Z Number: 23226341

Name: Carlos Gross-Martinez

Professor: Dr. Taghi M. Khoshgoftaar

Course: CAP6778 – Advanced Data Mining & Machine Learning

Assignment 4: Feature Selection II

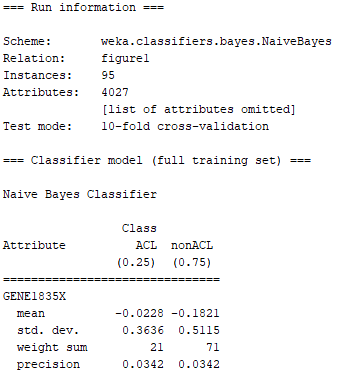
Dataset Analysis

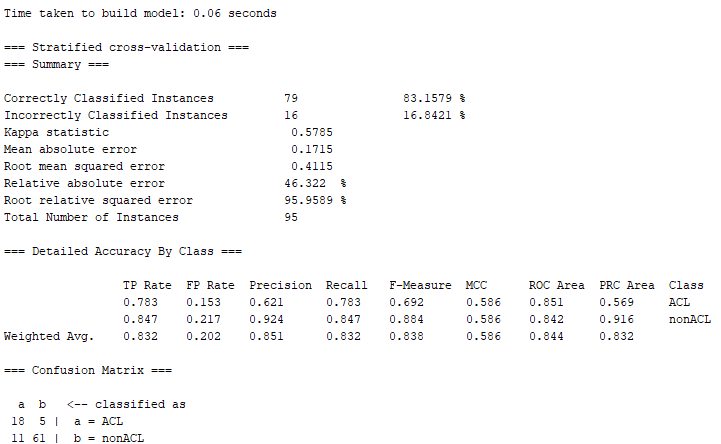
The dataset used in this assignment consisted of 95 instances/samples and 4027 attributes. From the 4027 attributes 4026 were input attributes to the model while 1 attribute was the class label. Furthermore, out of the 95 samples provided in the dataset, 23 were from the minority class labeled “ACL” 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: Full Feature Selection

In this segment of the assignment two classification models are trained using the full dataset described above. Afterwards, the same models will be trained with datasets where different feature selection techniques will be implemented to reduce the dataset dimensionality. The feature ranking techniques that will be implemented include: Gain Ratio, Symmetric Uncertainty, Information Gain, Chi Square, and two different types of Relief to include ReliefF and ReliefF-W. The newly created datasets will not only use different feature selection techniques, but it will also build the datasets with a predefined number of features. The number of features for each feature ranking technique include: 5, 6, 7, 8, 9, 10, 20, 50, 100, and 200. Hence, each model will be trained with 10 different subsets with different sizes in feature dimensionality for each feature ranking technique. The algorithms used to build the classifiers include the Naïve Bayes and 5-Nearest Neighbor. Once these two models are built for each dataset in each feature ranking technique, 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. In order to make it simpler, the first model to train and compare will be the Naïve Bayes with all different datasets and feature selection techniques to conduct the analysis, and then do the same for the 5-Nearest Neighbor. Finally, the results from both classifiers will be compared.

***Naïve Bayes Full Dataset Training Results:***





Picture 1: Naïve bayes Classifier with Full Dataset

Based on the information from the picture above which displays the Naïve Bayes classifier training information, we can quickly determine the FPR, FNR, and AUC information. This data can be observed in table 1 below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.153 | 0.217 | 0.844 |

Table 1: Naïve Bayes Classifier with Full Dataset Performance Metrics

From table 1 above it can be noted that this classifier performed better at classifying the secondary (negative) class when compared to the primary class since it had a higher false negative rate of 0.217, while the secondary (negative) class had a lower misclassification rate 0.153. 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 majority class in the dataset.

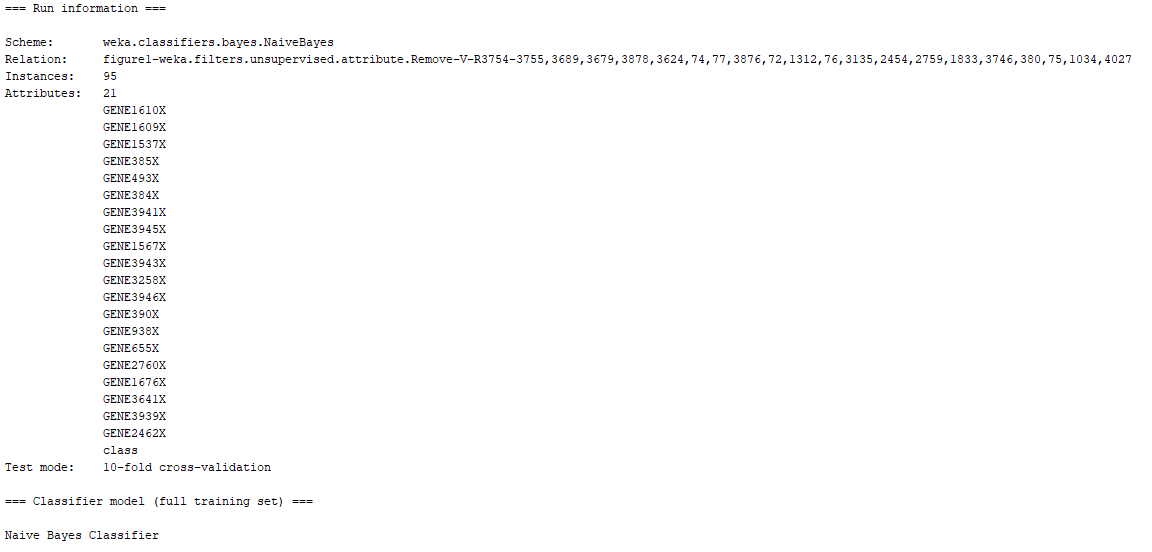
***Naïve Bayes Chi Square with all Datasets Results:***

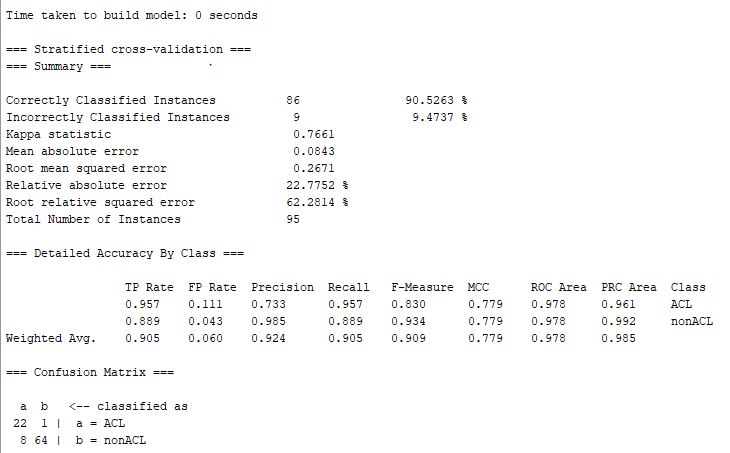
Next, the Naïve Bayes classifier will be trained using the dataset and feature sizes of the Chi Square feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the Chi Square technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the chi square datasets for the Naïve bayes classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| Naïve Bayes | 5 | 0.125 | 0.130 | 0.951 |
| Naïve Bayes | 6 | 0.139 | 0.130 | 0.953 |
| Naïve Bayes | 7 | 0.111 | 0.130 | 0.958 |
| Naïve Bayes | 8 | 0.097 | 0.087 | 0.958 |
| Naïve Bayes | 9 | 0.097 | 0.043 | 0.961 |
| Naïve Bayes | 10 | 0.125 | 0.043 | 0.965 |
| Naïve Bayes | 20 | 0.111 | 0.043 | 0.978 |
| Naïve Bayes | 50 | 0.083 | 0.043 | 0.970 |
| Naïve Bayes | 100 | 0.125 | 0.043 | 0.968 |
| Naïve Bayes | 200 | 0.125 | 0.043 | 0.959 |

