Z Number: 23226341

Name: Carlos Gross-Martinez

Professor: Dr. Taghi M. Khoshgoftaar

Course: CAP6778 – Advanced Data Mining & Machine Learning

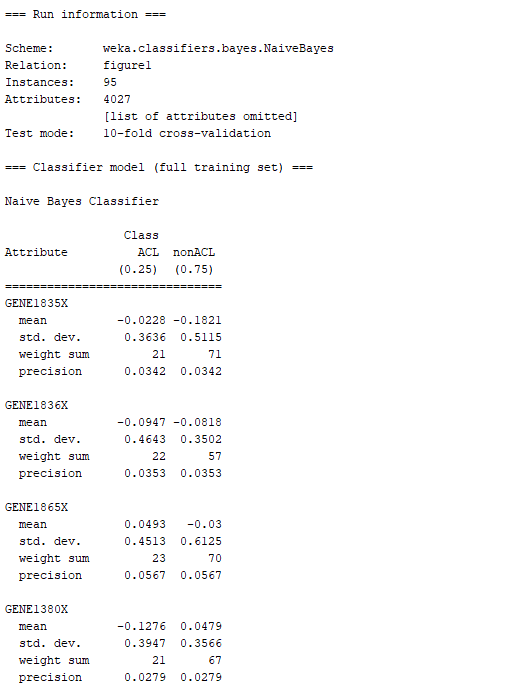
Assignment 3: Feature Selection I

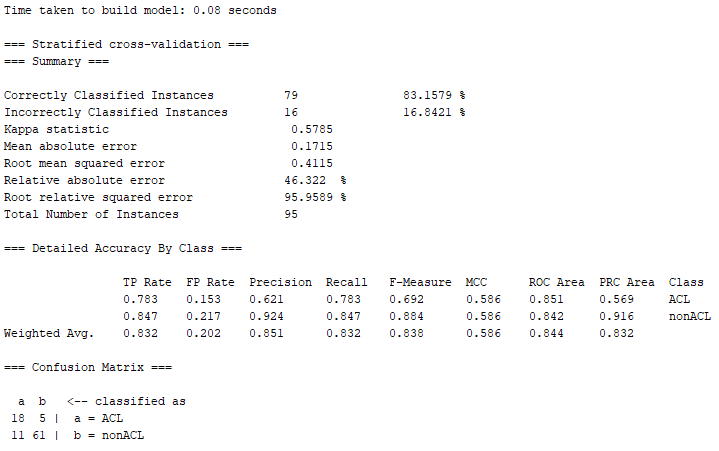
Dataset Analysis

The dataset used in this assignment consisted of 95 instances/samples and 4027 attributes. From the 4027 attributes 4026 were input attributes to the model while 1 attribute was the class label. Furthermore, out of the 95 samples provided in the dataset, 23 were from the minority class labeled “ACL” while 72 were from the majority class labeled “nonACL”. With this information, it can be noted that percentage wise, the minority class represents of the full dataset, while the majority class represents of the full dataset. With this data distribution in mind is important to note that the data set contains contain high dimensionality due to the high number of attributes serving as an input as well as class imbalance, as the data contains a great number of samples for the secondary class “nonACL” while containing a smaller number of samples for the primary class “ACL”.

Part 1: Preliminary Classification

In this segment of the assignment two classification models are trained using the full dataset described above. The algorithms used to build the classifiers include the Naïve Bayes and 5-Nearest Neighbor. Once these two models are built an evaluation of their result will be conducted with emphasis on the false positive rate, false negative rate, and area under ROC curve performance metrics. It is important to note that the only change from the default setting of the models will be conducted in the KNN algorithm where the K value will be updated to 5 in the Weka tool as directed in the assignment. All other settings from both classifiers will remain the defaults. Additionally, the number of folds for cross validation will be set up to 10. With this in mind, the first classifier to be trained was the Naïve Bayes. The results from the training of the model can be seen below:





Picture 1: Naïve bayes Classifier with Full dataset

Based on the information from picture 1 which displays the Naïve Bayes classifier training information, the confusion matrix from the model can be observed in table 1 below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | ACL | nonACL |
| ACL | 18  (TN) | 5  Type II (FN) |
| nonACL | 11  Type I (FP) | 61  (TP) |

Table 1: Confusion Matrix Naïve Bayes Classifier with Full Dataset

With a clear understanding of the confusion matrix for the model, it is not possible to calculate the false positive rate and false negative rate. These calculations can be observed below:

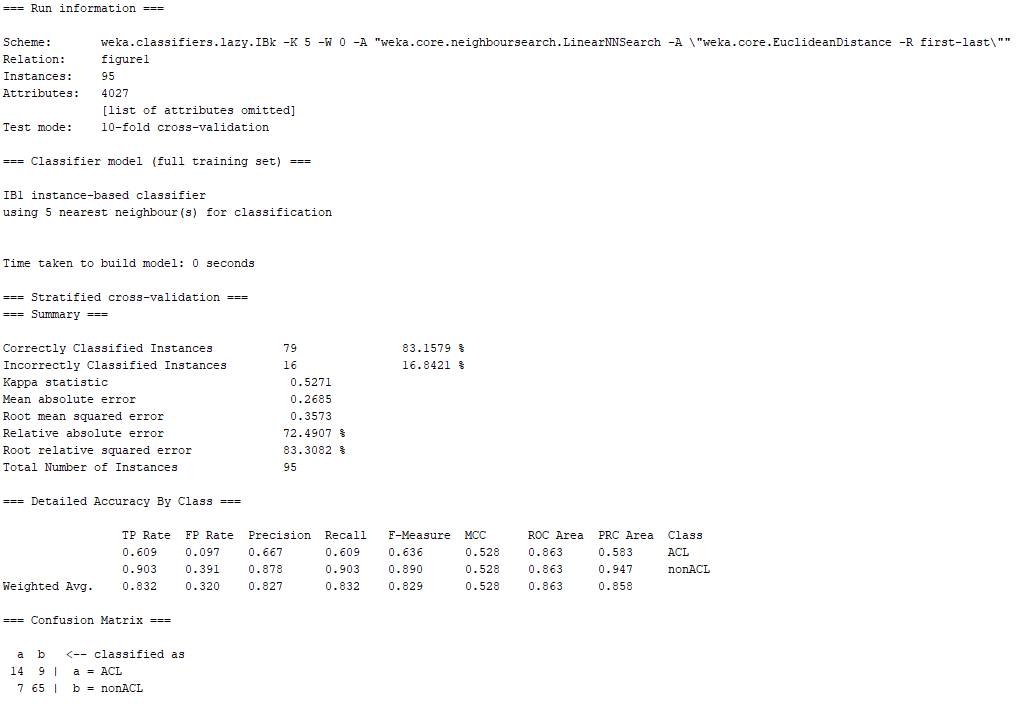
Additionally, based on the performance results from the model it can be noted that AUC is 0.844. Table 2 below displays a summary of the performance metric for the Naïve Bayes classifier:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.379 | 0.076 | 0.844 |

Table 2: Naïve Bayes Classifier with Full Dataset Performance Metrics

From table 2 above it can be noted that this classifier performed better at classifying the primary (negative) class when compared to the secondary class since it only had a false negative rate of 0.076, while the secondary (positive) class had a higher misclassification with 0.379. It is important to note that the overall performance of the classifier under the ROC value was satisfactory at 0.844 although this model greatly favors the proper classification of the minority class in the dataset.

The second classifier to be trained was the k nearest neighbor with a K value of 5. The results from the training of the model can be seen below:



Picture 2: 5-Nearest Neighbor Classifier with Full dataset

Based on the information from picture 2 which displays the 5-Nearesrt Neighbor classifier training information, the confusion matrix from the model can be observed in table 3 below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | ACL | nonACL |
| ACL | 14  (TN) | 9  Type II (FN) |
| nonACL | 7  Type I (FP) | 65  (TP) |

Table 3: Confusion Matrix 5-Nearest Neighbor Classifier with Full Dataset

With a clear understanding of the confusion matrix for the model, it is not possible to calculate the false positive rate and false negative rate. These calculations can be observed below:

Additionally, based on the performance results from the model it can be noted that AUC is 0.863. Table 4 below displays a summary of the performance metric for the 5-Nearest Neighbor classifier:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| 5-Nearesrt Neighbor | 0.333 | 0.122 | 0.863 |

Table 4: 5-Nearest Neighbor Classifier with Full Dataset Performance Metrics

From table 4 above it can be noted that this classifier performed better at classifying the primary (negative) class when compared to the secondary class since it only had a false negative rate of 0.122, while the secondary (positive) class had a higher misclassification with 0.333. It is important to note that the overall performance of the classifier under the ROC value was satisfactory at 0.863 although this model greatly favors the proper classification of the majority class in the dataset.

With this information in mind from both models, it is now possible to conduct an evaluation of performance between the two models. Table 5 below shows the performance between the two classifiers:

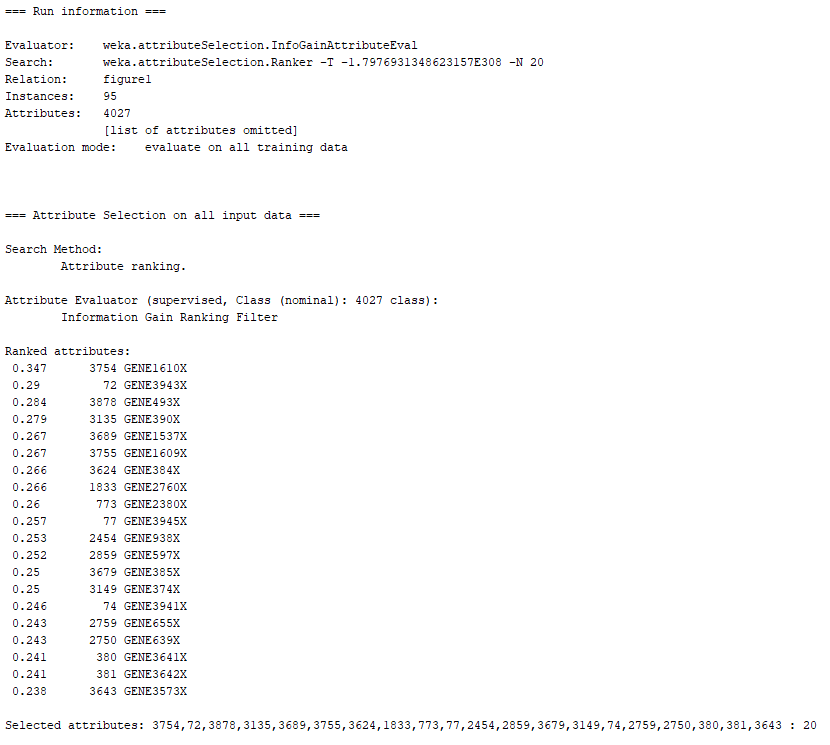
|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.379 | 0.076 | 0.844 |
| 5-Nearesrt Neighbor | 0.333 | 0.122 | 0.863 |

Table 5: Classifiers with Full Dataset Performance Metrics

From table 5 above it can easily determined that the 5-Nearesrt Neighbor had a better classification of both classes when compares to the Naïve Bayes classifiers. This cannot only be noted on the higher value of the ROC performance metrics which is slightly higher, but it can also be noted in the slightly lower false positive rate and higher false negative rate. These values can help us determine that although both classifiers were favoring the classification of the majority class in the dataset, the 5-Nearesrt Neighbor model had a more robust classification of both classes overall.

