Machine Learning Course

Decision Tree Assignment Report

Professor:
Dr. Mahdi Eftekhari

Written by:
Amirhossein Abolhassani

Fall 2024

Contents

1 Introduction

A decision tree is a supervised learning model that is widely used in classification problems. The ID3 algorithm is one of the most widely used algorithms for building decision trees. This algorithm uses the Entropy criterion [1] to select the best feature for splitting a node and recursively repeats this process until a base condition is met. In this report, we first examine the dataset and its preprocessing steps, then we explain the Feature Selection method, and finally, we evaluate the results of each tree on a subset of features.

2 Dataset Analysis

2.1 Feature Overview

In this assignment, the dataset named Salary is used. This dataset consists of 32561 instances, with 15 features of individuals along with their annual income class.

Feature Name	Feature Type	Number of Unique Values	Value Example
age	Numerical		50
workclass	Discrete	9	Federal-gov
fnlwgt	Numerical		77516
education	Discrete	16	HS-grad
education-num	Discrete	16	3
marital-status	Discrete	7	Married-spouse-absent
occupation	Discrete	15	Tech-support
relationship	Discrete	6	Wife
race	Discrete	5	White
sex	Discrete	2	Male
capital-gain	Numerical		10566
capital-loss	Numerical		974
hours-per-week	Numerical		88
native-country	4Discrete	2	England
salary	2Discrete	2	<=50K, >50K

Table 1: Features of the salary dataset

2.2 Missing Values

Fortunately, this dataset does not have any missing values.

[1] Entropy

2.3 Charts

The distribution of some features in the dataset has been analyzed. As can be seen in figure 1 , the male population is twice the female population in this dataset.

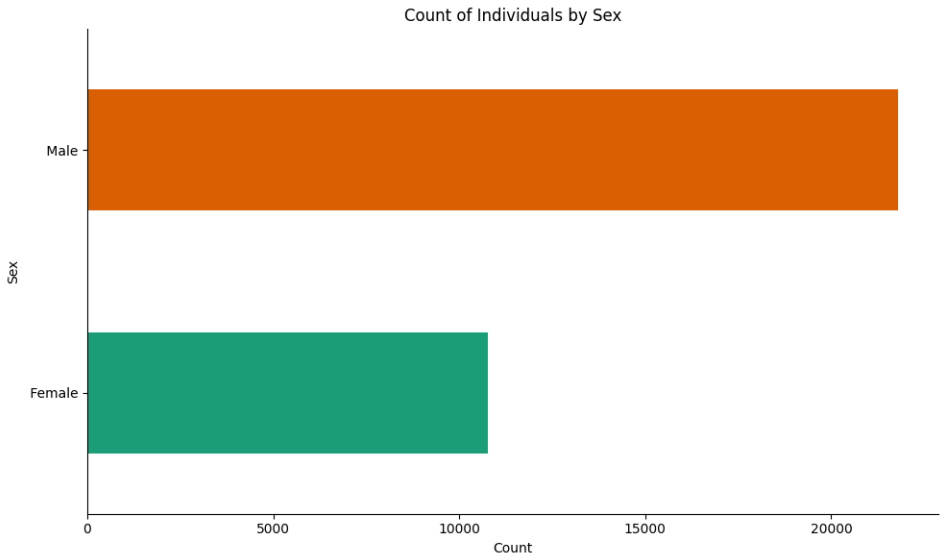

Figure 1: Distribution of the Sex feature

Another feature is the race of each instance in the dataset. As shown in figure 2 , white people are the most frequent and Indian-Eskimo people are the least frequent race specified in this dataset.

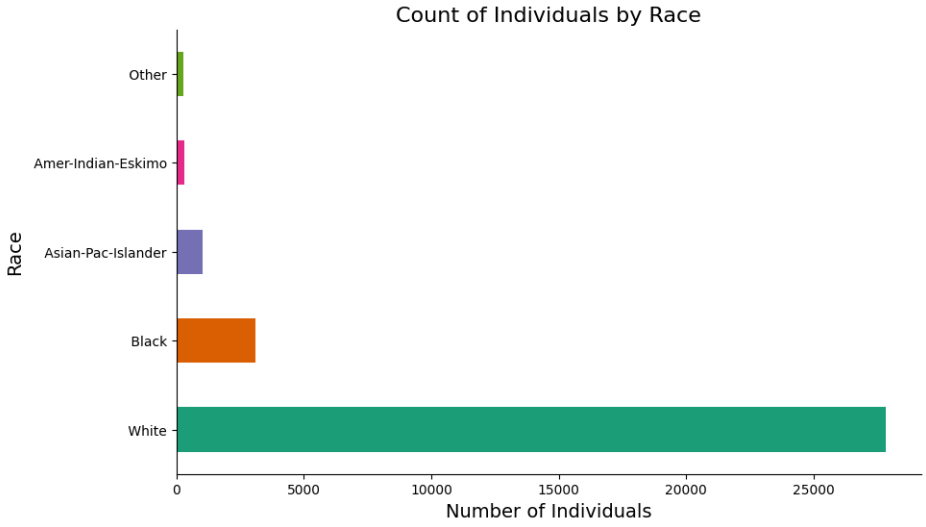

Figure 2: Distribution of the Race feature

One of the most important distributions of this dataset is the distribution of the Age variable. As shown in figure 3 , most instances are in their 30s and 40s. Also, individuals under 10 and over 90 have a very small membership in this dataset.