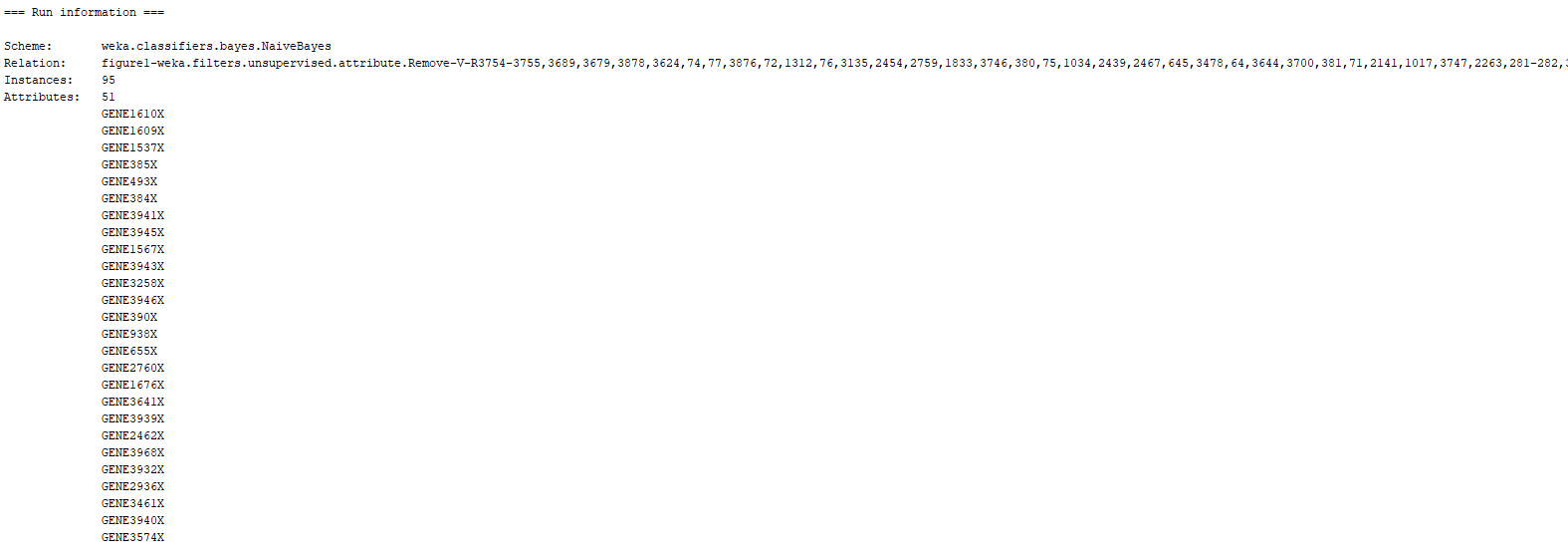
Table 2: Naïve Bayes Classifier with all Datasets for Chi Square Performance Metrics

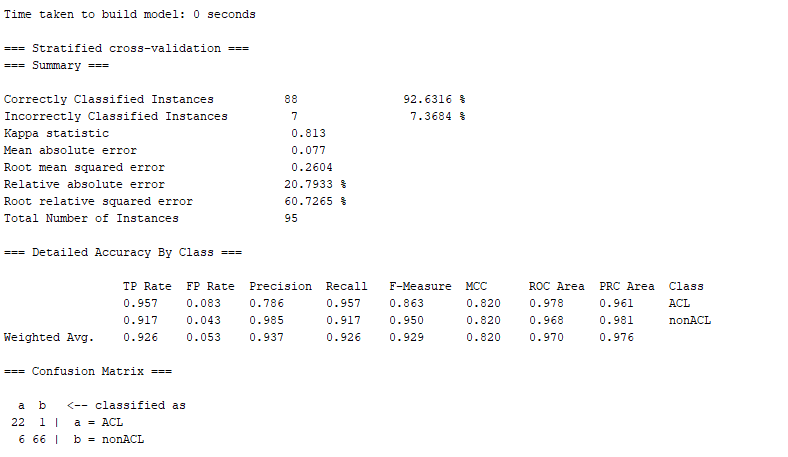
Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed very well with a strong AUC of 0.951. On the other side of the spectrum, when the model was trained with 200 the AUC did not improve greatly only delivering a value of 0.959. It is necessary to note the two patterns from the AUC and FNR values from the table above. From an FNR perspective, when the number of features in the dataset were small, the value for the FNR was the highest. Nevertheless, as the value of the features increased in the dataset, the FNR value decreased on every time the model was trained with a higher feature dimensionality until it finally stopped decreasing when it reached the value of 0.043. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC had very accurate but far from the best value it could get. As the dimensionality of the dataset increased, the AUC value also increased until reaching a value 0.978 when the model was trained with 20 features. After the number of features increased from 20, then the value of AUC decreased as more dimensionality was added decreasing the classification performance of the model. Moreover, it is important to note that FPR value did not really display any type of trend. Nevertheless, it is essential to note that from all the values that it possessed the middle value was the one that provided assisted in delivering the highest AUC. In other words, from all the values which it had: 0.083, 0.097, 0.111, 0.125, and 0.139, the value of 0.111 was the one that assisted in delivering the highest AUC. Furthermore, based on the results from the table above, it can be concluded that when the model was trained using a feature dimensionality of 20, the classifiers had the best AUC and FNR performance metrics when compared to all other models, Although the FPR was not the best. Finally, when we compare the 20-feature model with the base model, it can be easily noted that the 20 feature possesses better performance metrics results when compared with its counterpart. The pictures of the top 3 models for the Chi Square feature ranking technique which included a feature dimensionality of 20, 50, and 100 can be observed below:



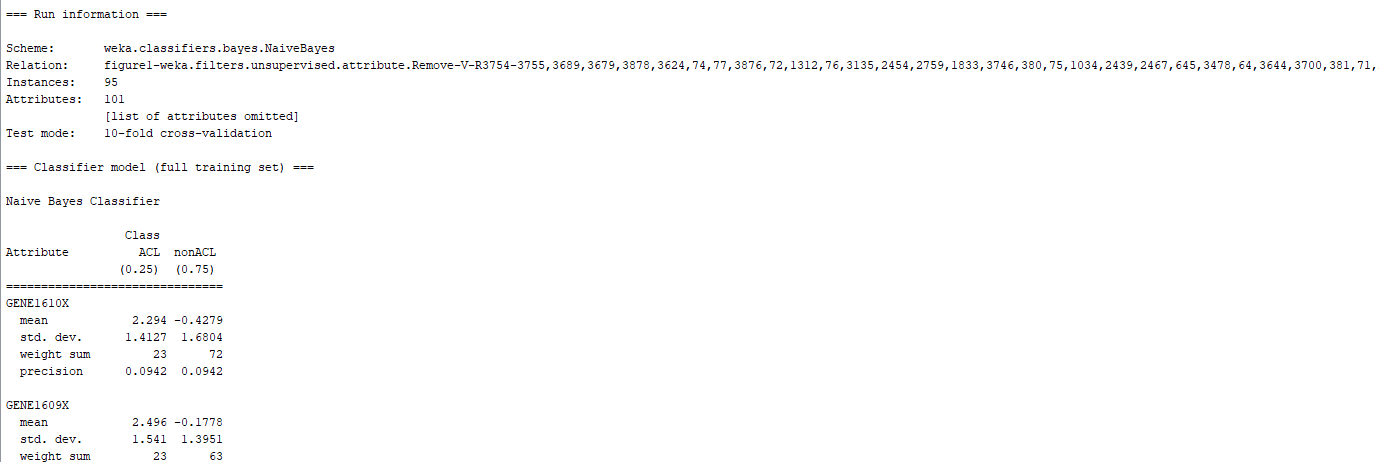


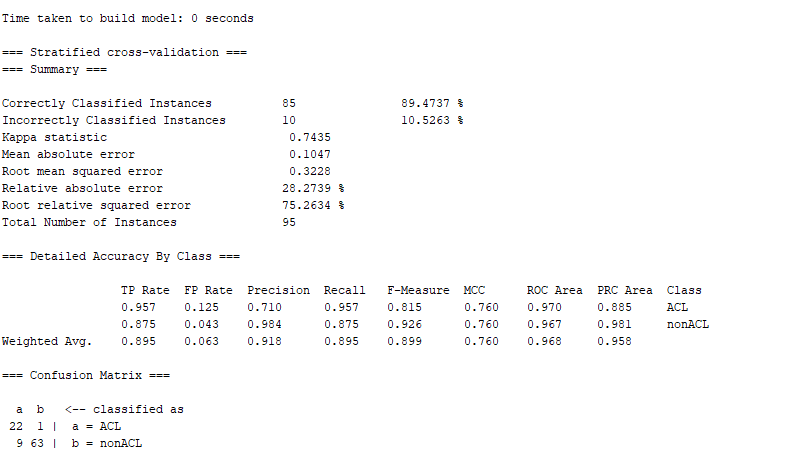
Picture 2: Naïve bayes Classifier with 20 Feature Dataset – Chi Square





