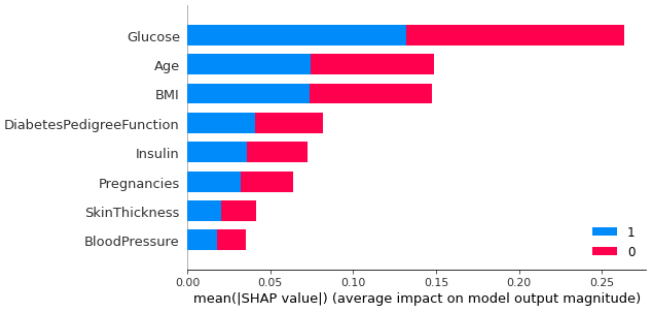
XGBoost: The XGBoost has 78% accuracy, but the accuracy slightly improved to about 81% with important features from grid search cross-validation.



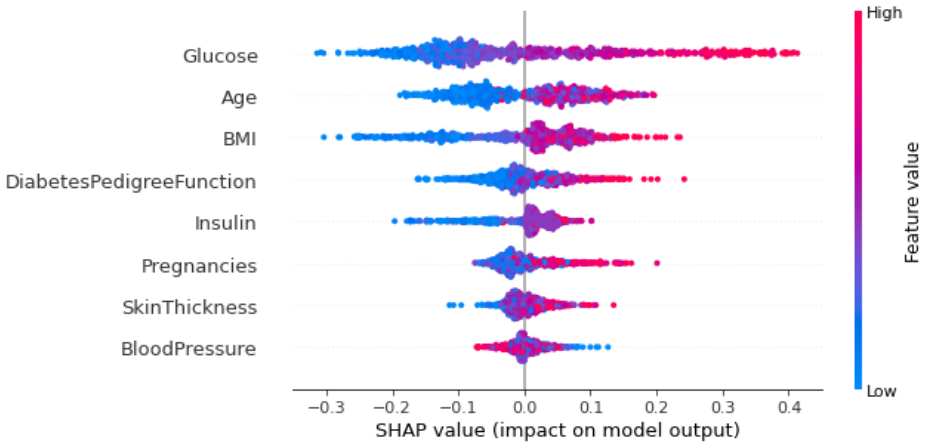
Features Importance Using Explainable AI (SHAP):

SHAP (Shapley Additive Explanations) connects optimal credit allocation with local explanations using the game-theoretic approach and classical Shapley values to explain individual predictions or output of the machine learning model. The observation-based random forest model performance with SHAP is as follows:

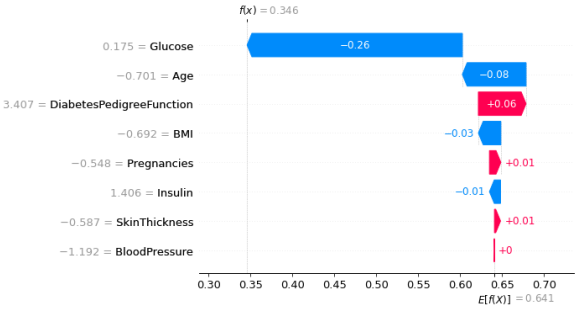
Mean SHAP calculates the mean of absolute SHAP values for each feature across all observations. Then, each feature plot with a separate bar. The figure below shows that Glucose has the largest mean SHAP compared to other features.



SHAP Beeswarm overcame some of the issues derived from the SHAP Mean. It combines feature importance with feature effects to highlight important relationships. Again it turns out Glucose highlight an important relationship in predicting diabetes case.



SHAP Waterfall Plot shows how much each feature has increased or decreased the predicted number of rings for a specific observation. For example, the x-axis indicates that the base value is E[f(x)] = 0.641 is the average predicted number of rings across all 768 observations.



SHAP Force Plot provides the same information as the SHAP Waterfall plot. The figure below starts at the same base value of 0.640, and each feature increases/decreases the predicted value to give us the final prediction of 0.35.

APPENDIX:

Model Summary:

Model Name	Accuracy Score
Decision Tree	0.7597
Decision Tree Pruned	0.7792
Bagging	0.8247
Random Forest	0.7987
Random Forest with Feature Importance	0.8052
XGBoost	0.7792
XGBoost with Feature Importance	0.8052

SHAP Dependence Plot shows the marginal effect of feature(s) on the predicted outcome of a machine learning model.