#### Implementation and Evaluation of a Decision Tree Model for Income Classification Using the ID3 Algorithm on the Census Income Dataset

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 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

To assess the performance of the decision tree model, several evaluation metrics were employed, including accuracy, confusion matrix, precision, recall, and F1-score. Each of these metrics offers unique insights into different aspects of the model's predictive capabilities, allowing for a more comprehensive evaluation.

**RESULTS**

The decision tree model was evaluated using several performance metrics, including accuracy, the confusion matrix, precision, recall, and F1-score. Below are the detailed results obtained from testing the model on the Census Income dataset.

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Sample of True Labels | ['<=50K', '<=50K', '>50K', '>50K', '<=50K', '<=50K', '<=50K', '>50K', '<=50K', '<=50K'] |
| Sample of Predicted Labels | ['<=50K', '<=50K', '<=50K', '<=50K', '<=50K', '<=50K', '<=50K', '>50K', '<=50K', '<=50K'] |
| True Labels Distribution | {'<=50K': 12435, '>50K': 3846} |
| Predicted Labels Distribution | {'<=50K': 12603, '>50K': 3678} |
| Correct Predictions | 13,205 |
| Total Predictions | 16,281 |
| Accuracy | 0.81107 |
| Confusion Matrix | [[10,981 1,454], [1,622 2,224]] |
| Precision | [0.8713, 0.6047] |
| Recall | [0.8831, 0.5783] |
| F1 Score | [0.8771, 0.5912] |

# DISCUSSION

The model achieved an accuracy of 81.11%, which demonstrates that it was generally effective in predicting whether an individual's income falls above or below $50K. When evaluating precision, the model demonstrated a high precision for the <=50K class (87.13%), indicating that most of the instances predicted as <=50K were indeed correct. This suggests that the model was effective at minimizing false positives for individuals with income <=50K. However, precision for the >50K class was notably lower (60.47%), meaning that there was a higher rate of false positives when predicting individuals as earning more than $50K. This discrepancy highlights the challenge the model faced in correctly identifying higher-income individuals. In terms of recall, the model achieved 88.31% for the <=50K class, meaning it was highly effective at capturing true positives in this class. Conversely, the recall for the >50K class was significantly lower, at 57.83%, indicating that the model struggled to identify many of the individuals with incomes greater than $50K. This imbalance in recall reflects the difficulty the model had in recognizing individuals with higher incomes, likely due to the fewer number of samples in this category compared to the <=50K class. The F1-score, which balances precision and recall, was 87.71% for the <=50K class and 59.12% for the >50K class. The high F1-score for the <=50K class underscores the model’s overall success in identifying lower-income individuals. However, the relatively lower F1-score for the >50K class confirms the trade-off between precision and recall in this category. The model's performance in predicting higher-income individuals could be improved through further adjustments, potentially by employing techniques such as rebalancing the dataset, fine-tuning hyperparameters, or using ensemble methods to improve decision boundaries.

Overall, the decision tree model demonstrated good performance in predicting lower-income individuals but struggled with identifying those earning more than $50K. Despite its high accuracy, the lower recall and F1-score for the >50K class suggest that improvements could be made in how the model handles the minority class.

**CONCLUSION**

This report presented the application of a decision tree model, utilizing the ID3 algorithm, to address a classification task using the Census Income dataset. The goal of this task was to predict whether an individual's income exceeds $50K or is less than or equal to $50K, based on various demographic and employment attributes.

The resulting model achieved an accuracy of 81.11%, indicating that it performed reasonably well in classifying the income categories. This level of accuracy demonstrates the model's ability to make correct predictions in the majority of cases. However, while the overall accuracy is satisfactory, a more detailed analysis of the evaluation metrics reveals that the model performs better for the majority class (<=50K) compared to the minority class (>50K).

The implementation of the decision tree algorithm can be found in the following GitHub repository: [RiverLiangH/DecisionTree (github.com)](https://github.com/RiverLiangH/DecisionTree)

**REFERENCES**

[1] Prajwala, T. R. (2015). A comparative study on decision tree and random forest using R tool. International journal of advanced research in computer and communication engineering, 4(1), 196-199.

[2] Anyanwu, M. N., & Shiva, S. G. (2009). Comparative analysis of serial decision tree classification algorithms. International Journal of Computer Science and Security, 3(3), 230-2