Picture 3: Naïve bayes Classifier with 50 Feature Dataset – Chi Square





Picture 4: Naïve bayes Classifier with 100 Feature Dataset – Chi Square

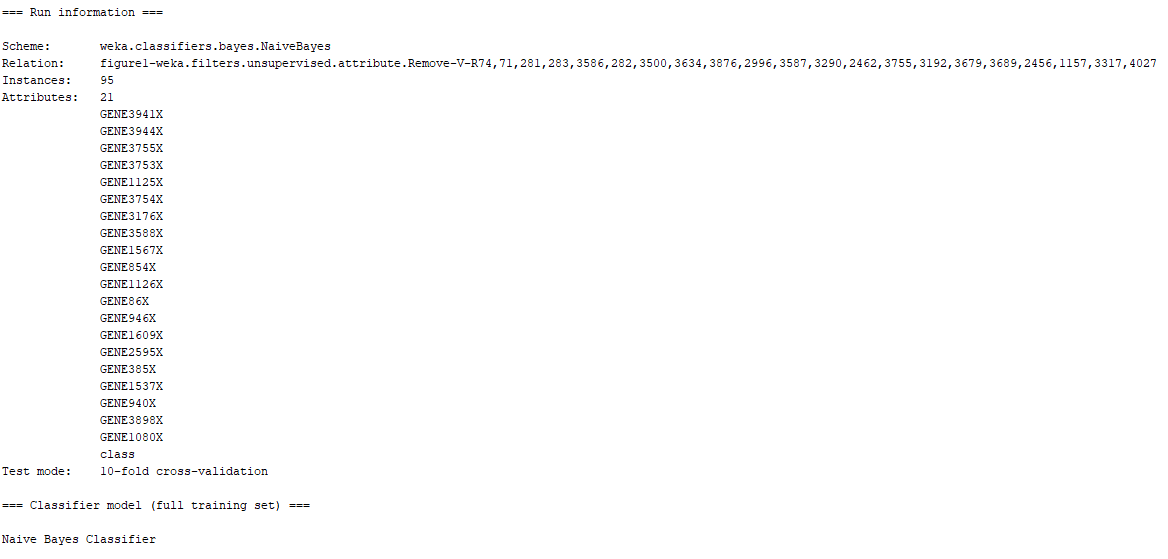
***Naïve Bayes Gain Ratio with all Datasets Results:***

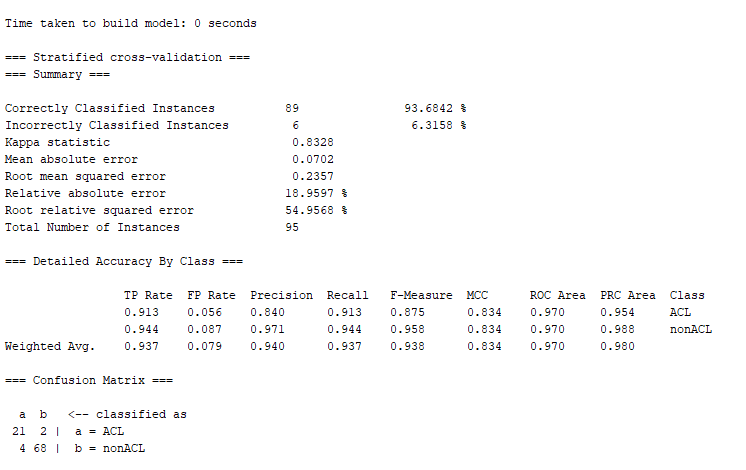
Next, the Naïve Bayes classifier will be trained using the dataset and feature sizes of the Gain Ratio feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the Gain Ratio technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the gain ratio datasets for the Naïve bayes classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| Naïve Bayes | 5 | 0.125 | 0.478 | 0.830 |
| Naïve Bayes | 6 | 0.167 | 0.435 | 0.817 |
| Naïve Bayes | 7 | 0.139 | 0.348 | 0.827 |
| Naïve Bayes | 8 | 0.139 | 0.435 | 0.830 |
| Naïve Bayes | 9 | 0.111 | 0.174 | 0.926 |
| Naïve Bayes | 10 | 0.097 | 0.174 | 0.927 |
| Naïve Bayes | 20 | 0.056 | 0.087 | 0.970 |
| Naïve Bayes | 50 | 0.056 | 0.043 | 0.969 |
| Naïve Bayes | 100 | 0.111 | 0.043 | 0.970 |
| Naïve Bayes | 200 | 0.125 | 0.043 | 0.967 |

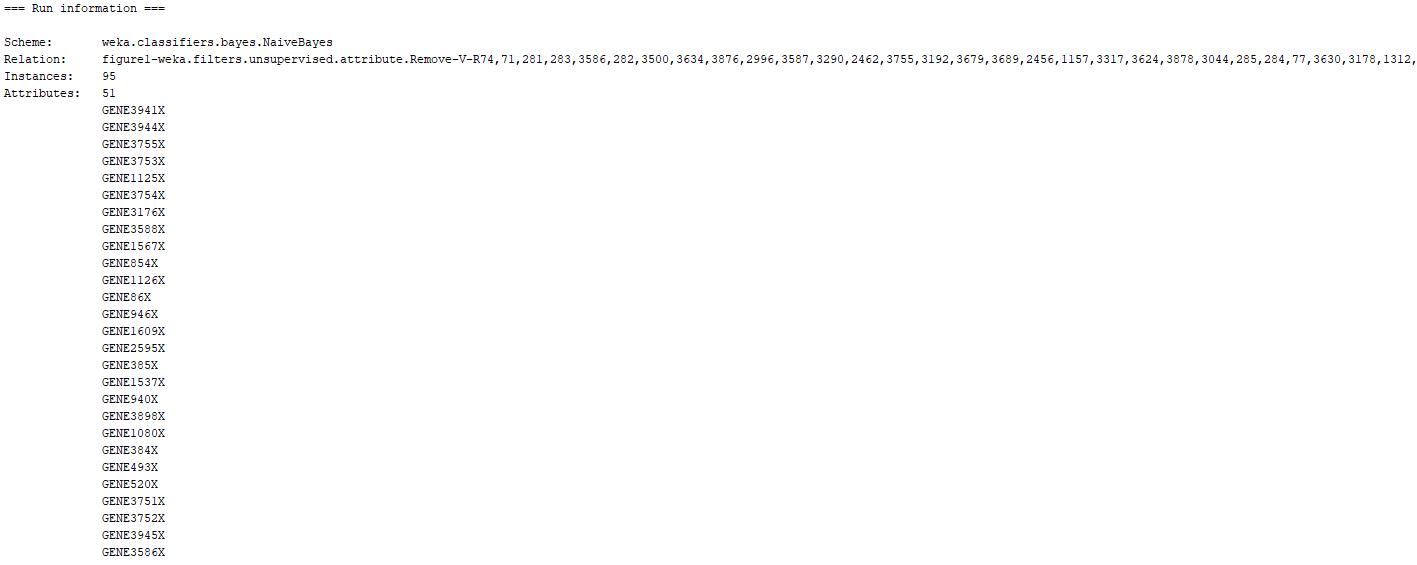
Table 3: Naïve Bayes Classifier with all Datasets for Gain ratio Performance Metrics

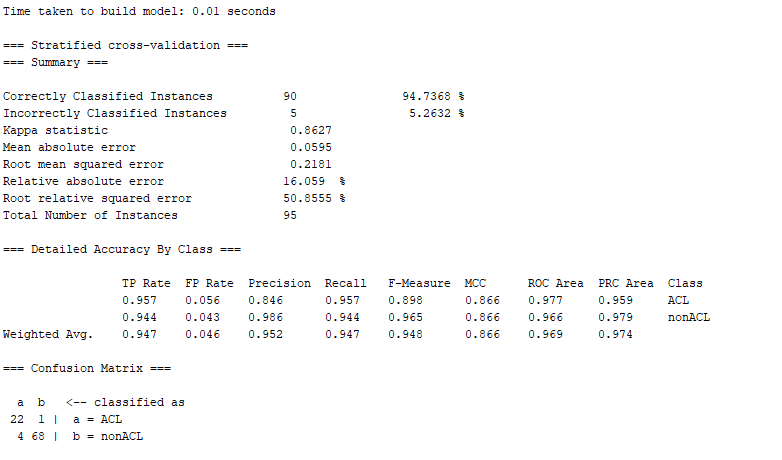
Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model did not perform very well with an AUC of 0.830. On the other side of the spectrum, when the model was trained with 200 the AUC did improve greatly delivering a strong value of 0.967. It is necessary to note the two patterns from the AUC and FNR values from the table above. From an FNR perspective, when the number of features in the dataset were small, the value for the FNR was the highest although when the 8-feature dataset was used, the FNR spiked to 0.435 even though the 7-feature dataset had an FNR of 0.348. Nevertheless, overall, as the value of the features increased in the dataset, the FNR value decreased on every time the model was trained with a higher feature dimensionality until it finally stopped decreasing when it reached the value of 0.043. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC was decently accurate but far from the best value it could get. As the dimensionality of the dataset increased, the AUC value also increased until reaching a value 0.970 when the model was trained with 20 features. Although the AUC was slightly reduced when the dataset with 50 features was used, the same peak AUC was reached again when the training set possessed a dimensionality of 100 features. Moreover, it is important to note that FPR value did not really display any type of trend. However, the best FPR that delivered the highest AUC were 0.056 and 0.111 for the 20 and 100 feature datasets respectively. It is necessary to make the distinction that although the 100-feature dataset obtained with the gain ratio feature selection technique had the same exact values for FPR and FNR as the model which the delivered the highest AUC with the chi square technique, the AUC results for the gain ratio model did not achieve the same level of AUC performance. Furthermore, it is also necessary to make mention of the fact that although the 20 and 100 feature datasets for the gain ratio had the same AUC, the 20-feature dataset provided a more balance FPR and FNR when compared to the 100-feature dataset. The pictures of the top 3 models for the gain ratio feature ranking technique which included a feature dimensionality of 20, 50, and 100 can be observed below:



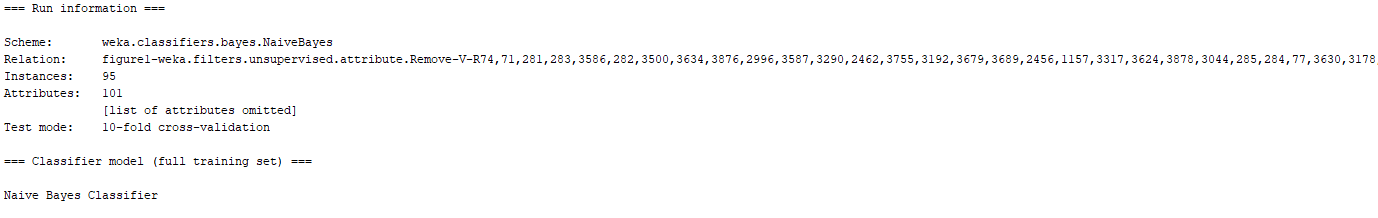


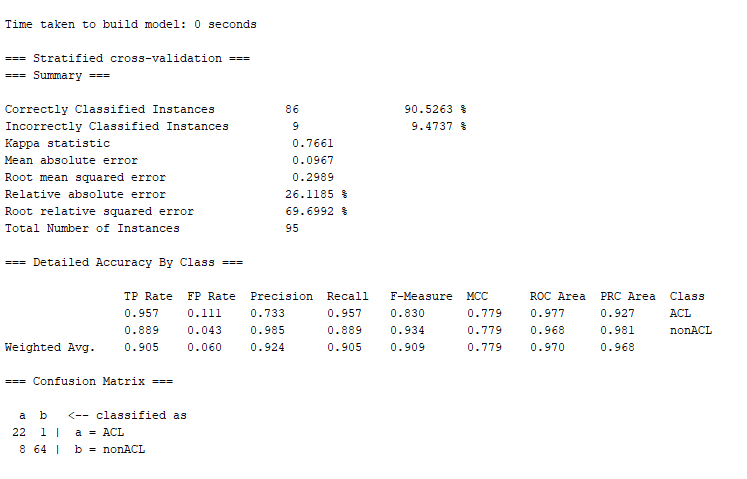
Picture 5: Naïve bayes Classifier with 20 Feature Dataset – Gain Ratio





Picture 6: Naïve bayes Classifier with 50 Feature Dataset – Gain Ratio





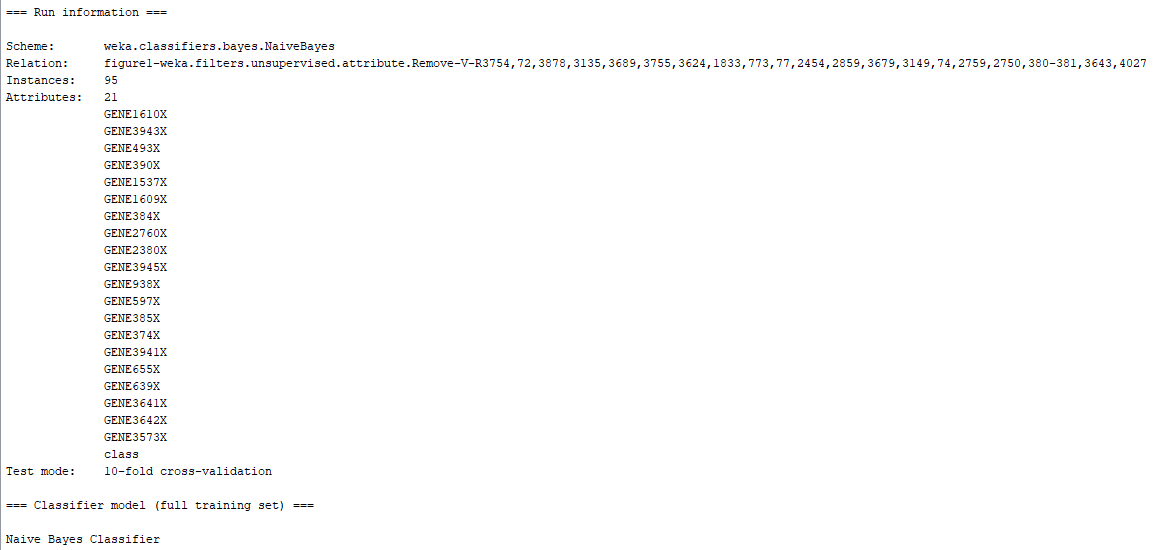
Picture 7: Naïve bayes Classifier with 100 Feature Dataset – Gain Ratio

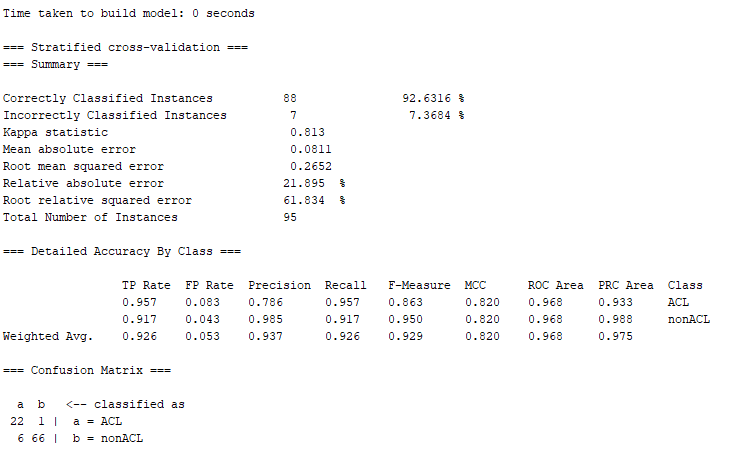
***Naïve Bayes Information gain with all Datasets Results:***

Next, the Naïve Bayes classifier will be trained using the dataset and feature sizes of the Information Gain feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the Information Gain technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the information gain datasets for the Naïve bayes classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| Naïve Bayes | 5 | 0.125 | 0.087 | 0.944 |
| Naïve Bayes | 6 | 0.139 | 0.174 | 0.947 |
| Naïve Bayes | 7 | 0.125 | 0.087 | 0.949 |
| Naïve Bayes | 8 | 0.111 | 0.087 | 0.960 |
| Naïve Bayes | 9 | 0.125 | 0.087 | 0.944 |
| Naïve Bayes | 10 | 0.111 | 0.087 | 0.944 |
| Naïve Bayes | 20 | 0.083 | 0.043 | 0.968 |
| Naïve Bayes | 50 | 0.139 | 0.043 | 0.968 |
| Naïve Bayes | 100 | 0.111 | 0.043 | 0.971 |
| Naïve Bayes | 200 | 0.153 | 0.043 | 0.955 |

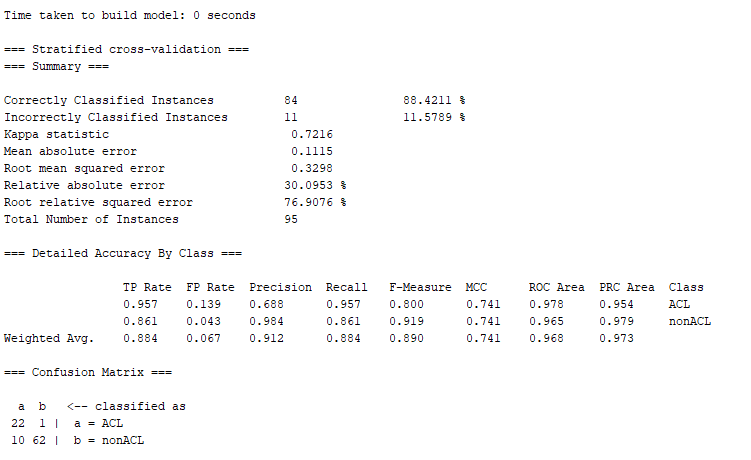
Table 4: Naïve Bayes Classifier with all Datasets for Information Gain Performance Metrics

Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed very well with an AUC of 0.944. On the other side of the spectrum, when the model was trained with 200 the AUC did not improve greatly delivering a value of 0.955. Through this experimentation, it can be noted that the FNR values remained fairly constant throughout all datasets with values 0.087 and 0.043. There was once instance on the 6-feature dataset which had a different value of 0.174. Nevertheless, between the two FNR values that repeated the most, the 0.043 value was the one that assisted in delivering the best AUC. It is also important to note that the FPR was random and did not have any type of pattern that can be determined. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC was very accurate but far from the best value it could get. As the dimensionality of the dataset increased, the AUC values fluctuated back and forth creating new values or repeating some of them until it reached its peak value at 0.971. There was not really a trend that could have been observed from the AUC as the numbers either repeated or provided different values to which a relationship could not be concluded. Moreover, it is important to note that FPR value did not really display any type of trend. However, the best FPR that delivered the highest AUC was 0.111 for the 100-feature dataset. It is necessary to make the distinction that although the 100-feature dataset obtained with the information gain feature selection technique had the same exact values for FPR and FNR as the model which the delivered the highest AUC with the chi square technique, the AUC results for the information gain model did not achieve the same level of AUC performance. The pictures of the top 3 models for the gain ratio feature ranking technique which included a feature dimensionality of 20, 50, and 100 can be observed below:

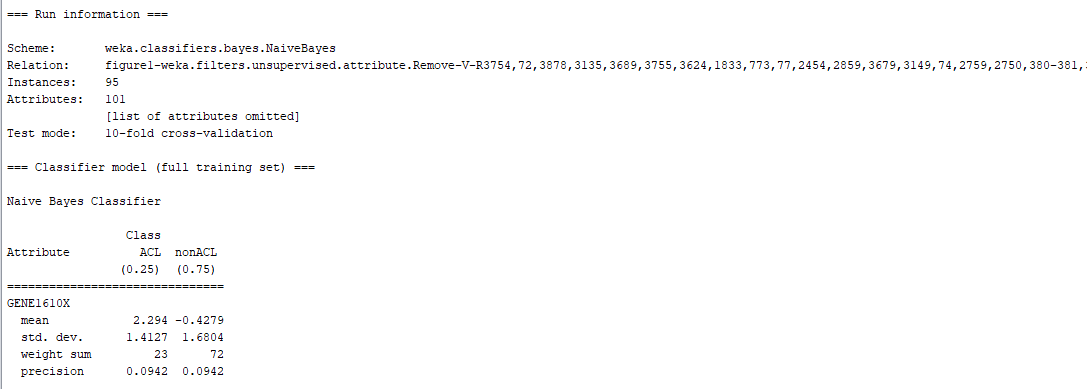


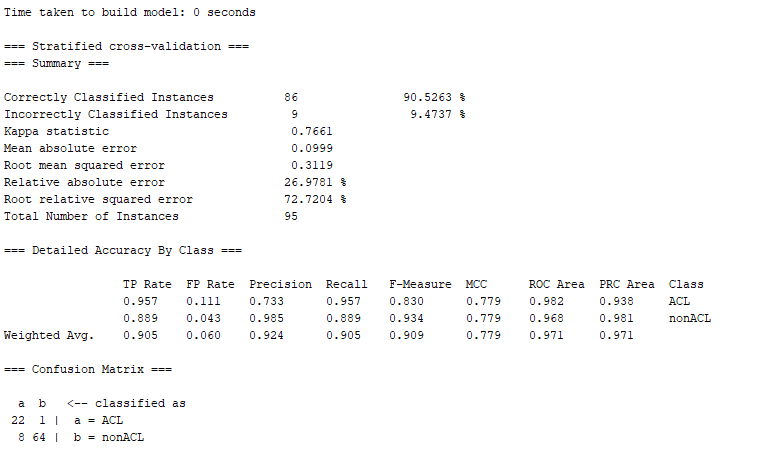
Picture 8: Naïve bayes Classifier with 20 Feature Dataset – Information Gain





Picture 9: Naïve bayes Classifier with 50 Feature Dataset – Information Gain





Picture 10: Naïve bayes Classifier with 100 Feature Dataset – Information Gain

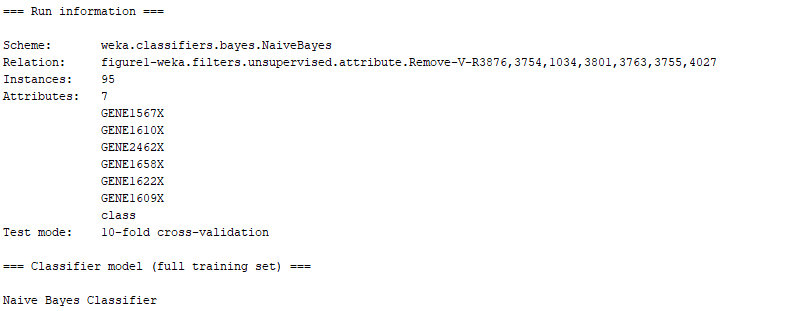
***Naïve Bayes ReliefF with all Datasets Results:***

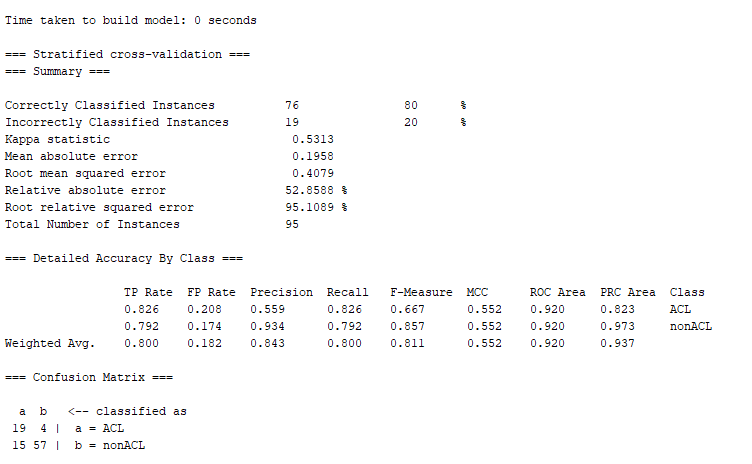
Next, the Naïve Bayes classifier will be trained using the dataset and feature sizes of the ReliefF feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the ReliefF technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the ReliefF datasets for the Naïve bayes classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| Naïve Bayes | 5 | 0.208 | 0.174 | 0.896 |
| Naïve Bayes | 6 | 0.208 | 0.174 | 0.920 |
| Naïve Bayes | 7 | 0.236 | 0.174 | 0.908 |
| Naïve Bayes | 8 | 0.236 | 0.174 | 0.908 |
| Naïve Bayes | 9 | 0.250 | 0.261 | 0.900 |
| Naïve Bayes | 10 | 0.236 | 0.217 | 0.894 |
| Naïve Bayes | 20 | 0.236 | 0.217 | 0.901 |
| Naïve Bayes | 50 | 0.194 | 0.130 | 0.910 |
| Naïve Bayes | 100 | 0.208 | 0.087 | 0.921 |
| Naïve Bayes | 200 | 0.222 | 0.087 | 0.931 |

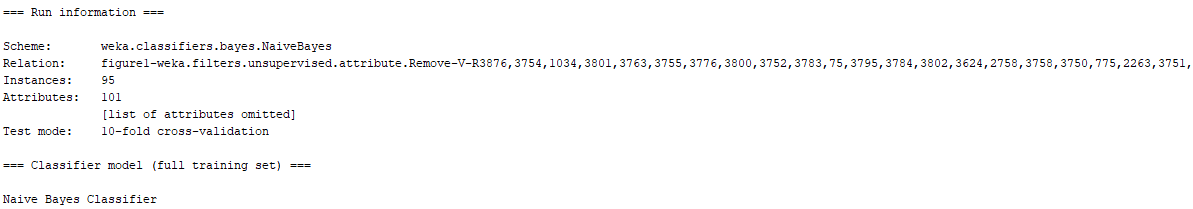
Table 5: Naïve Bayes Classifier with all Datasets for ReliefF Performance Metrics

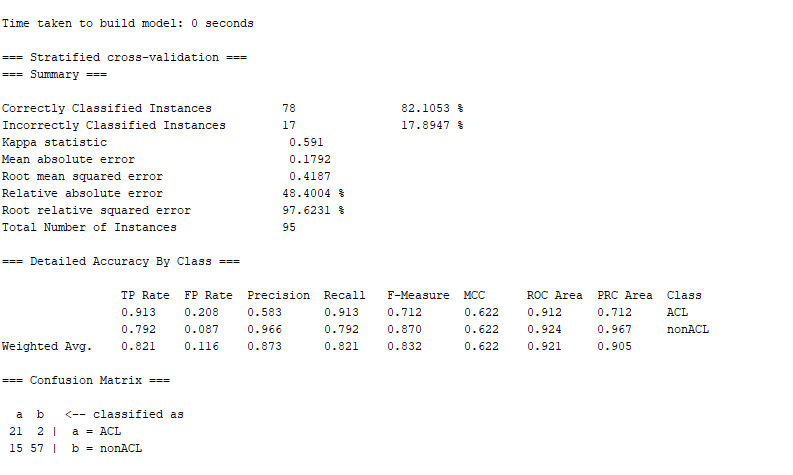
Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed decently with an AUC of 0.896. On the other side of the spectrum, when the model was trained with 200 the AUC did not improve greatly delivering a value of 0.931. Through this experimentation, it can be noted that the FNR values remained fairly constant throughout all datasets with values 0.087, 0.174, and 0.217. There were two instances on the 9 and 50 feature dataset which had a different value of 0.261 and 0.130 respectively. Nevertheless, between the two FNR values that repeated the most, the 0.087 value was the one that assisted in delivering the best AUC. It is also important to note that the FPR was very constant as well and did not vary a lot from all the values. For the FPR, the values that repeated the most include 0.208 and 0.236. Nevertheless, none of the two values were used when delivering the highest AUC value for this model. Another observation was in regards to the AUC value. This value fluctuated to different magnitudes depending on the size o the features. There was no concise pattern that could be derived from when the value of features was the smallest to when the value of the number of features was the highest. With this in mind, it can be concluded that there was not really a trend that could have been observed from the AUC as the numbers either repeated or provided different values to which a relationship could not be concluded. Finally, it is necessary to note that the model which contained a value of 200 for the number of features provided the highest AUC value of 0.931. Although, this was the highest accuracy, it is important to note that when the number of features was 20 and above, the model started to show an upwards trend for the AUC value as more features were added to the training dataset which leads to the assumption that with this technique, if the number of features continues to increase, maybe the AUC will also improve. Finally, it is essential to point that even with the highest AUC for this model 0.931, this model is not as precise in the classification as others which hovered around the 0.97 AUC value. The pictures of the top 3 models for the ReliefF feature ranking technique which included a feature dimensionality of 6, 100, and 200 can be observed below:



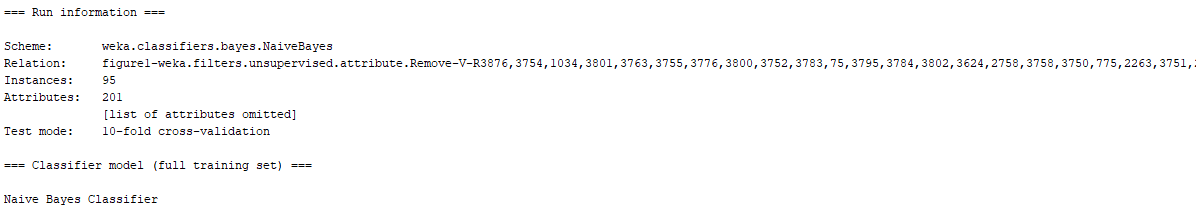


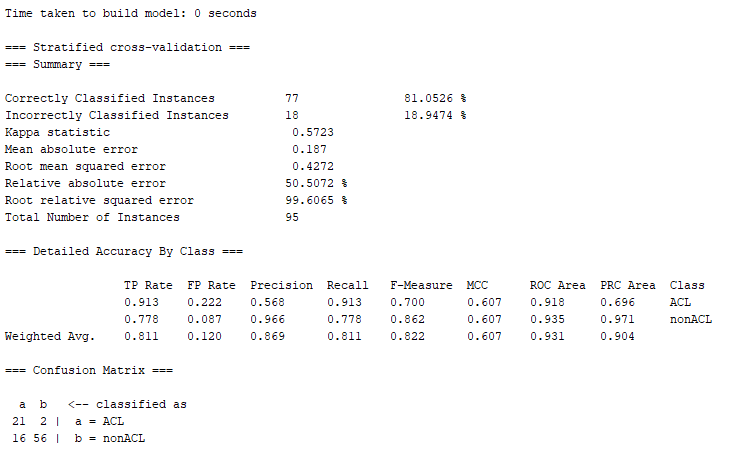
Picture 11: Naïve bayes Classifier with 6 Feature Dataset – ReliefF





Picture 12: Naïve bayes Classifier with 100 Feature Dataset – ReliefF





Picture 13: Naïve bayes Classifier with 200 Feature Dataset – ReliefF

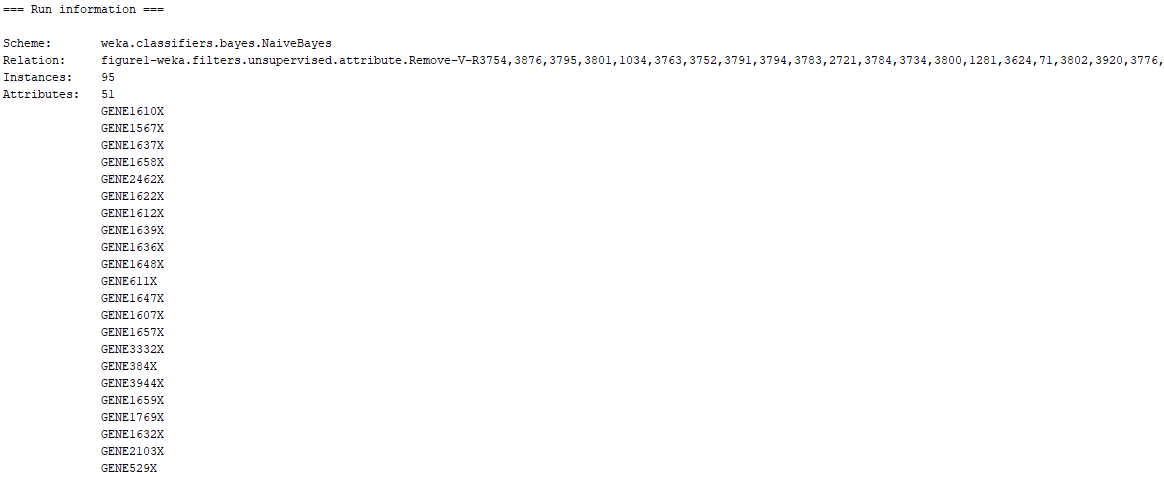
***Naïve Bayes ReliefW with all Datasets Results:***

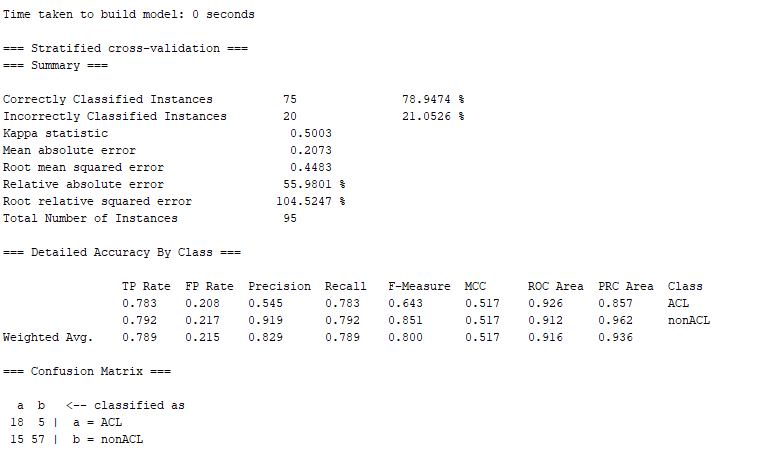
Next, the Naïve Bayes classifier will be trained using the dataset and feature sizes of the ReliefW feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the ReliefW technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the ReliefW datasets for the Naïve bayes classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| Naïve Bayes | 5 | 0.208 | 0.217 | 0.889 |
| Naïve Bayes | 6 | 0.222 | 0.217 | 0.885 |
| Naïve Bayes | 7 | 0.208 | 0.261 | 0.884 |
| Naïve Bayes | 8 | 0.222 | 0.261 | 0.876 |
| Naïve Bayes | 9 | 0.250 | 0.217 | 0.872 |
| Naïve Bayes | 10 | 0.264 | 0.217 | 0.870 |
| Naïve Bayes | 20 | 0.194 | 0.217 | 0.915 |
| Naïve Bayes | 50 | 0.208 | 0.217 | 0.916 |
| Naïve Bayes | 100 | 0.208 | 0.217 | 0.920 |
| Naïve Bayes | 200 | 0.125 | 0.130 | 0.940 |

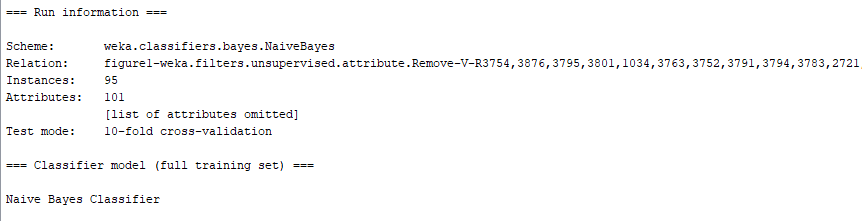
Table 5: Naïve Bayes Classifier with all Datasets for ReliefW Performance Metrics