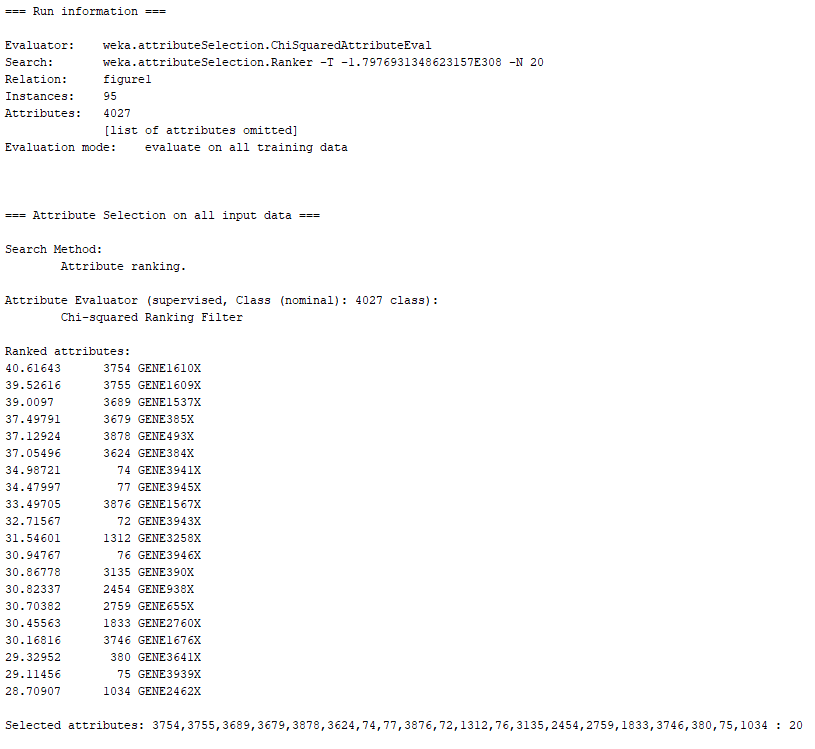
Part 2: Preliminary Feature Selection

In this section of the assignment, we train the same models that were used in part one with the same options. The only difference in this section of the assignment is that both models will be trained using two different datasets which will be derived from the main dataset. These new datasets will be a subset each of the main dataset and to create them we will implement two distinct feature ranking techniques. The feature selection techniques used in to generate the two subsets to train the models include the information gain feature selection technique and the Chi Squared feature selection technique. It is important to note that only the top 20 features from each feature selection technique will be used to build the new datasets used to trained the models. The first sub dataset created used the Information Gain feature ranking technique. The results from the information gain ranking can be seen below:



Picture 3: Information Gain Feature Ranking Information

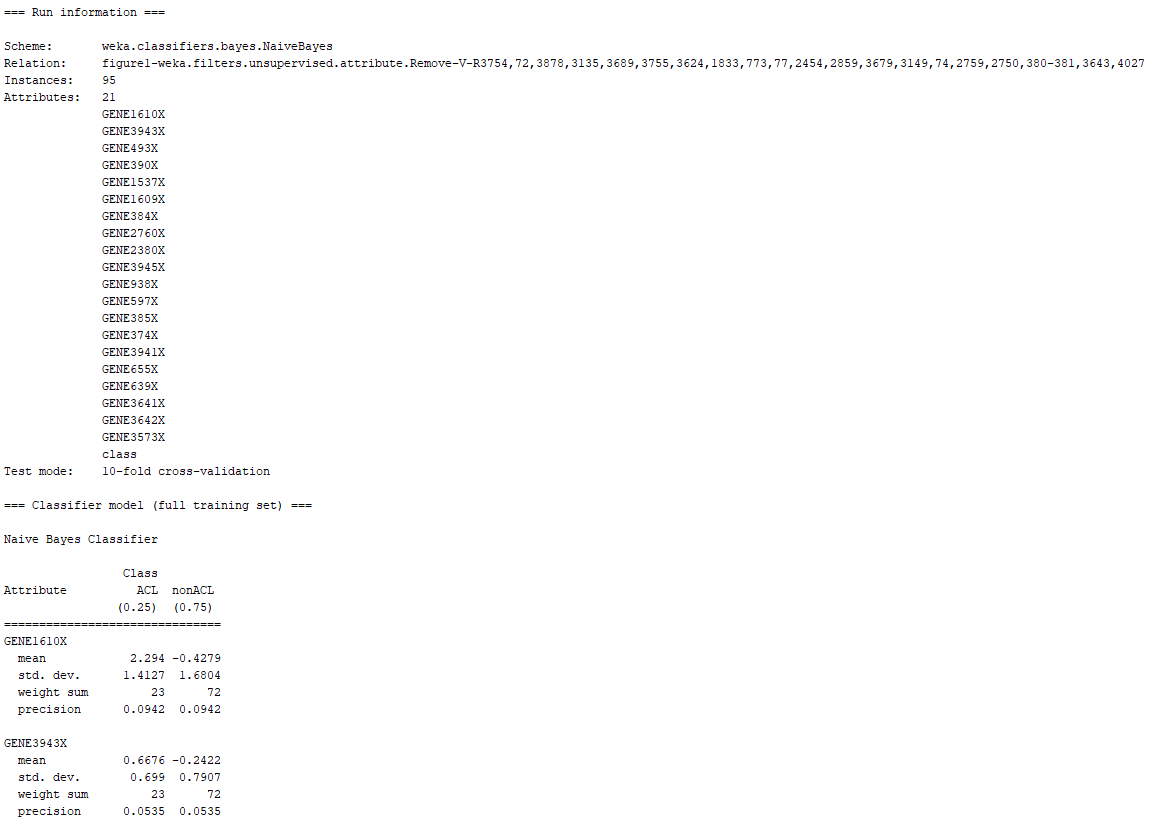
The second sub dataset created used the Chi Squared feature ranking technique. The results from the chi squared ranking can be seen below:

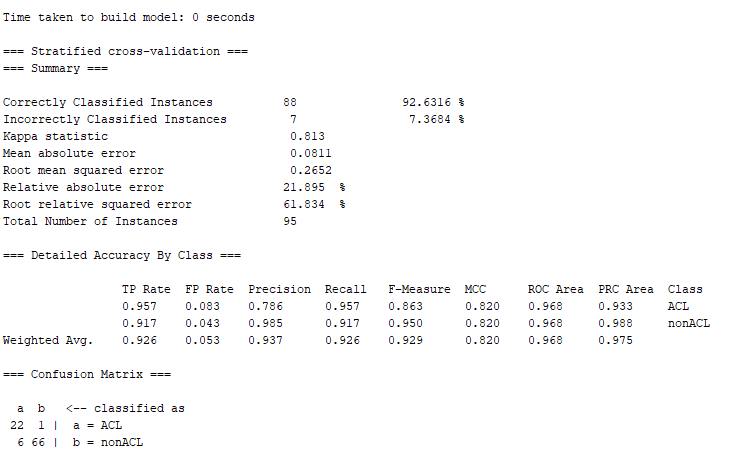


Picture 4: Chi Squared Feature Ranking Information

It is essential to note that the only distinction between the original dataset and the subsets is the reduction of the feature dimensionality to 20. Furthermore, the distinction between the two subsets are the different techniques implemented in order to create them. On one side we have the subset created with the information gain feature ranking technique. While on the other side, the feature ranking technique implemented was chi squared. With this information in mind, the step is to train the models from part 1 using the new datasets and comparing and contrasting their performance.

The first model built was the Naïve Bayes classifier trained with the information gain feature ranking technique dataset. The results from the training can be seen below:





Picture 5: Naïve bayes Classifier with Information Gain dataset

Based on the information from picture 5 which displays the Naïve Bayes classifier training information with the information gain dataset, the confusion matrix from the model can be observed in table 6 below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | ACL | nonACL |
| ACL | 22  (TN) | 1  Type II (FN) |
| nonACL | 6  Type I (FP) | 66  (TP) |

Table 6: Confusion Matrix Naïve Bayes Classifier with Information Gain Dataset

With a clear understanding of the confusion matrix for the model, it is not possible to calculate the false positive rate and false negative rate. These calculations can be observed below:

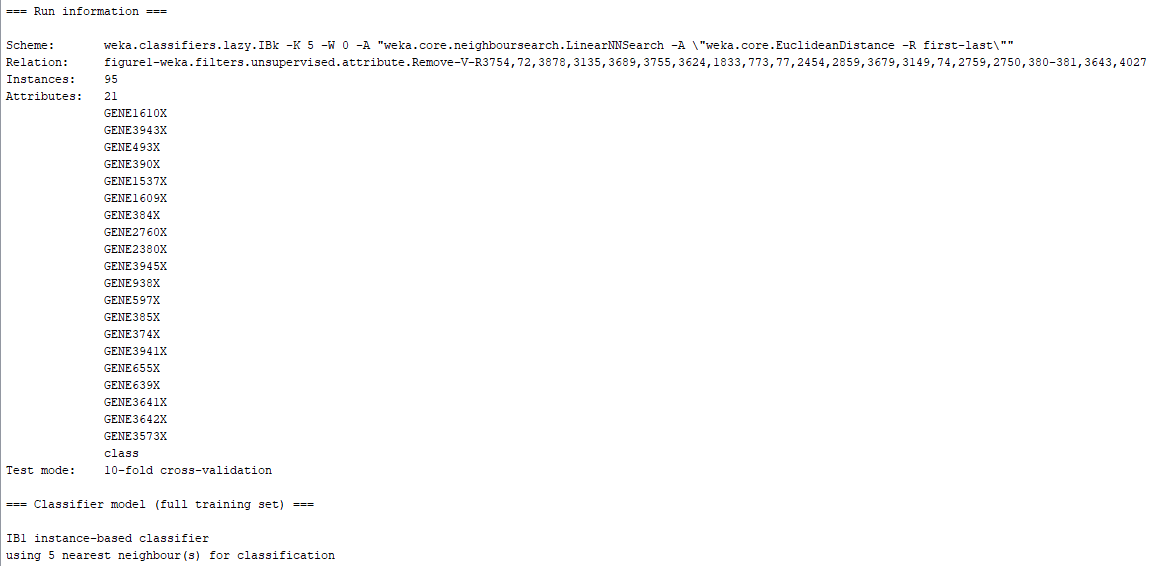
Additionally, based on the performance results from the model it can be noted that AUC is 0.968. Table 7 below displays a summary of the performance metric for the Naïve Bayes classifier:

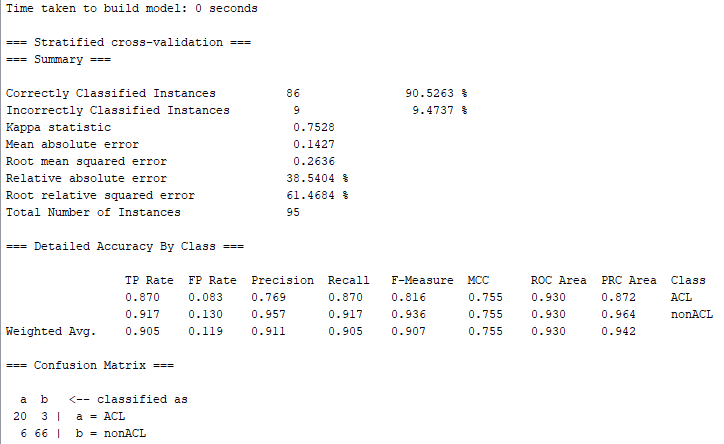
|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.214 | 0.015 | 0.968 |

Table 7: Naïve Bayes Classifier with Information Gain Dataset Performance Metrics

From table 7 above it can be noted that this classifier performed better at classifying the primary (negative) class when compared to the secondary class since it only had a false negative rate of 0.015, while the secondary (positive) class had a higher misclassification with 0.214. It is important to note that the overall performance of the classifier under the ROC value was outstanding at 0.968 although this model slightly favors the proper classification of the minority class in the dataset. Nevertheless, it can be concluded that this model is very balanced in the overall classification of the information gain dataset.

The next classifier to be trained was the k nearest neighbor with a K value of 5 and the information gain feature ranked dataset. The results from the training of the model can be seen below:





Picture 6: 5-Nearest Neighbor Classifier with Information Gain Dataset

Based on the information from picture 6 which displays the 5-Nearesrt Neighbor classifier training information, the confusion matrix from the model can be observed in table 8 below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | ACL | nonACL |
| ACL | 20  (TN) | 3  Type II (FN) |
| nonACL | 6  Type I (FP) | 66  (TP) |

Table 8: Confusion Matrix 5-Nearest Neighbor Classifier with Information Gain Dataset

With a clear understanding of the confusion matrix for the model, it is not possible to calculate the false positive rate and false negative rate. These calculations can be observed below:

Additionally, based on the performance results from the model it can be noted that AUC is 0.930. Table 9 below displays a summary of the performance metric for the 5-Nearest Neighbor classifier:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| 5-Nearesrt Neighbor | 0.231 | 0.045 | 0.930 |

Table 9: 5-Nearest Neighbor Classifier with Information gain Dataset Performance Metrics

From table 9 above it can be noted that this classifier performed better at classifying the primary (negative) class when compared to the secondary class since it only had a false negative rate of 0.045, while the secondary (positive) class had a higher misclassification with 0.231. It is important to note that the overall performance of the classifier under the ROC value was very good with a value of 0.930.

With this information in mind from both models, it is now possible to conduct an evaluation of performance between the two models. Table 10 below shows the performance between the two classifiers:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.214 | 0.015 | 0.968 |
| 5-Nearesrt Neighbor | 0.231 | 0.045 | 0.930 |

Table 10: Classifiers with Information Gain Dataset Performance Metrics

From table 10 above it can determined easily determine that the Naïve Bayes had a better classification of both classes when compares to the 5-Nearesrt Neighbor classifier. This cannot only be noted on the higher value of the ROC performance metrics which is slightly higher for the Naïve Bayes classifier, but it can also be noted in the slightly lower false positive rate and lower false negative rate. These values can help us determine that although both classifiers were very balanced in the classification of both classes in the information gain dataset. Nevertheless, the Naïve Bayes model had a more robust classification of both classes overall.