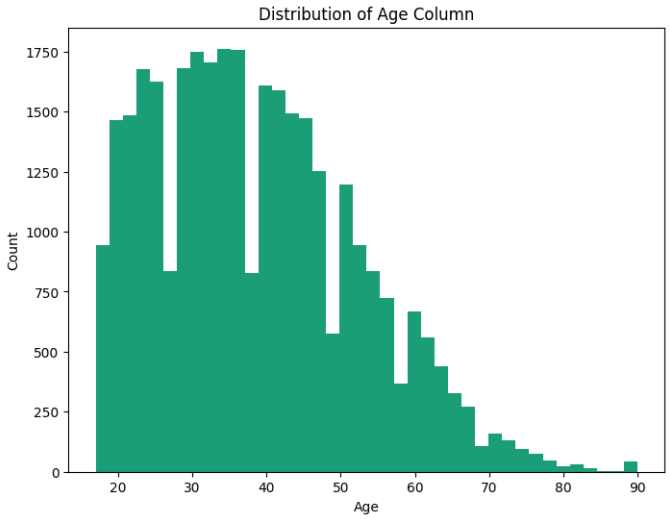

Figure 3: Distribution of the Age feature

Additionally, the distribution of the capital gain and capital loss features can be examined in figures 4 and 5 . Given the financial connection to the topic, these features seem to be related to the target variable.

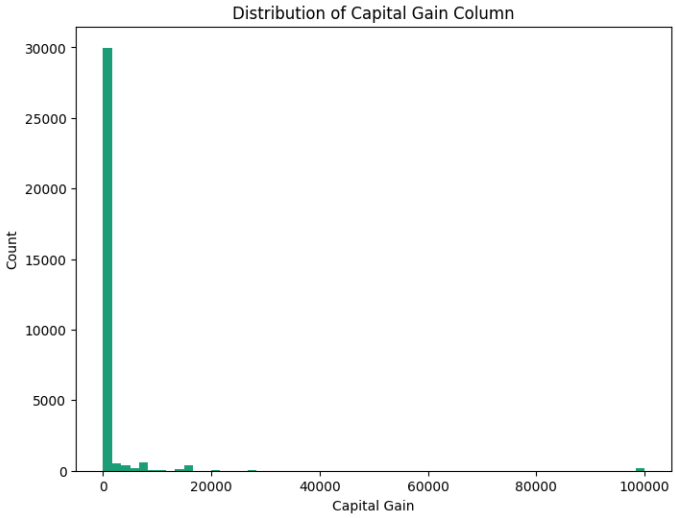

Figure 4: Distribution of the Capital Gain feature

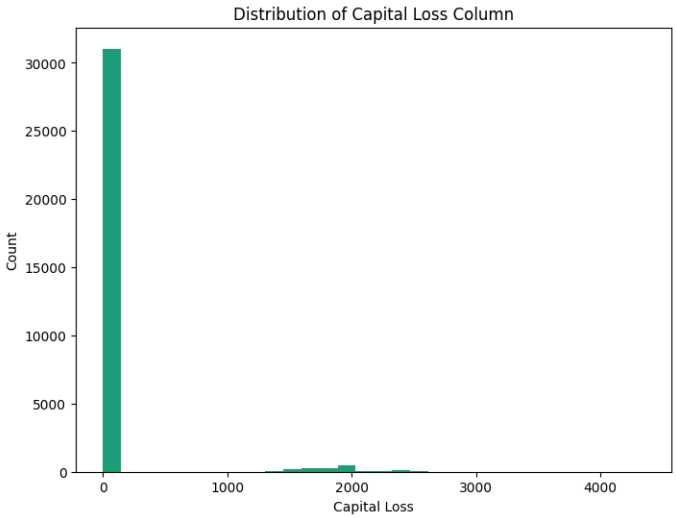

Figure 5: Distribution of the Capital Loss feature

Another important feature is an individual's education level. In countries where logical relationships are reasonably established, people with a higher level of education usually have a good income (figure 7), although the reverse is not always true.