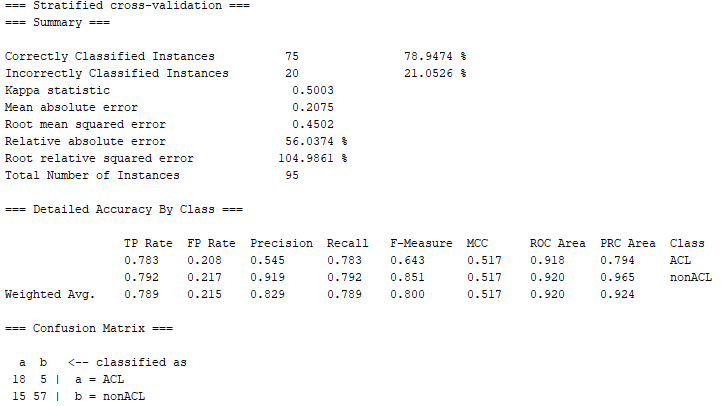
Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed decently with an AUC of 0.889. On the other side of the spectrum, when the model was trained with 200 the AUC did improve greatly delivering a value of 0.940. Through this experimentation, it can be noted that the FNR values remained fairly constant throughout all datasets with values 0.261 and 0.217. There weas one instances on the 200-feature dataset which had a different value 0.130 which concluded in the highest value for the AUC. Nevertheless, between the two FNR values that repeated the most, none of them assisted in delivering the best AUC. It is also important to note that the FPR was not constant as well and did vary a lot from all the values. For the FPR, the values that repeated the most include 0.208 and 0.222. Nevertheless, none of the two values were used when delivering the highest AUC value for this model. Another observation was in regards to the AUC value. This value fluctuated to different magnitudes depending on the size of the features. When the number of features was the smallest, the AUC was at 0.889. Then as the number of features increase, the AUC value started to decreased until it reached the value of 20 features in the training dataset. From this point onward, as the dimensionality of the dataset continued to increased, the AUC value increased as well. It is necessary to note that the model which contained a value of 200 for the number of features provided the highest AUC value of 0.940. Although, this was the highest accuracy, it is important to note that when the number of features was 20 and above, the model started to show an upwards trend for the AUC value as more features were added to the training dataset which leads to the assumption that with this technique, if the number of features continues to increase, maybe the AUC will also improve. Finally, it is essential to point that even with the highest AUC for this model 0.940, this model is not as precise in the classification as others which hovered around the 0.97 AUC value. The pictures of the top 3 models for the ReliefW feature ranking technique which included a feature dimensionality of 50, 100, and 200 can be observed below:



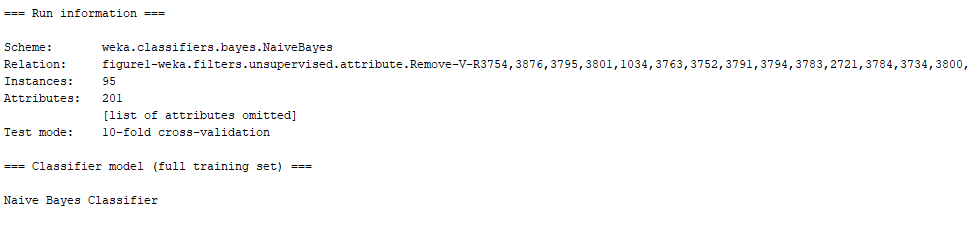


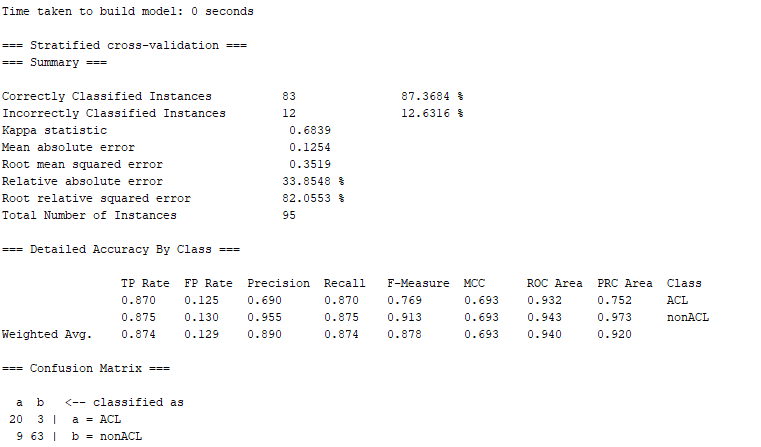
Picture 14: Naïve bayes Classifier with 50 Feature Dataset – ReliefW





Picture 15: Naïve bayes Classifier with 100 Feature Dataset – ReliefW





Picture 16: Naïve bayes Classifier with 200 Feature Dataset – ReliefW

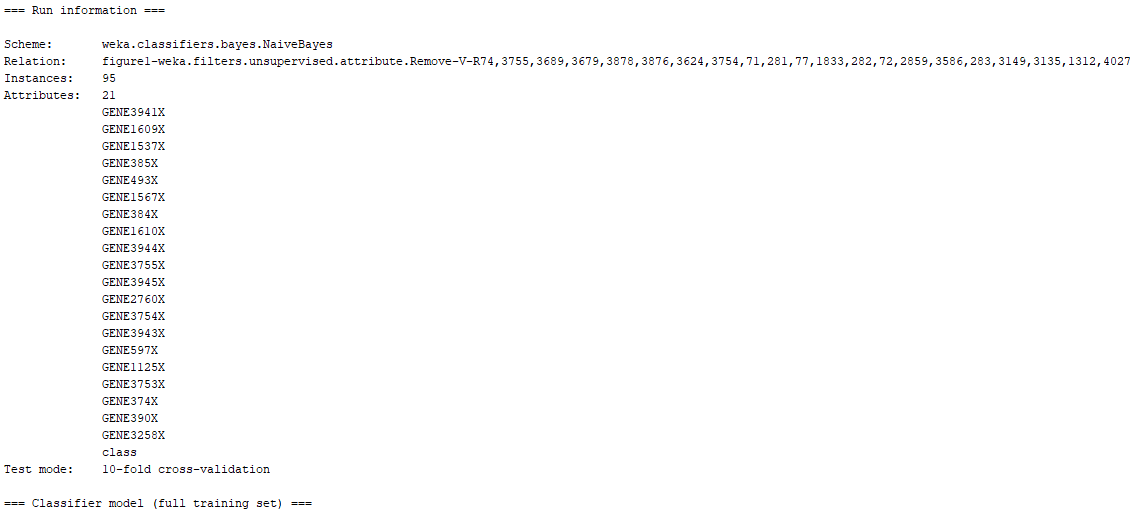
***Naïve Bayes Symmetric Uncertainty with all Datasets Results:***

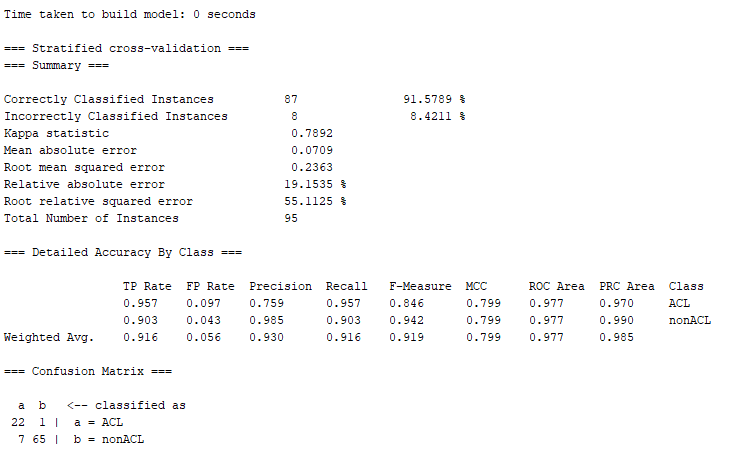
Next, the Naïve Bayes classifier will be trained using the dataset and feature sizes of the Symmetric Uncertainty feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the Symmetric Uncertainty technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the Symmetric Uncertainty datasets for the Naïve bayes classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| Naïve Bayes | 5 | 0.069 | 0.130 | 0.961 |
| Naïve Bayes | 6 | 0.097 | 0.087 | 0.969 |
| Naïve Bayes | 7 | 0.125 | 0.087 | 0.963 |
| Naïve Bayes | 8 | 0.125 | 0.043 | 0.966 |
| Naïve Bayes | 9 | 0.125 | 0.043 | 0.970 |
| Naïve Bayes | 10 | 0.111 | 0.043 | 0.969 |
| Naïve Bayes | 20 | 0.097 | 0.043 | 0.977 |
| Naïve Bayes | 50 | 0.111 | 0.043 | 0.971 |
| Naïve Bayes | 100 | 0.111 | 0.043 | 0.972 |
| Naïve Bayes | 200 | 0.139 | 0.043 | 0.964 |

