A blue and white logo

Description automatically generated

Adventist University of Central Africa

MSc in Big Data analytics

Data Mining

Lab: 4

2.1. Decision trees are often used to transform a set of observations into a specific recommended action. Describe the components (nodes, branches) of a decision tree. Why might it be necessary to prune the tree? Why are decision trees an attractive method for classification in practical applications?

2.2. Suppose an organization has built a rule-based classifier using domain knowledge. After collecting a large amount of data, outline the steps required to improve upon the existing approach by constructing a data-driven classifier. How would you advise to test the validity of the new model?

2.3. Consider the challenge of classifying the likelihood of survival using the Titanic dataset (<https://www.kaggle.com/datasets/yasserh/titanic-dataset>) Construct a decision tree and display the structure of this tree using a graphic.

2.4. Evaluate the performance of the tree (before and after pruning) and provide results using cross-validation.

2.5. Compare the final tree with logistic regression and comment on the advantages and disadvantages of both. Which model is best for competing in the Kaggle competition?

2.6. Go to this link <https://www.kaggle.com/competitions/titanic> Started and follow the instructions to register and enter the challenge. After this assignment, you should have two new approaches for classifying survival on the Titanic: a tree and a KNN model.