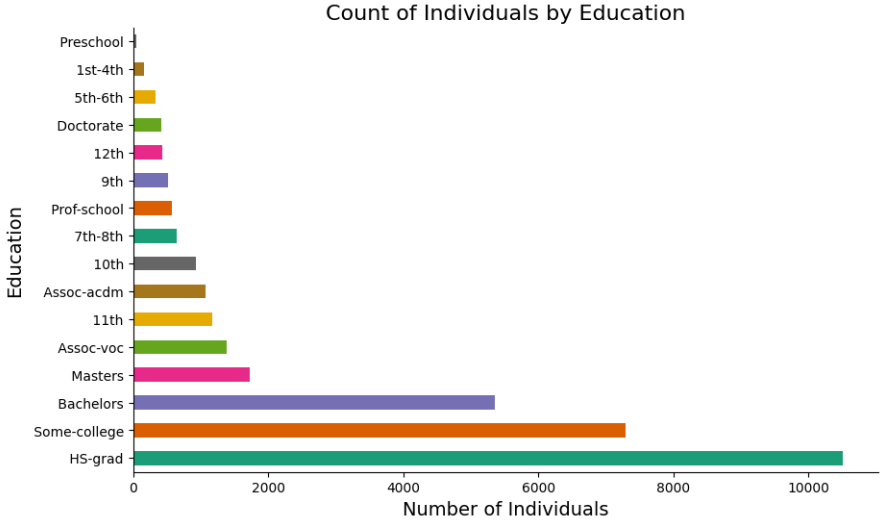

Figure 6: Distribution of the Education feature

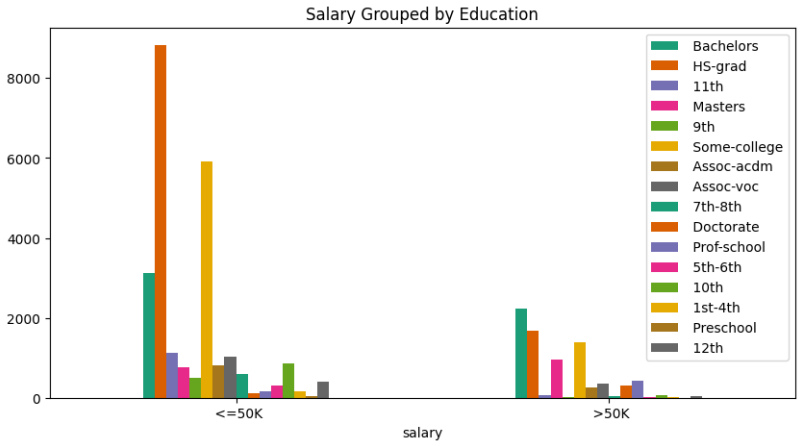

Figure 7: Distribution of the Salary feature based on Education

The distribution of samples' weekly working hours is also shown in figure 8 .

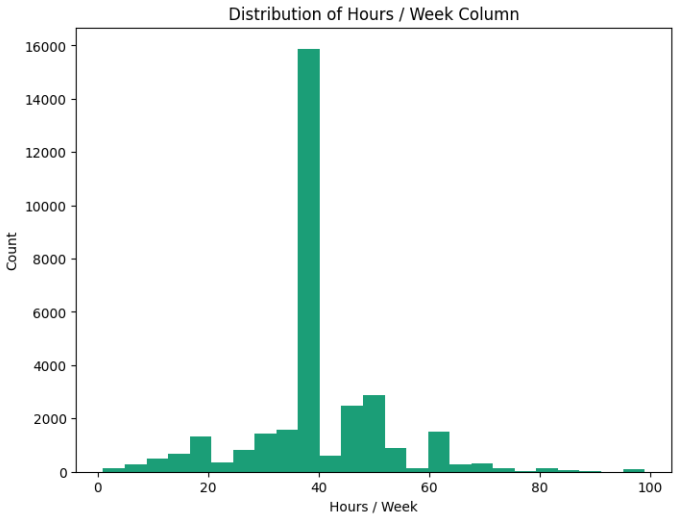

Figure 8: Distribution of the Hours Per Week feature

Another somewhat related feature is people's occupation, whose distribution is shown in figure 9 .

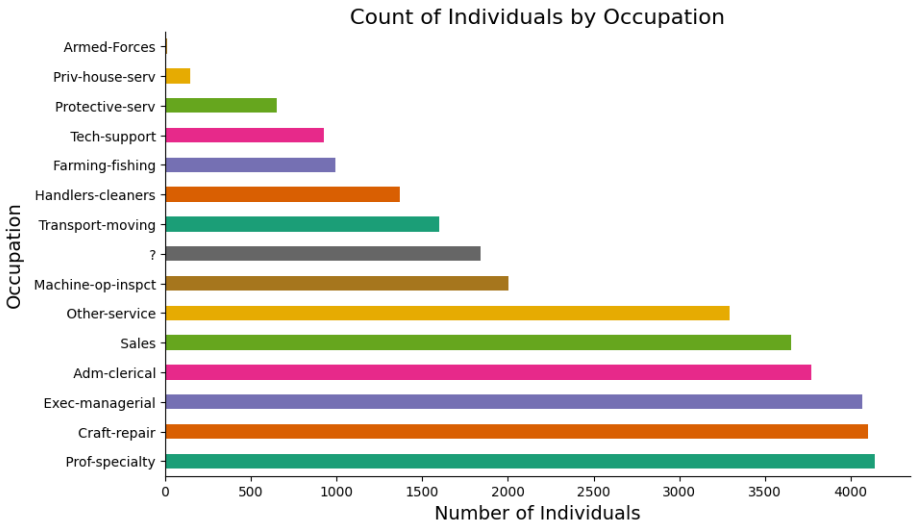

Figure 9: Distribution of the Occupation feature

One of the key features that will be discovered by the tree later on is the Relationship feature (figure 10).