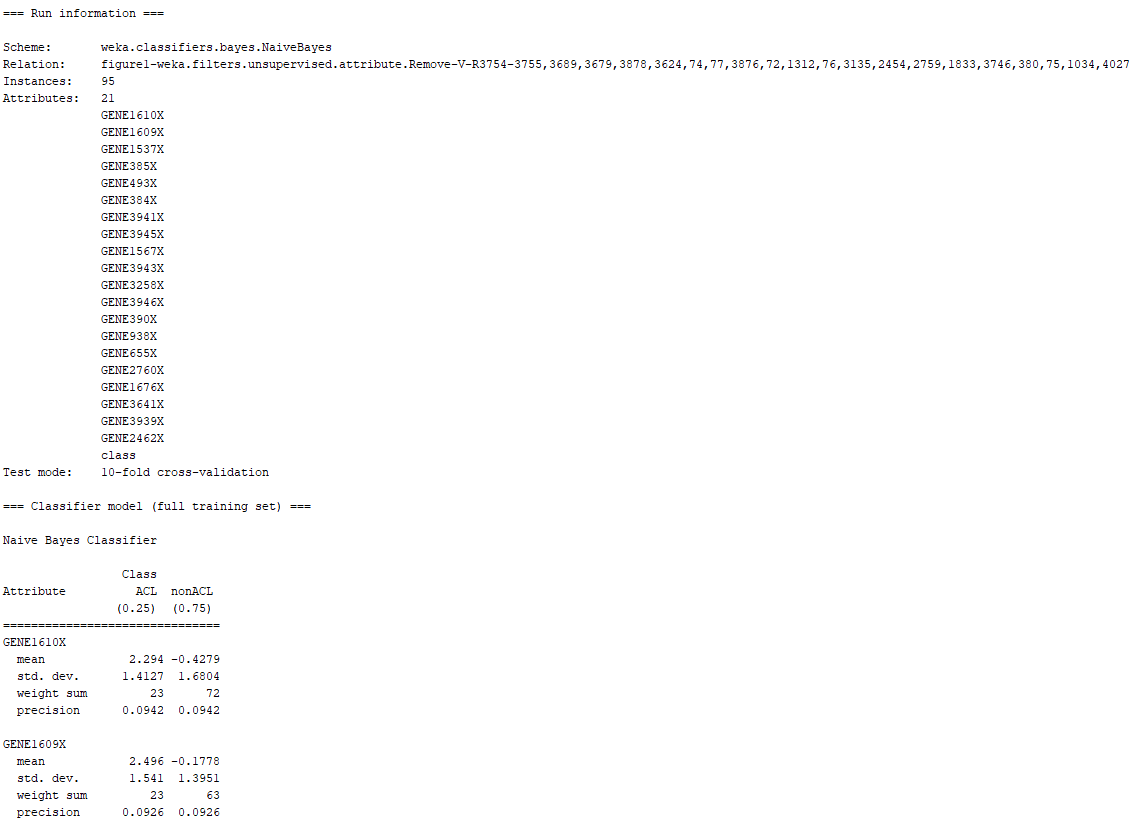
With this information available, it is now possible to compare the performance metrics between the models built with the full dataset and the models built with the information gain dataset. The performance metrics for all the models can be seen below:

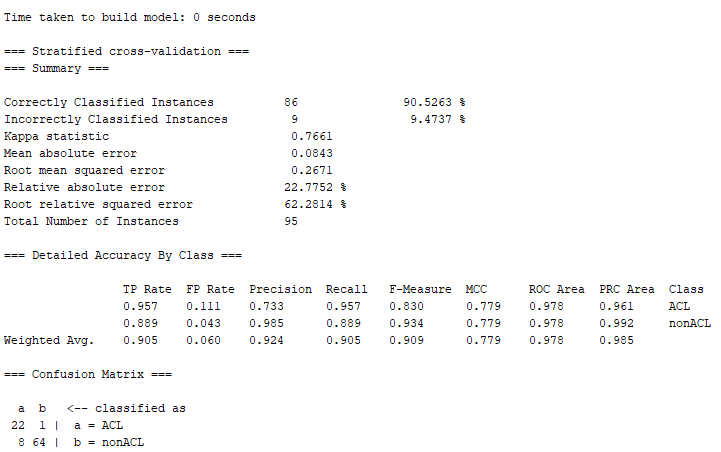
|  |  |  |  |
| --- | --- | --- | --- |
| Model  (Dataset) | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes  (Full Dataset) | 0.379 | 0.076 | 0.844 |
| Naïve Bayes  (Info Gain Dataset) | 0.214 | 0.015 | 0.968 |
| 5-Nearest Neighbor  (Full Dataset) | 0.333 | 0.122 | 0.863 |
| 5-Nearesrt Neighbor  (Info Gain Dataset) | 0.231 | 0.045 | 0.930 |

Table 11: Classifiers with Information Gain Dataset and Full Dataset Performance Metrics

From table 11 above, it can be noted that from the Naïve Bayes classifiers, the model which had the best performance was the one that was trained using the information gain dataset. This can be noted because the aforementioned classifier had the lowest FPR and FNR between the model as well as the highest ROC value. On the other hand, from the two 5-Nearest Neighbor models, it can be noted as well that the classifier that was built with the information gain dataset performed better than its counterpart as it obtained a lower FPR and FNR with the highest value for the ROC. It is important to note as well that between the Naïve Bayes and 5-Nearest Neighbor classifiers trained with the information gain dataset, the Naïve Bayes model slightly outperformed the 5-Nearest Neighbor model in all performance metrics by having a lower FPR and FNR with a higher value for the ROC. With this information in mind, it can be concluded that it did not matter which model was used, whenever the information gain dataset was used to train the classifiers, the models substantially decreased the misclassification of samples for the dataset when compared to the models trained with the full dataset and no feature ranking technique. Furthermore, the value of ROC for the models with the information gain dataset was significantly higher when compared to the models trained with the full dataset.

