Project-1 Report

Overall Status:

The project was done by splitting the data into test and training sets using the seed value and sample size given. In the case that an attribute needed to be held, the attribute column was removed from the training and test sets. The training set was used to create a model as per the requirements for Naïve Bayes, Information Gain and GINI index. Trees were plotted for the models of Information Gain and GINI index. The model was used to predict the outcome using the test set as input. The result generated was and the income column as a table was used along with the result to generate the confusion matrix and scores for the evaluation of the classifier.

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.
- When applying the formula in the rpart() function as formula = income~.-(attribute to exclude), some attributes were unable to be excluded due to a mismatch in the levels of factors for nominal attributes such as marital-status, occupation and relationship. A quick workaround for this problem was to previously exclude that column in the training and test sets.
- I encountered errors when applying the confusion matrix for labelled vectors (the result was in the form of a vector) which would result in an error when trying to print the $overall data of the confusion matrix. To solve this, I converted the result to a new vector using as.vector(result) for an unlabeled vector.
- Other minor issues were just regarding the compatibility of data structures used in R which were relatively simpler to resolve.

Results:

Information Gain:

	<=50K	>50K
<=50K	**581**	**65**
>50K	**50**	**104**

Accuracy = 0.85625
Precision = 0.8993808
Recall = 0.9207607
F1 Score = 0.9099452
Specificity = 0.6153846
Negative Predictive Value = 0.6753247
Balanced Accuracy = 0.7680727

GINI index:

	<=50K	>50K
<=50K	**580**	**65**
>50K	**51**	**104**

Accuracy = 0.855
Precision = 0.8992248
Recall = 0.9191759
F1 Score = 0.9090909
Specificity = 0.6153846
Negative Predictive Value = 0.6709677
Balanced Accuracy = 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
Specificity = 0.5325444
Negative Predictive Value = 0.7258065
Balanced Accuracy = 0.7393308

Results 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
Specificity = 0.6094675
Negative Predictive Value = 0.6167665
Balanced Accuracy = 0.7540206

Results after withholding one column for GINI index:

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

	<=50K	**>50K**
<=50K	**559**	**64**
>50K	**72**	**105**

Accuracy = 0.83
Precision = 0.8972713
Recall = 0.8858954
F1 score = 0.8915470
Specificity = 0.6213018
Negative Predictive Value = 0.5932203
Balanced Accuracy = 0.7535986

Analyses:

GINI vs Information gain:

GINI

	<=50K	>50K
<=50K	**580**	**65**
>50K	**51**	**104**

Information Gain

	<=50K	>50K
<=50K	**581**	**65**
>50K	**50**	**104**

Both GINI and Information Gain give almost identical results for the test data have different structures for each of their trees

On observing the tree structures of Information Gain and GINI index, they vary in the order of the parent and children nodes. However, the split points for each of the nodes are identical to both trees although they are located at different levels of the tree.

There is a small difference in the number of true positives and the false negatives for the test data even though the split points for both trees are the same. This can be attributed to the difference in tree structures i.e., for the test data, some small percentage of values may be decided early or late due to the difference in sequence of decisions giving us different results.

This difference allows for Information Gain to give slightly better results over GINI index in terms of the difference in F1 score (0.0008543 greater than GINI index)

Information gain vs Naïve Bayes:

Information Gain

	<=50K	>50K
<=50K	**581**	**65**
>50K	**50**	**104**

Naïve Bayes

	<=50K	>50K
<=50K	**597**	**79**
>50K	**34**	**90**

Result Difference (Information Gain – Naïve Bayes):

Accuracy = -0.0025
Precision = +0.0162447
Recall = -0.0253566
F1 Score = -0.0035973
Specificity = +0.0828402
Negative Predictive Value = -0.0504818
Balanced Accuracy = +0.0287419

From the confusion matrix, we can see that Naïve Bayes gives us a significantly better result in terms of number of true positives and the overall F1 score.

However, for naïve bayes the number of false positives and true negatives also has a significant difference over Information Gain which affect the Specificity (True negative rate) and Precision.

For the given training set, naïve bayes has a relatively higher tendency to classify data as a positive result i.e., >=50 resulting in a relatively higher number of true positives as well as false positives. As a direct consequence, the true negatives and false negatives are comparatively lesser and hence the net accuracy sees very little change (0.0025 greater accuracy for Naïve Bayes)

GINI vs Naïve Bayes:

Due to the almost identical results between GINI index and Information Gain, the same comparison and analyses applied between Information Gain with Naïve Bayes can be also applied to the comparison and analysis between GINI index and Naïve Bayes.

Information Gain vs Information Gain after holding one attribute

Information Gain

	<=50K	>50K
<=50K	**581**	**65**
>50K	**50**	**104**

IG – capital-gain

	<=50K	>50K
<=50K	**567**	**66**
>50K	**64**	**103**

Result Difference (Information Gain – Information Gain w/o capital-gain):

Accuracy = 0.01875
Precision = 0.0036462
Recall = 0.022187
F1 Score = 0.0127933
Specificity = 0.0059171
Negative Predictive Value = 0.0585582
Balanced Accuracy = 0.0140521

Capital gain was chosen as it has the highest entropy among the other attributes.

Because of this selection it is expected that there will be an increase in misclassification. The results show a definite increase in misclassification due to the decrease in true positives and true negatives. However, the overall score of the classifier does not significantly change (only by about 3-5% max). This suggests that there exist multiple attributes with a high entropy and capital gain is simply one of them and the highest among them. The other attributes can maintain the quality of the decision tree despite the absence of capital-gain.

On comparing the decision tree structures for both cases, some nodes cannot be further classified due to the absence of capital-gain, but in another instance in the tree age is used as a node for splitting which had not been used when capital gain was included. The sequence of nodes has also changed, hours-per-week now appears before occupation as well as after occupation. As seen in the previous example of GINI vs Information gain, these change suggest a compensation for the absence of the capital-gain attribute.

GINI index vs GINI index after holding one attribute

GINI index

	<=50K	>50K
<=50K	**580**	**65**
>50K	**51**	**104**

GINI index – capital gain

	<=50K	>50K
<=50K	**559**	**64**
>50K	**72**	**105**

Result Difference (GINI index – GINI index w/o capital-gain):

Accuracy = 0.025
Precision = 0.0019535
Recall = 0.0332805
F1 Score = 0.0175439
Specificity = -0.0059172
Negative Predictive Value = 0.0777474
Balanced Accuracy = 0.0136817

Capital gain was chosen as it has the highest GINI index among the other attribute.

Like the above analysis for information gain, it is expected for a decrease in overall score of the classifier. According to the results, there is a significant decrease in the number if true positives, an increase in the number of false negatives and very little change in the in the number of false positives and true negatives.

The above changes suggest that the absence of capital gain as an attribute is related to the increase in number of values classified as negatives instead of positives. The change, however, does not significantly decrease the overall rating of the classifier suggesting that there are multiple attributes which have a high GINI index and capital gain is the highest among them (like information gain).

Like the changes seen for information gain, the GINI index tree structure changes with the removal of capital gain as the change in sequence of nodes, i.e., hours of week appearing before occupation and as a split at the leaf node, and age as a new node to split the data. As mentioned earlier, these changes are a possible compensation for the absence of the capital-gain attribute.