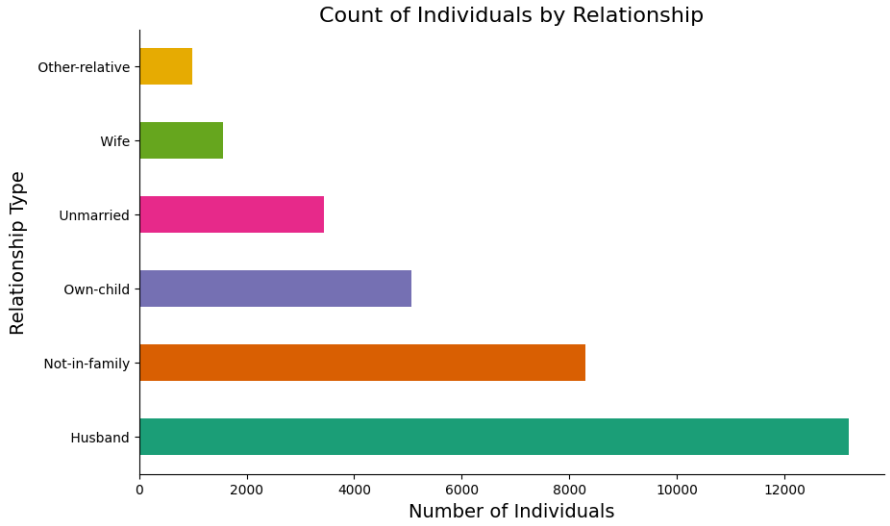

Figure 10: Distribution of the Relationship feature

Finally, to summarize the charts, an attempt has been made to examine the distribution of the target feature classes. As can be seen, the dataset is by no means balanced, and the minor class data corresponds to the higher income class. Also, in figure 11 , the distribution of the target class is shown with respect to three features to gain a better understanding of the relationship between each feature and each target feature class.

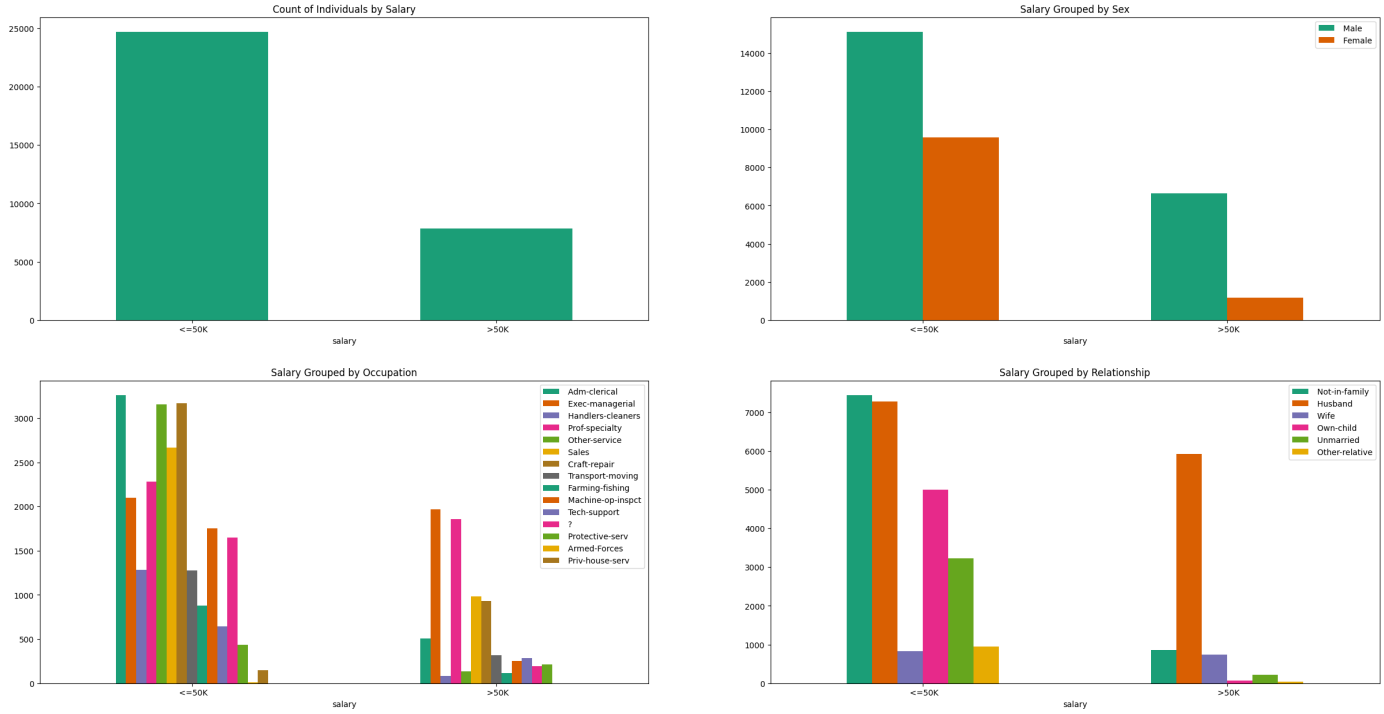

Figure 11: Distribution of the Salary feature according to the Occupation, Relationship, Sex features

2.4 Feature Discretization

To work with a decision tree, the data needs to be discrete. By defining intervals, the Age, Hours per Week, Capital Gain features were discretized. In tables 2 , 3 , and 4 , the values of each feature and the discretization intervals are shown.

