DAV 6150 Module 11 Assignment

Decision Trees & Random Forests

***** You may work in small groups of no more than three (3) people for this Assignment *****

We've learned that decision trees and random forest models can both be very effective when applied to classification problems, and that random forests can often improve upon the performance of a single individual tree by constructing a large number of individual decision trees via bootstrap aggregation and using their collective output to arrive at a predicted data value. While the performance of a random forest model is usually expected to be better than that of a single decision tree, there is an obvious complexity vs. performance tradeoff we must assess when deciding whether to implement a single decision tree vs. a random forest: random forests are much more computationally complex and generally more difficult to explain/interpret than are individual decision trees.

Your task for the **Module 11 Assignment** is to compare the performance of a decision tree vs. a random forest. The data set you will be working with is a well-known set of attributes that describe the physical characteristics of mushrooms. The data is sourced from the UCI data repository:

- https://archive.ics.uci.edu/ml/datasets/mushroom

The data set is comprised of a total of 23 attributes. Please refer to the UCI web page for further details on these variables. This data set has been widely used for demonstrating machine learning classification algorithms. One of the attributes within the data set is a binary variable that indicates whether or not a given mushroom is poisonous, and the vast majority of machine learning examples derived from the data set have typically been directed at predicting whether or not a given mushroom is poisonous. However, for this assignment the **veil-color** attribute will serve as the response variable for your models. As such, your decision tree and random forest models should be designed for purposes of predicting which of the four **veil-color** values is most likely to apply to a given observation.

Get started on the Assignment as follows:

1) Load the provided **M11_Data.csv** file to your DAV 6150 Github Repository.
2) Then, using a Jupyter Notebook, read the data set from your Github repository and load it into a Pandas dataframe. Ensure your data attributes are properly labeled within the data frame.
3) Using your Python skills, perform some basic exploratory data analysis (EDA) to ensure you understand the nature of each of the variables (including the response variable). Your EDA writeup should include any insights you are able to derive from your statistical analysis of the attributes and the accompanying exploratory graphics you create (e.g., bar plots, box plots, histograms, line plots, etc.). You should also try to identify some preliminary predictive inferences, e.g., do any of the explanatory variables appear to be relatively more "predictive" of the response variable? There are a variety of ways you can potentially identify such relationships between the explanatory variables and the response variable. It is up to you as the data science practitioner to decide how you go about your EDA, including selecting appropriate statistical metrics to be calculated + which types of exploratory graphics to make use of. Your goal should be to provide an EDA that is thorough and succinct without it being so detailed that a reader will lose interest in it.
4) Using your Python skills, apply your knowledge of feature selection and dimensionality reduction to the 22 candidate explanatory variables to identify variables that you believe will prove to be relatively

useful within your models. Your work here should reflect some of the knowledge you have gained via your EDA work. While selecting your features, be sure to consider the tradeoff between model performance and model simplification, e.g., if you are reducing the complexity of your model, are you sacrificing too much in the way of accuracy (or some other performance measure)? The ways in which you implement your feature selection and/or dimensionality reduction decisions are up to you as a data science practitioner to determine: will you use filtering methods? PCA? Stepwise search? etc. It is up to you to decide upon your own preferred approach. Be sure to include an explanatory narrative that justifies your decision making process.

5) After splitting the data into training and testing subsets, use the training subset to construct at least two different decision tree models and two different random forest models using different combinations of the explanatory variables (or the same variables if they have been transformed via different transformation methods). **Your models must each include at least four (4) explanatory variables.**

6) After training your various models, decide how you will select the "best" classification model from those you have constructed. For example, are you willing to select a model with slightly lower performance if it is easier to interpret or less complicated to implement? What metrics will you use to compare/contrast your models? Evaluate the performance of your models via cross validation using the training data set. Then apply your preferred model to the testing subset and assess how well it performs on that previously unseen data.

Your deliverable for this Assignment is your Jupyter Notebook. It should contain a combination of Python code cells and explanatory narratives contained within properly formatted Markdown cells. The Notebook should contain (at a minimum) the following sections (including the relevant Python code for each section):

1) **Introduction (5 Points)**: Summarize the problem + explain the steps you plan to take to address the problem

2) **Exploratory Data Analysis (15 Points)**: Explain + present your EDA work including any conclusions you draw from your analysis, including any preliminary predictive inferences. This section should include any Python code used for the EDA.

3) **Data Preparation (10 Points)**: Describe + show the steps you have taken to address the data integrity + usability issues you identified in your EDA, including any feature engineering techniques you have applied to the data set. This section should include any Python code used for Data Preparation.

4) **Prepped Data Review (5 Points)**: Explain + present your post-Data Prep EDA analysis. This section should include any Python code used for re-running your EDA on the variables adjusted during your Data Preparation work.

5) **Decision Tree + Random Forest Modeling (40 Points)**: Explain + present your decision tree and random forest modeling work, including your feature selection / dimensionality reduction decisions and the process by which you selected the hyperparameters for your models. This section should include any Python code used for feature selection, dimensionality reduction, and model building.

6) **Select Models (15 Points)**: Explain your model selection criteria. Identify your preferred model. Compare / contrast its performance with that of your other models. Discuss why you've selected that specific model as your preferred model. Apply your preferred model to the testing subset and discuss

your results. Did your preferred model perform as well as expected? Be sure include any Python code used as part of your model selection work and to frame your discussion within the context of the classification performance metrics you have derived from the models.

7) **Conclusions (10 Points)**

Your Jupyter Notebook deliverable should be similar to that of a publication-quality / professional caliber document and should include clearly labeled graphics, high-quality formatting, clearly defined section and sub-section headers, and be free of spelling and grammar errors. Furthermore, your Pythion code should include succinct explanatory comments.

Upload your Jupyter Notebook to your online DAV6150 GitHub directory. Be sure to save your Notebook using the following nomenclature: **first initial_last name_M11_assn**" (e.g., J_Smith_M11_assn). Then submit the resulting web link via Canvas within the Module 11 Assignment page. ***Small groups should identity all group members at the start of the Jupyter Notebook and each team member should submit their own copy of the team's work within Canvas.***