For the second section of part 2 of the assignment we will conduct the training of the Naïve Bayes and 5-Nearest Neighbor classifiers with the dataset constructed using the chi square feature ranking technique. The first model built in this section was the Naïve Bayes classifier trained with the chi squared feature ranking technique dataset. The results from the training can be seen below:





Picture 7: Naïve bayes Classifier with Chi Squared dataset

Based on the information from picture 7 which displays the Naïve Bayes classifier training information with the chi squared dataset, the confusion matrix from the model can be observed in table 12 below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | ACL | nonACL |
| ACL | 22  (TN) | 1  Type II (FN) |
| nonACL | 8  Type I (FP) | 64  (TP) |

Table 12: Confusion Matrix Naïve Bayes Classifier with Chi Square Dataset

With a clear understanding of the confusion matrix for the model, it is not possible to calculate the false positive rate and false negative rate. These calculations can be observed below:

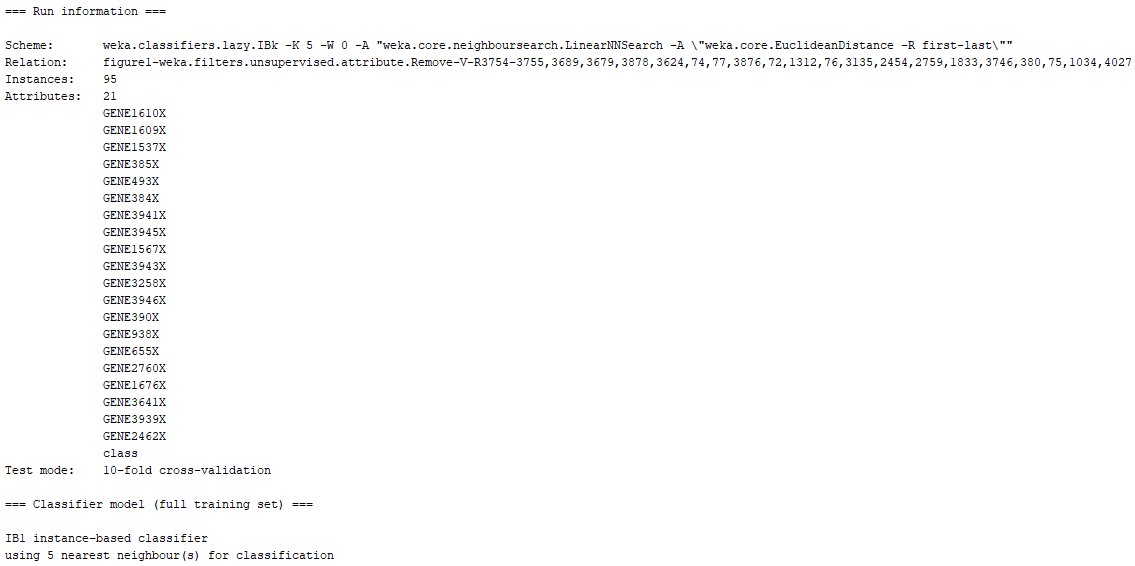
Additionally, based on the performance results from the model it can be noted that AUC is 0.978. Table 13 below displays a summary of the performance metric for the Naïve Bayes classifier:

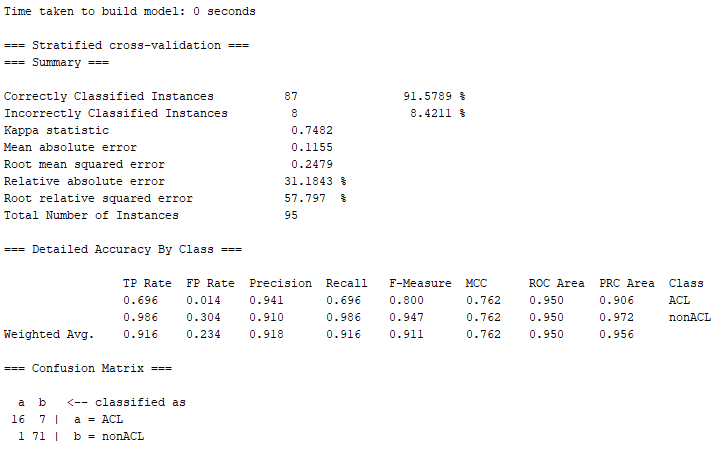
|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.267 | 0.015 | 0.978 |

Table 13: Naïve Bayes Classifier with Chi Squared Dataset Performance Metrics

From table 13 above it can be noted that this classifier performed better at classifying the primary (negative) class when compared to the secondary class since it only had a false negative rate of 0.015, while the secondary (positive) class had a higher misclassification with 0.267. It is important to note that the overall performance of the classifier under the ROC value was outstanding at 0.978 although this model slightly favors the proper classification of the minority class in the dataset. Nevertheless, it can be concluded that this model is very balanced in the overall classification of the chi squared dataset.

The next classifier to be trained was the k nearest neighbor with a K value of 5 and the chi squared feature ranked dataset. The results from the training of the model can be seen below:





Picture 8: 5-Nearest Neighbor Classifier with Chi Squared Dataset

Based on the information from picture 8 which displays the 5-Nearesrt Neighbor classifier training information, the confusion matrix from the model can be observed in table 14 below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | ACL | nonACL |
| ACL | 16  (TN) | 7  Type II (FN) |
| nonACL | 1  Type I (FP) | 71  (TP) |

Table 14: Confusion Matrix 5-Nearest Neighbor Classifier with Chi Squared Dataset

With a clear understanding of the confusion matrix for the model, it is not possible to calculate the false positive rate and false negative rate. These calculations can be observed below:

Additionally, based on the performance results from the model it can be noted that AUC is 0.950. Table 15 below displays a summary of the performance metric for the 5-Nearest Neighbor classifier:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| 5-Nearesrt Neighbor | 0.059 | 0.090 | 0.950 |

Table 15: 5-Nearest Neighbor Classifier with Chi Squared Dataset Performance Metrics

From table 15 above it can be noted that this classifier performed better at classifying the secondary (positive) class when compared to the primary class since it only had a false positive rate of 0.059, while the primary (negative) class had a higher misclassification with 0.09. It is important to note that the overall performance of the classifier under the ROC value was very good with a value of 0.950.