$(50, \infty)$	$(30, 50]$	$(0, 30]$
Over 50	31 - 50	1 - 30

Table 2: Discretization of Age

$(60, \infty)$	$(40, 60]$	$(20, 40]$	$(0, 20]$
Very High	High	Average	Low

Table 3: Discretization of $\frac{\text{Hours}}{\text{Week}}$

$(15000, \infty)$	$(0, 15000]$
>15K	<=15K

Table 4: Discretization of Capital Gain

2.5 Manual Removal of Some Features

Here, the reasons for removing the three features fnlwgt , education-num , and capital-loss are explained.

- education-num: This feature is removed because it is identical to the Education feature, causing redundancy.
- capital-loss: By looking at figure 5 , one can infer that the amount of loss individuals have incurred is not significant enough to affect their annual income. On the other hand, the amount of capital gain, especially for some individuals where it is very high, can have a significant impact on a person's annual income.

3 Feature Selection

For this section, a criterion called the Chi-Square Test was used to select a set of features that have the strongest relationship with the target variable. The Chi-Square Test is a widely used statistical method for analyzing categorical data and examining the relationship between discrete variables. This test is extensively used in various fields, including machine learning, data analysis, and scientific research. In the context of feature selection for decision trees and the ID3 algorithm, the Chi-Square Test is used to measure the degree of dependence between features and the target variable. This test helps us determine which features have a stronger relationship with the target variable. The formula for the Chi-Square Test is as follows:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The Chi-Square value for each feature is shown in table 5.

Feature	χ^2
Relationship	3659
Capital Gain	1866
Marital Status	1123
Age	1113
Occupation	504
Sex	502
Education	297
Hours per Week	199
Work class	47
Race	33
Native Country	13

Table 5: χ^2 values for each feature

4 Model Training

Before training the model, we modified the code of the id3 function to add the ability to prune based on a threshold for Information Gain . This helps the model prevent overfitting. It also reduces the tree's depth, which ultimately helps with the tree's memory complexity.

4.1 Dataset Split

For a fair Evaluation phase, the dataset is split into two sections: Train (80%) and Test (20%).

4.2 Model Training on All Features

To have a baseline for comparison, we first train the tree on all features. Also, to see the effect of changing the Information Gain threshold (and thus the effect of overfitting on accuracy and F1 Score), this value was increased from 0 to 2.0 with a step of 0.2 at each stage (figures 12 and 13).

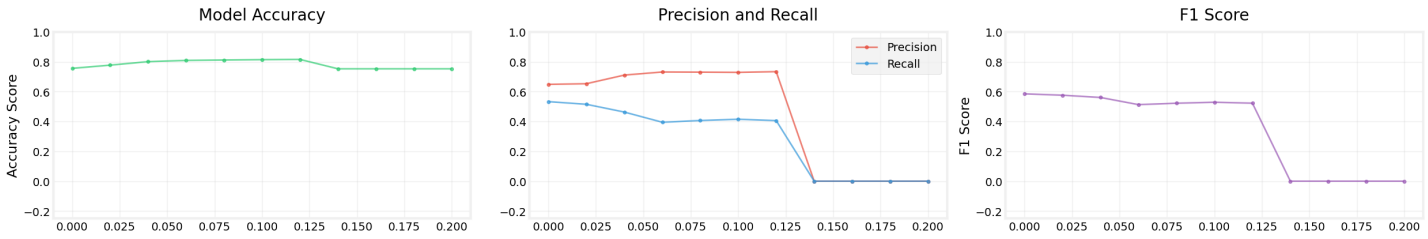

Figure 12: Accuracy, F1 Score, Precision & Recall charts with respect to the Information Gain threshold

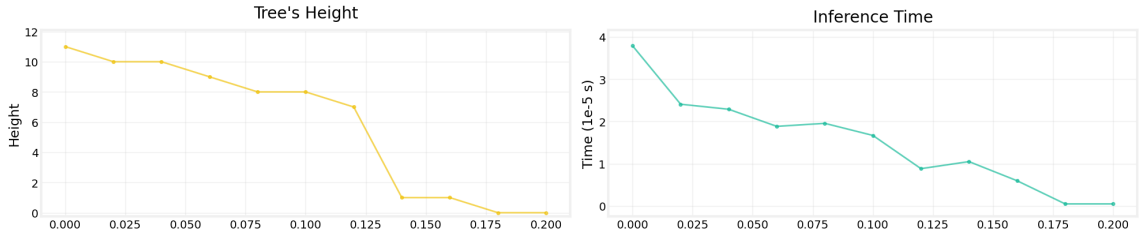

Figure 13: Inference Time and tree height charts with respect to the Information Gain threshold

4.3 Model Training on 4 Features

After classifying the features using the χ^2 criterion, we select four of the best features in this section and build a new decision tree with them. In figures 14 and 15 , different plots are shown for various Information Gain threshold values.

