Assignment #4
Machine Learning – Ensemble Methods & Fairness and Bias Evaluation

Overview:

The goal of this assignment are:

1. Further develop your understanding of the data science/ML process
2. Familiarize yourself with Python libraries including: numpy, sklearn, matplotlib, seaborn and others.
3. Be able to handle common tasks such as file parsing and feature transformation
4. Gain familiarity with Ensemble classifiers.
5. Handle a varied data with bias present in the data.
6. Be able to plot and visually explore the different data characteristics
7. Gain familiarity with the importance of ethics in AI.

You may collaborate on this with others but you must follow the rules as outlined in the syllabus and not examine or share code or copy/plagiarize from others online. Cite any references.

Submission:

Upload a zip file named your 'fname_lastname_A4' containing:

1. Output directory to store all outputs your code generates in terms of results, charts, etc. that you want to persist.
2. Jupyter Notebook with code and appropriate functions and comments. Please make sure your notebook can access the data needed without the instructor having to make changes to file paths. i.e. set relative file paths and assume there is an input directory.
3. A pdf file with a summary of your analysis as outlined later in the assignment tasks below.

Background (For Reference Only):

1. Anaconda Installation: https://www.anaconda.com/products/individual
   This is the easiest toolkit that millions of data scientists use for individual work. It is open-source and one stop shop for most libraries you will need. Once you install conda you get an environment with notebooks etc. It has both UI portal or you can use CLI of your OS of choice be it windows, mac or linux.
2. Jupyter Notebooks: The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. (jupyter.org description)
3. A notebook is made up of sequence of cells. Cell can be Code or cell can be Markdown. We use code cells running python 3.9 or the default version that comes with your anaconda install.
4. You can save jupyter notebooks as html files.
5. Download and review closely this example machine learning notebook by Randall S. Olson:
   https://nbviewer.jupyter.org/github/rhiever/Data-Analysis-and-Machine-Learning-Projects/blob/master/example-data-science-notebook/Example%20Machine%20Learning%20Notebook.ipynb
   This notebook is a great starting point to start thinking through the process of data science and machine learning. It uses the IRIS data.

Additional Library you will need:

- The fairMLHealth Library – This is an open source library tool for evaluating fairness and bias in machine learning.

Assignment #4
Machine Learning – Ensemble Methods & Fairness and Bias Evaluation

Dataset:

In this assignment the dataset we will use is the **South German Credit Dataset.**
https://archive.ics.uci.edu/ml/datasets/South+German+Credit
The basic ML task is to build classifiers to predict good or bad credit standing.

Assignment Tasks:

Part 1 (60 points):

(10 points) Output the Table 1 for the dataset that describes the characterics of the dataset.
More about what is Table 1 here:
https://academic.oup.com/jamiaopen/article/1/1/26/5001910
https://cran.r-project.org/web/packages/table1/vignettes/table1-examples.html

(50 points) Using the South German Credit dataset from homework 2, build the following classification models. Use 10 fold cross-validation.

i) Logistic Regression (10 points)
ii) Decision Tree (5 points)
iii) Naïve Bayes (5 points)
iv) Random Forest (15 points)
v) XGBoost (15 points)

(points given as above) Compute the relative predictive performance (Precision, Recall, F-Score, AUC) and report the averages after 10-fold cross validation in the following table.

ML Method	AUC	Precision	Recall	F-Score
LR				
DT (best max depth)				
NB				
RF (list params)				
XGBoost (list params)				

Part 2 (40 points): Evaluate for fairness and bias

First please review the Study Material from canvas on Fairness issues in machine learning.

Then compute the fairness metrics for the models you developed in part 1 using the fairMLHealth Library.

Outline your conclusions regarding fairness from the results in a separate document.

Guidelines:

Usage to invoke and test your code:

Assignment #4
Machine Learning – Ensemble Methods & Fairness and Bias Evaluation

If downstream functions need the output data then you can persist the output in a dataframe or other data structure. If you create new features you may want to persist them if you find them useful to augment the dataset.

Style requirements:

- a great primer From Google: https://developers.google.com/machine-learning/guides/rules-of-ml
- Coding Style Guidelines for deep learning and python based ML in general. http://deeplearning.net/software/pylearn/v2_planning/API_coding_style.html
- Break your code into multiple functions. Represent decision nodes and decision trees as objects. Here is a decent primer on how to do OOP in Python 3: https://realpython.com/python3-object-oriented-programming/

Checks before you submit:

1. Did you include a README.md?
2. Did you remove debug comments and code and print statements for your own development work?
3. Did you clean up your notebooks and remove any non-trivial functions/portions?
4. Is your code commented for others to consume?
5. Is it easy to follow?
6. Is your code able to run if you upload and package everything in a zip and then open that zip on another machine or in any other directory?

References:

1. https://archive.ics.uci.edu/ml/datasets/South+German+Credit#
2. https://www.kensci.com/blog/fairmlhealth-a-step-in-the-right-direction
3. https://slideslive.com/38923130/the-measure-and-mismeasure-of-fairness-a-critical-review-of-fair-machine-learning