#### Reports Template: Title Here

Liang Hou, G2401631D  
IN6227-2024-Assignment-1

**INTRODUCTION**

This report presents the implementation of a decision tree algorithm to perform a classification task on the Census Income dataset. The objective is to predict an individual’s income, categorized as either greater than $50K or less than or equal to $50K, based on a range of demographic and employment attributes. The dataset contains 14 features, including: age, work-class, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country. The target variable, income, is divided into two categories: ">50K" and "<=50K". For further details on the dataset, please refer to the Census Income data in the UCI Machine Learning Repository.

In this implementation, the ID3 (Iterative Dichotomiser 3) algorithm is employed to construct the decision tree. The ID3 algorithm selects the optimal feature for splitting at each node by maximizing information gain, which measures the reduction of uncertainty in the target variable. This method builds a tree by recursively partitioning the dataset until each leaf node corresponds to a homogeneous class or other stopping criteria are met.

To evaluate the performance of the decision tree model, the metrics of precision, recall, and F1-score will be utilized. These metrics provide insight into the model's ability to correctly classify the target variable and balance between sensitivity and accuracy in the classification task.

**METHODS OR PROCEDURES**

## Data Preprocessing

The data preprocessing workflow consists of handling missing values, splitting features and labels, normalizing numerical features, and encoding categorical features.

First, for handling missing values, since all features with missing data are categorical, a new category named "other" is created to represent records with missing values. This is achieved by replacing the "?" in the dataset with "other."

Second, the features and labels are separated. The label is the target variable, income, while the remaining 14 attributes serve as features.

Third, the numerical features are normalized. For each numerical column, the mean () are computed. Since numerical features are not ideal for decision tree classification, the next step is to convert these numerical features into categorical ones. Based on the computed mean and standard deviation, each record is assigned to one of the following categories: , , , , , where represents the average value of the column and denotes the standard deviation.

Finally, categorical features are encoded by converting descriptive values, such as "Tech-support" into numerical representations like "1".

## Building the Decision Tree

The ID3 decision tree algorithm is introduced in this report to address the classification task. The algorithm initiates at the root node with the entire training dataset and subsequently selects the best attribute for splitting based on a specified criterion. In this case, the splitting criterion employed is entropy.

Entropy serves as a measure of disorder within a dataset. A dataset exhibiting high entropy indicates that the data points are evenly distributed across various categories, while a dataset with low entropy signifies that the data points are concentrated within one or a few categories. The formula for entropy is expressed as:

* Where represents the fraction of the sample within a particular node.
* – The current dataset.
* – Set of classes in .

In essence, a low entropy value suggests that the data is well understood, whereas a high entropy value indicates that additional information is necessary for clarity. Preprocessing the data prior to utilizing the ID3 algorithm can enhance the overall accuracy of the model. In summary, ID3 aims to reduce uncertainty and facilitate informed decision-making by selecting attributes that provide the most significant insights into the dataset.

Upon selecting the best attribute, the algorithm proceeds to split the data based on the chosen attribute and assesses whether the resulting subsets are pure. If the split data is determined to be pure, the algorithm terminates; otherwise, it continues to perform further splitting using another attribute (feature).

The stopping criteria for the algorithm include two conditions: the current node is pure, or the depth of the tree has exceeded the pre-defined maximum depth threshold.

## Model Evaluation

**RESULTS**

Results are where things can get difficult. If you want to use tables and other approaches go ahead but keep tables within the margins, they do not go after the references but in the text. Moreover, you may use symbols for various statistics. Make sure you are using the correct program to put these in the text.

We strongly encourage authors to carefully review the material posted here to avoid problems with incorrect files or poorly formatted graphics.

# DISCUSSION

You want the brilliance of the work to shine and other things like that. So just wrap it up with how great the findings are. Bring it.

**CONCLUSION**

Although a conclusion may review the main points of the paper, do not replicate the introduction as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions. Since the conclusion is the last part of the paper read, make it memorable.

**REFERENCES**

[1] Seeger, M. W., Sellnow, T. and Ulmer, R. R. 2003. *Communication and organizational crisis*, Westport, CT: Praeger.

[2] Benoit, W. (2018). Crisis and image repair at united airlines: Fly the unfriendly skies. *Journal of International Crisis and Risk Communication Research*, *1*, 11-26. https://doi.org/10.30658/jicrcr.1.1.2