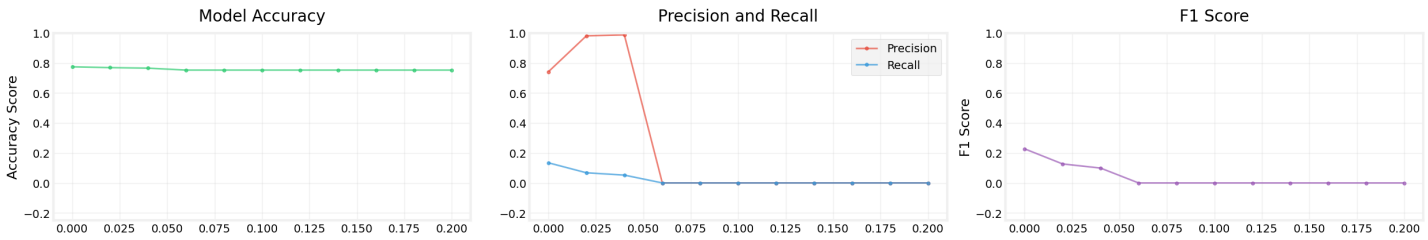

Figure 14: Accuracy, F1 Score, Precision & Recall charts with respect to the Information Gain threshold

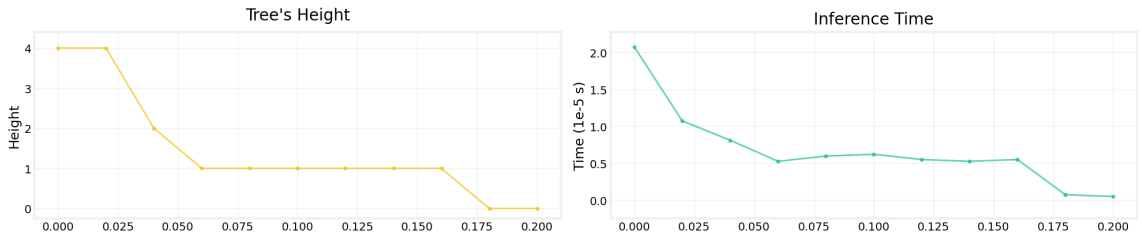

Figure 15: Inference Time and tree height charts with respect to the Information Gain threshold

4.4 Model Training on 8 Features

This time, we select eight of the best features and build a new decision tree with them. In figures 16 and 17 , different plots are shown for various Information Gain threshold values.

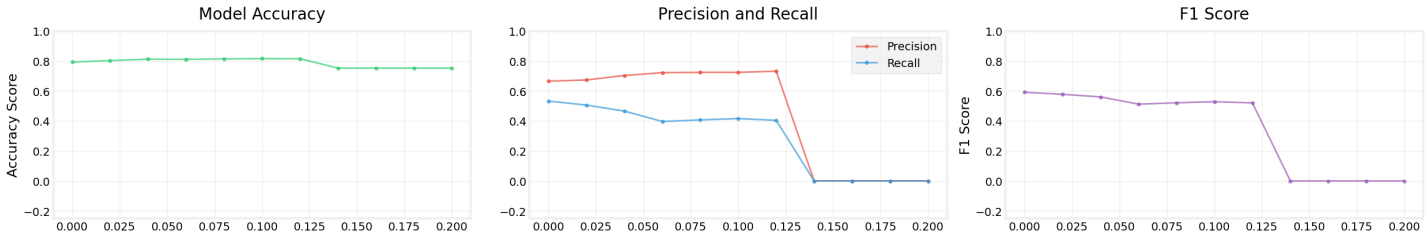

Figure 16: Accuracy, F1 Score, Precision & Recall charts with respect to the Information Gain threshold

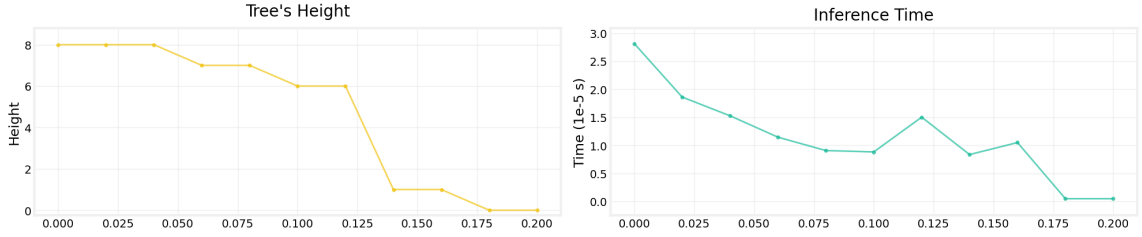

Figure 17: Inference Time and tree height charts with respect to the Information Gain threshold

5 Results

The scores of the best models from the section 4 are compiled in table 6.

