**DATA 430 – Assignment 3: Decision Trees**

**Introduction**

A decision tree is a tool that uses a tree-like model of decisions and their possible outcomes. Each node represents a test criterion on an attribute and each branch represents an outcome of that test. In this assignment, you will be applying a decision tree analysis as a classification tool. An example where a decision tree classifier could be used is to identify loan applicants as low, medium, or high credit risk.

1. **General steps:**

* Review theoretical background and implementation examples, resources are provided on the course content page.
* Select appropriate dataset from the list at the end of this document. Note that you will use the same dataset for Assignment 3: Decision Trees and Assignment 4: Clustering.
* If you are using a dataset that is not from list below, make sure that the data is suitable for the assignment. Ask your instructor if you are unsure. If you choose a dataset from your organization, be sure to obtain permission. Furthermore, be sure to remove any confidential information and all identifying and sensitive data such as names, addresses, social security numbers, passport numbers, etc.
* Perform your analysis and fit a decision tree model in python.
* Run an analysis, perform evaluation, and capture the results.
* Document your findings and analysis in a technical report using the template that accompanies these instructions.

1. **Deliverables with critical areas:**

**Overview**: areas to address:

* **Problem Domain**: give some background and context about the problem domain (application area). For instance, if you are doing the analysis for predicting heart disease, provide some context about the disease and include some interesting statistics about it. Also, discuss how the method is relevant for the chosen problem.
* **Objective**: clearly state the objective of the analysis in relation to the kind of algorithm you are employing. Use specific language as to what question(s) you are trying to answer using the specific analysis/modeling type.

**Analysis**: areas to address:

* **Exploratory Analysis**: describe the data including the source, the collection method and variables. Perform exploratory analysis. Also, select few key variables (including the target variable for supervised learning) and study their distributions using plots such as histograms, box plot, bar chart, etc.
* **Preprocessing**: armed with the exploratory analysis, perform the necessary preprocessing, both general and specific types appropriate for the modeling type being employed.
* **Model Fitting**: explain the key steps and activities you perform to fit the model. Experiment (as appropriate) with parameters tuning. This is key, what separates highly accurate model from a less accurate ones is the amount of performance tuning performed.

**Results**: areas to address:

* **Model Properties:** explain the components of the fitted model and their characteristics. Leverage functions to summarize the model properties. Also, leverage visualization as required.
* **Output Interpretation**: explain the result and interpret the final model output using terms that reflect the application area and in relation to the stated objective. This is where you check whether or not the stated objective is met.
* **Evaluation**: employ appropriate metrics to quantitatively evaluate the performance of the fitted model. For supervised classification, this includes simple accuracy, precision & recall (or sensitivity & specificity), all of which can be generated from a confusion matrix, or ROC.

**Conclusion**: areas to address:

* **Summary**: highlight the main findings in relation to the stated objective. You don’t need to discuss the details of the analysis and the model such as accuracy here, just focus on the key findings.
* **Limitations & Improvement areas**: discuss the limitations of the analysis and identify potential improvement areas for future work. This could be related to the data, algorithm or a combination of the two.

**Miscellaneous**:

* Use the template that accompanies these instructions to submit your responses to each section, with Python code and any extended model outputs in the Appendix section
* Proofread your report for correct structure, grammar, and spelling.
* The report should be entirely in your own words, no direct quotes from any source. However, keep in mind that any original ideas, information or interpretation of your dataset or regarding the general use of any algorithm, method, or model that you may discover from a source must be cited. Follow appropriate APA format and provide all necessary references. If you have any questions about this requirement, please ask your instructor for clarification.
* Graphics, figures, or tables should be titled and explained. For example, screen captures generated should be assigned a figure title and label (e.g. Figure 1.xxx) and have a description associated with that figure providing details and context for the image.
* In addition to any python code that you may include in the Appendix, you should submit all your python script as a separate file when you submit the assignment to LEO

***The total length of the report should be 6-9 pages, single-spaced within the text areas of the template provided, excluding the appendix, and python script. Large code snippets and graphs should be in the appendix.***

**Suitable datasets for Assignment 3: Decision Trees**

The dataset you choose for this assignment will also be used for Assignment 4: Clustering.

<https://archive-beta.ics.uci.edu/ml/datasets/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition>

[http://archive.ics.uci.edu/ml/datasets/Phishing+Websites#](http://archive.ics.uci.edu/ml/datasets/Phishing+Websites)

<https://www.kaggle.com/datasets/prathamtripathi/customersegmentation>

<https://www.kaggle.com/datasets/kukuroo3/body-performance-data>

<https://www.kaggle.com/datasets/alukosayoenoch/cyber-security-breaches-data>

<https://www.kaggle.com/datasets/ruthgn/wine-quality-data-set-red-white-wine>

<https://www.kaggle.com/sujaradha/online-education-system-review>