With this information in mind from both models, it is now possible to conduct an evaluation of performance between the two models. Table 16 below shows the performance between the two classifiers:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.267 | 0.015 | 0.978 |
| 5-Nearesrt Neighbor | 0.059 | 0.090 | 0.950 |

Table 16: Classifiers with Chi Squared Dataset Performance Metrics

From table 16 above, it can be easily determined that the Naïve Bayes had a better classification of both classes when compared to the 5-Nearesrt Neighbor classifier. This cannot only be noted on the higher value of the ROC performance metric which is slightly higher for the Naïve Bayes classifier, but it can also be noted in the slightly lower false negative rate. Although the FPR was higher in the Naïve bayes classifier, this rate represents the class which contains the majority samples in the dataset and it is the least important between the two classes. These values can help us determine that although both classifiers were very balanced in the classification of both classes in the chi squared dataset, the Naïve Bayes model had a more robust classification of the primary class which is the important class of the dataset.

With this information available, it is now possible to compare the performance metrics between the models built with the full dataset and the models built with the chi square dataset. The performance metrics for all the models can be seen below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model  (Dataset) | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes  (Full Dataset) | 0.379 | 0.076 | 0.844 |
| Naïve Bayes  (Chi Squared Dataset) | 0.267 | 0.015 | 0.978 |
| 5-Nearest Neighbor  (Full Dataset) | 0.333 | 0.122 | 0.863 |
| 5-Nearesrt Neighbor  (Chi Squared Dataset) | 0.059 | 0.090 | 0.950 |

Table 17: Classifiers with Chi Squared Dataset and Full Dataset Performance Metrics

From table 17 above, it can be noted that from the Naïve Bayes classifiers, the model which had the best performance was the one that was trained using the chi squared dataset. This can be noted because the aforementioned classifier had the lowest FPR and FNR between the model as well as the highest ROC value. On the other hand, from the two 5-Nearest Neighbor models, it can be noted as well that the classifier that was built with the chi squared dataset performed better than its counterpart as it obtained a lower FPR and FNR with the highest value for the ROC. It is important to note as well that between the Naïve Bayes and 5-Nearest Neighbor classifiers trained with the chi squared dataset, the Naïve Bayes model slightly outperformed the 5-Nearest Neighbor model in the majority of the performance metrics by having a lower FNR with a higher value for the ROC, although the Naïve Bayes had a considerably higher FPR. With this information in mind, it can be concluded that it did not matter which model was used, whenever the chi squared dataset was used to train the classifiers, the models substantially decreased the misclassification of samples for the dataset when compared to the models trained with the full dataset and no feature ranking technique. Furthermore, the value of ROC for the models with the chi squared dataset was significantly higher when compared to the models trained with the full dataset.

With this information in mind, it is now possible to compare the performance metrics from all the different models that were trained in the experiment to draw comprehensive performance conclusions based on the results. Table 18 below shows the performance metrics from all the classifiers in the assignment:

|  |  |  |  |
| --- | --- | --- | --- |
| Model  (Dataset) | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes  (Full Dataset) | 0.379 | 0.076 | 0.844 |
| Naïve Bayes  (Info Gain Dataset) | 0.214 | 0.015 | 0.968 |
| Naïve Bayes  (Chi Squared Dataset) | 0.267 | 0.015 | 0.978 |
| 5-Nearest Neighbor  (Full Dataset) | 0.333 | 0.122 | 0.863 |
| 5-Nearesrt Neighbor  (Info Gain Dataset) | 0.231 | 0.045 | 0.930 |
| 5-Nearesrt Neighbor  (Chi Squared Dataset) | 0.059 | 0.090 | 0.950 |

Table 18: Classifiers with All Datasets Performance Metrics

From table 18 above, it can be noted that from the Naïve Bayes classifiers, the model which had the best performance was the one that was trained using the chi squared dataset. This can be noted because the aforementioned classifier had the exact same FNR between the models as well as the highest ROC value. Even tough, the chi squared had a slightly higher FPR, since this represents the misclassification of the majority class in the dataset it did not have a negative impact on the overall classification performance of the model. On the other hand, from the 5-Nearest Neighbor models, it can be noted as well that the classifier that was built with the chi squared dataset performed better than its counterpart as it obtained a slightly lower FPR with the highest value for the ROC. It is important to make the distinction that this classifier also obtained a slightly higher misclassification of the primary class in the FNR. Therefore, the classification performance of the models slightly favors the majority class of the dataset. It is important to note as well that between the Naïve Bayes and 5-Nearest Neighbor classifiers trained with all datasets, the Naïve Bayes with the chi squared model outperformed all other models. The aforementioned classifier obtained the highest ROC value as well as the lowest FNR value which delivered a more robust dataset for the classification of both classes.

With this information in mind, it can be concluded that it did not matter which model was used, whenever a sub dataset derived from the main dataset using a feature ranking technique, both models substantially decreased the misclassification of samples for the dataset when compared to the models trained with the full dataset and no feature ranking technique. Furthermore, the value of ROC for the models with the feature selection technique dataset was significantly higher when compared to the models trained with the full dataset. Therefore, it can be concluded from this experiment that implementing some type of feature ranking technique to a dataset with a high degree of dimensionality, can be very beneficial when training machine learning classifiers in order to increase the classification performance of the model for a dataset with a lot of features as well as class imbalance.