Method Used	IG Threshold	Height	($\times 10^{-5}$) Inference Time	Accuracy	Precision	Recall	F1 Score
-	12.0	7	8.0	81.0	73.0	40.0	52.0
χ^2 - Top 8	1.0	6	88.0	81.0	72.0	41.0	53.0
χ^2 - Top 4	0	4	2	77.0	74.0	13.0	22.0

Table 6: Comparison table of the best model scores

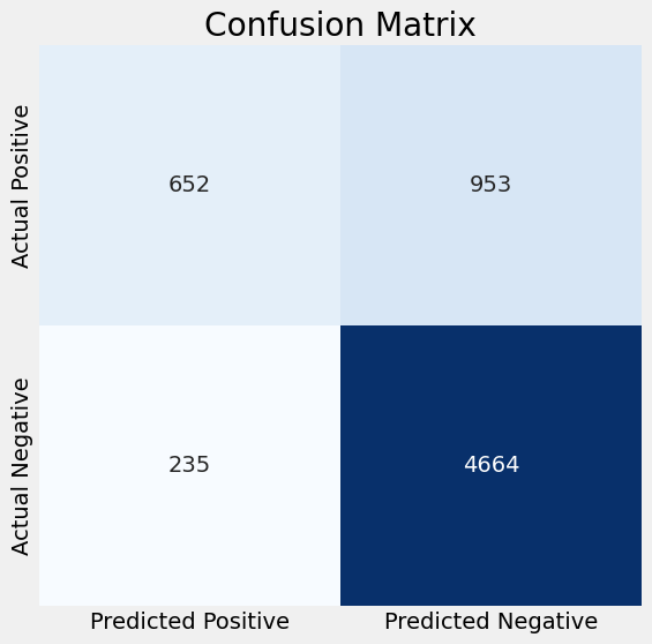

Figure 18: Confusion matrix of a normal tree

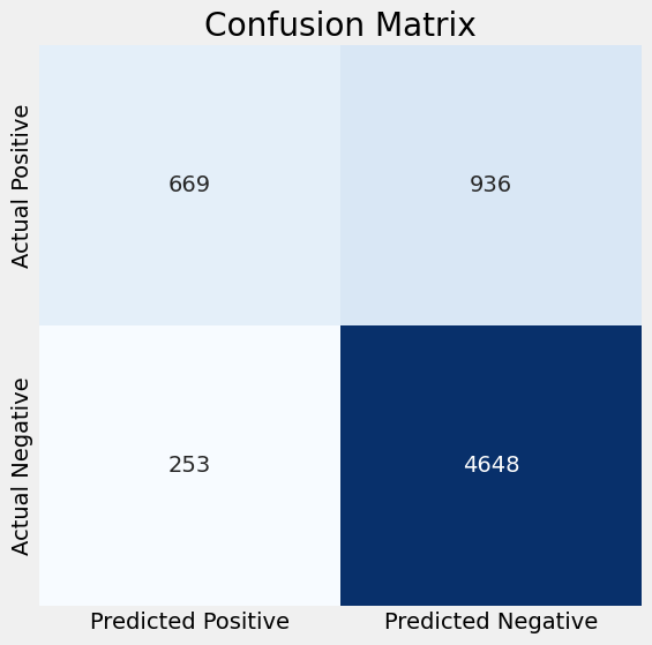

Figure 19: Confusion matrix of a tree built with 8 features

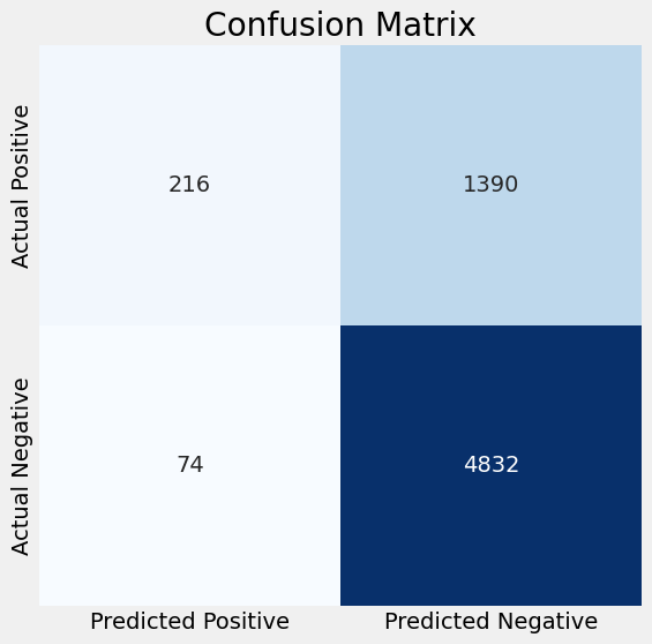

Figure 20: Confusion matrix of a tree built with 4 features

6 Conclusion

As seen in table 6, in this dataset, accuracy drops after 8 features, while the F1 Score initially shows an upward trend but then decreases sharply with 4 features. It can be concluded from table 6 that the model trained with the top 8 features from table 5 and with an IG threshold of 1.0 will perform best. One interpretation for the low Recall value could be the unbalanced nature of the data. The formula for Recall is as follows:

$$Recall = \frac{TP}{TP + FN}$$

When the data is unbalanced and the minor class is the positive class, a bias towards the negative class values occurs, causing many instances to be reported as negative, even if they belong to the positive class (figures 20, 19, and 18). This causes the False Negative value to increase, and as a result, the Recall value to decrease.

