**DATA 630**

**Assignment 3: Classification (Supervised)**

1. **Introduction**

Classification assigns items in a collection to target categories or classes. Its goal is to accurately predict the target class for each case in the data. One example where a classification model could be used is to identify loan applicants as low, medium, or high credit risks.

We will focus on the following classification approach: Decision Trees

A decision tree is a tool that uses a tree-like model of decisions and their possible outcomes. Each node represents a test on an attribute and each branch represents an outcome of that test. This assignment requires using decision tree analysis as a classification tool to be applied to a data mining study within your domain of interest using R and RStudio. You need to provide a report on this.

1. **Steps to Completion**

For each study the general procedure is to:

* Review theoretical background based on available resources in the course content
* Select a dataset from the module’s recommended datasets list
* Run an analysis, perform evaluation, and capture the results
* Document your findings and analysis in a data mining analytical report

1. **Deliverables**

Submit your analysis report by addressing the following critical areas:

* **Introduction**: give some background and context about the domain of application, provide the rationale for the type of analysis, and state the objective clearly.
* **Analysis**: describe the data both qualitatively and quantitatively through exploratory analysis, perform necessary preprocessing activities, give some intuition about the algorithm and core parameters, demonstrate the model building steps along with parameter tuning, and explain all your assumptions.
* **Result**: explain the result and interpret the model output using terms that reflect the application area, perform model evaluation using the appropriate metrics, and leverage visualization.
* **Conclusion**: summarize your main findings, discuss experimental limitations related to the data and/or implementation of the algorithm, and suggest improvement areas as a potentiation future work.
* **Miscellaneous**:
  + Proof read your report for correct structure, grammar, and spelling
  + Follow appropriate APA formatting and provide all references
  + Include your R script and extended model outputs in an Appendix section.

***The length of the report should be 7-10 pages excluding the title page, appendix and R script.***

1. **Grading Rubric**

|  |  |
| --- | --- |
| Criteria | Weight  (%) |
| Introduction, objective, rationale. | 10 |
| Analysis and Demonstration of Model Development | 40 |
| Result Interpretation and Model Evaluation | 40 |
| Conclusion, limitations, improvement suggestions | 10 |
| Total | 100 |