**Project-1 Report**

**Results and Analyses:**

Information Gain:

<=50K >50K

<=50K 581 65

>50K 50 104

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9207607 0.6153846 0.8993808 0.6753247

Precision Recall F1 Prevalence

0.8993808 0.9207607 0.9099452 0.7887500

Detection Rate Detection Prevalence Balanced Accuracy

0.7262500 0.8075000 0.7680727

GINI index:

<=50K >50K

<=50K 580 65

>50K 51 104

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9191759 0.6153846 0.8992248 0.6709677

Precision Recall F1 Prevalence

0.8992248 0.9191759 0.9090909 0.7887500

Detection Rate Detection Prevalence Balanced Accuracy

0.7250000 0.8062500 0.7672803

Naïve Bayes:

<=50K >50K

<=50K 597 79

>50K 34 90

Accuracy = 0.85875

Precision = 0.8831361

Recall = 0.9461173

F1 score = 0.9135425

Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull

8.587500e-01 5.303603e-01 8.326696e-01 8.821516e-01 7.887500e-01

AccuracyPValue McnemarPValue

2.475344e-07 3.485618e-05

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9461173 0.5325444 0.8831361 0.7258065

Precision Recall F1 Prevalence

0.8831361 0.9461173 0.9135425 0.7887500

Detection Rate Detection Prevalence Balanced Accuracy

0.7462500 0.8450000 0.7393308

GINI vs Information gain:

Information gain vs Naïve Bayes:

GINI vs Naïve Bayes:

Analysis after withholding one column for Information Gain:

The column withheld is Capital-gain since, it has the highest information gain among all the other attributes. The following are the results after withholding Capital-gain:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| <=50K >50K | |  |  |  |  |  |  |
| <=50K 567 66 | |  |  |  |  |  |  |
| >50K 64 103  Accuracy = 0.8375  Precision = 0.8957346  Recall = 0.8985737  F1 score = 0.8971519    0.8375000000 | |  |  |  |  |  |  |
| Sensitivity Specificity Pos Pred Value Neg Pred Value | | | | | | | |
| 0.8985737 0.6094675 0.8957346 0.6167665 | | | | | | | |
| Precision Recall F1 Prevalence | | | | | | | |
| 0.8957346 0.8985737 0.8971519 0.7887500 | | | | | | | |
| Detection Rate Detection Prevalence Balanced Accuracy | | | | | | |  |
| 0.7087500 0.7912500 0.7540206 |  |  |  |  |  |  |  |

Analysis after withholding one column for GINI index:

<=50K >50K

<=50K 580 65

>50K 51 104

Accuracy = 0.855

Precision = 0.8992248

Recall = 0.9191759

F1 score = 0.9090909

Kappa AccuracyLower AccuracyUpper AccuracyNull

8.550000e-01 5.512790e-01 8.286702e-01 8.786799e-01 7.887500e-01

AccuracyPValue McnemarPValue

1.058296e-06 2.274246e-01

Sensitivity Specificity Pos Pred Value Neg Pred Value

0.9191759 0.6153846 0.8992248 0.6709677

Precision Recall F1 Prevalence

0.8992248 0.9191759 0.9090909 0.7887500

Detection Rate Detection Prevalence Balanced Accuracy

0.7250000 0.8062500 0.7672803

**Overall Status:**

**File Descriptions:**

The .R files created have been separated by their functions as such:

naiveBayes.R – classification using Naïve Bayes

informationGain.R - Classification using Information Gain and creating the decision tree for it

informationGainHold.R - Classification using Information Gain while holding one attribute (capital-gain) and creating the decision tree for it

gini.R - Classification GINI index and creating the decision tree for it

giniHold.R - Classification using GINI index while holding one attribute (capital-gain) and creating the decision tree for it

Other files include .ps (PostScript files) generated to create the decision trees for as .pdf files.

gain.ps - generates the gain.pdf file showing the decision tree for classification using Information Gain.

gainHold.ps - generates the gainHold.pdf file showing the decision tree for classification using Information Gain while holding one attribute (capital-gain)

gini.ps - generates the gini.pdf file showing the decision tree for classification using GINI index.

giniHold.ps - - generates the giniHold.pdf file showing the decision tree for classification using GINI index while holding one attribute (capital-gain)

**Division of Labor:**

Since this project was completed as a single member team, the entire project was completed by me (Omar Hasan Mohiuddin). The amount of time spent on the project was about 1 week for learning the basics of R and about 10 hours (possibly lesser since this was done over the course of 2 weeks) in total to complete the coding of the project.

**Problems encountered:**

One of the major inconveniences during the project was to eliminate the rows that had a ‘?’ in them. The solution was relatively simple but did not work since each value had a whitespace before it. So, when omitting the rows consisting of ‘?’, it failed to omit any rows since the input had to be ‘ ?’ that is, a whitespace followed by a ‘?’.

cen.data = read.csv("census-adult.csv",na.strings = "?") did not work because of the whitespace before the ‘?’

cen.data = read.csv("census-adult.csv",na.strings = " ?") was the solution to it.

Another problem was whether I needed to do the entire classification procedures from scratch or simply use the algorithms as a means while understanding the process behind it. This was clarified in the group discussions.

Other minor issues were just regarding the compatibility of data structures used in R which were relatively simpler to resolve.