7 Curiosity: Dataset Visualization with PCA

Although this was not mentioned in the assignment instructions, the topic of Dimensionality Reduction is truly fascinating! In figure 21, it can be seen that if PCA is performed with two principal components, only a few of the positive class data points are separable from the negative class data.

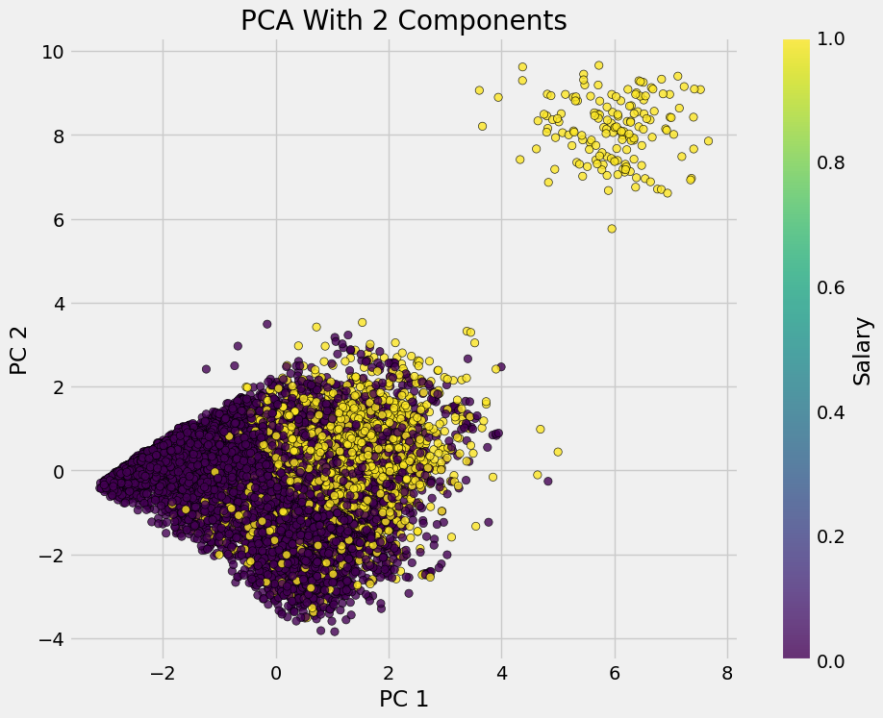

Figure 21: Scatter plot of data after performing PCA with two components

However, with 3 components, it can be seen in figure 22 that the separability of the data from each class increases, as if another aspect of the data is being shown.

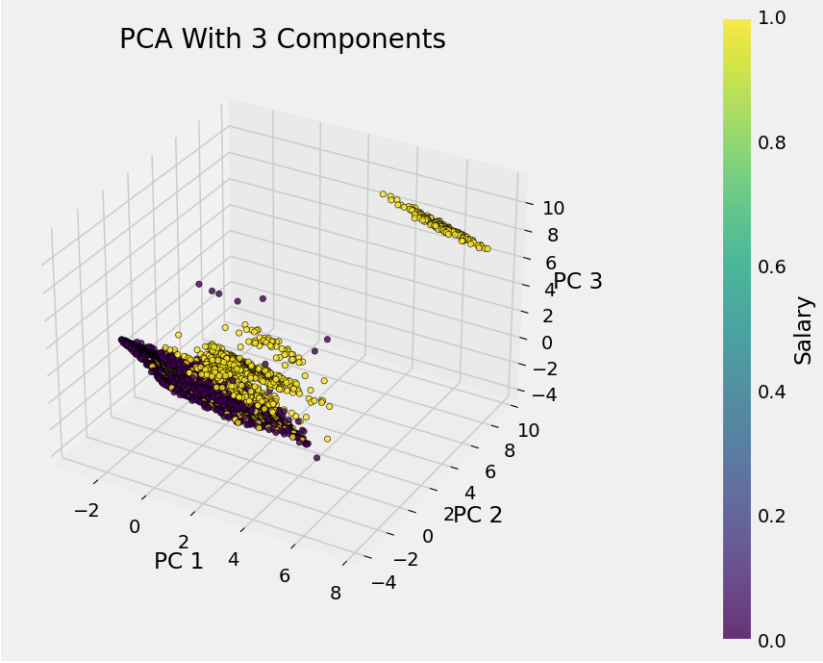

Figure 22: Scatter plot of data after performing PCA with three components