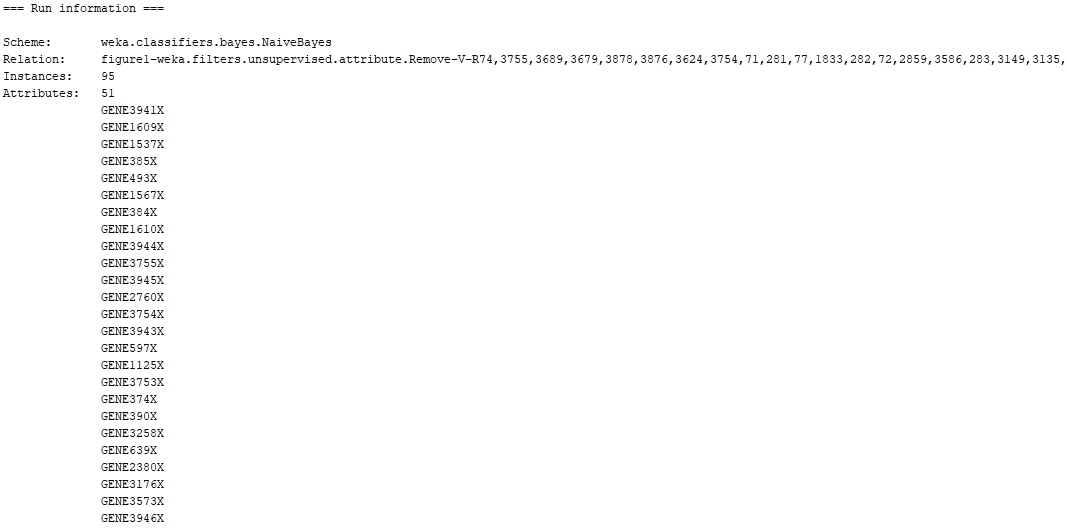
Table 6: Naïve Bayes Classifier with all Datasets for Symmetric Uncertainty Performance Metrics

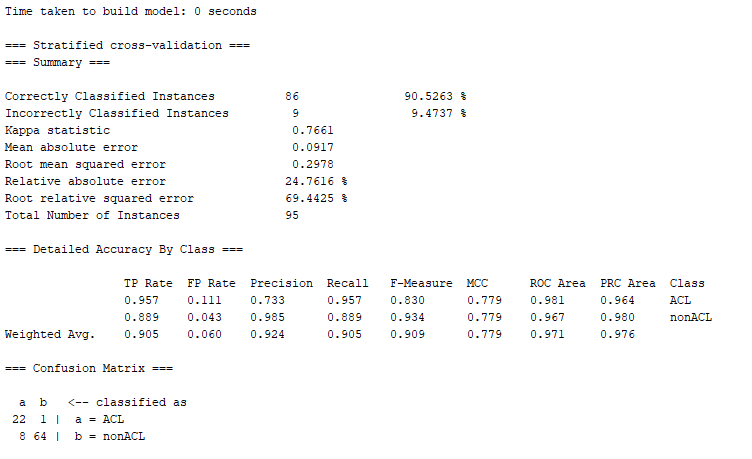
Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed very well with an AUC of 0.961. On the other side of the spectrum, when the model was trained with 200 the AUC did not improve greatly delivering a value of 0.964. Through this experimentation, it can be noted that the FNR values remained fairly constant throughout all datasets with values 0.087 and 0.043. There weas one instances on the 5-feature dataset which had a different value of 0.130. Nevertheless, it was the FNR value of 0.043 not only the one that repeated the most but also the one which led to the highest AUC. It is important to note as well the FNR value of 0.043 started to repeat itself after the dataset reached a number of features of 8. For the FPR, the values that repeated the most include 0.097, 0.111, and 0.125. It is essential to consider that the FPR of 0.097 assisted in delivering the highest AUC value of the model. Another observation was in regards to the AUC value. This value fluctuated to different magnitudes depending on the size of the features in the dataset. There was no clear trend in the AUC value on this model. As the number of features increased, the AUC value increased and decreased until it finally reached it highest value at 0.977. The pictures of the top 3 models for the Symmetric Uncertainty feature ranking technique which included a feature dimensionality of 20, 50, and 100 can be observed below:



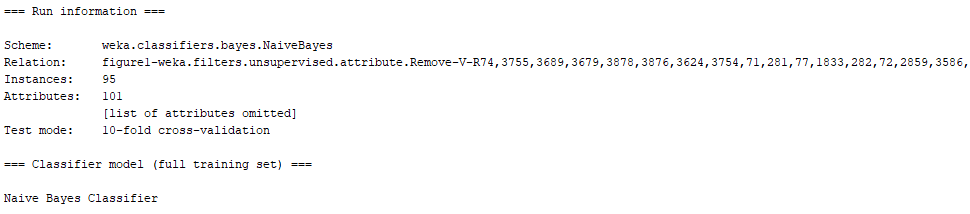


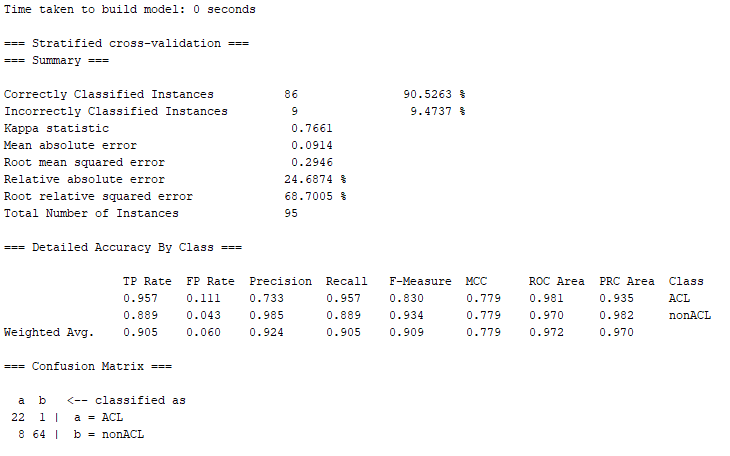
Picture 17: Naïve bayes Classifier with 20 Feature Dataset – Symmetric Uncertainty





Picture 18: Naïve bayes Classifier with 50 Feature Dataset – Symmetric Uncertainty

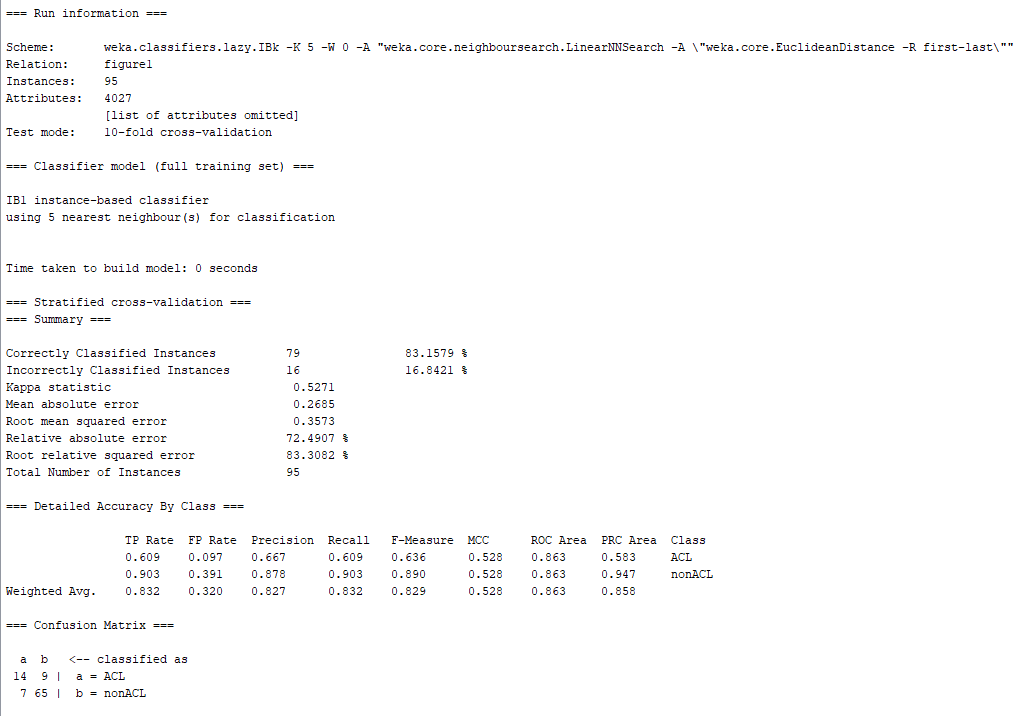




Picture 19: Naïve bayes Classifier with 100 Feature Dataset – Symmetric Uncertainty

Based on these results, it can be noted that the any implementation of feature ranking technique in a high dimensionality dataset will perform better and provide better performance in comparison to if no feature ranking technique is implemented. This can be concluded since the first model that was built using the naïve bayes classifier with the full dataset and no feature ranking technique performed the worst from all models built using the different raking techniques a feature dimensionality. Furthermore, it can be noted that the symmetric uncertainty feature ranking technique with a feature size of 20 provided the best AUC value of all models that were trained. On the next step, we conduct all the same training with the same data using the KNN classifier with a k value of 5.

***5-Nearest Neighbor Full Dataset Training Results:***



Picture 20: 5-Nearest Neighbor Classifier with Full Dataset

Based on the information from the picture above which displays the 5-Nearest Neighbor classifier training information, we can quickly determine the FPR, FNR, and AUC information. This data can be observed in table 1 below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.097 | 0.391 | 0.863 |

Table 7: 5-Nearest Neighbor Classifier with Full Dataset Performance Metrics

From table 1 above it can be noted that this classifier performed better at classifying the secondary (negative) class when compared to the primary class since it had a higher false negative rate of 0.391, while the secondary (negative) class had a lower misclassification rate 0.097. 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.

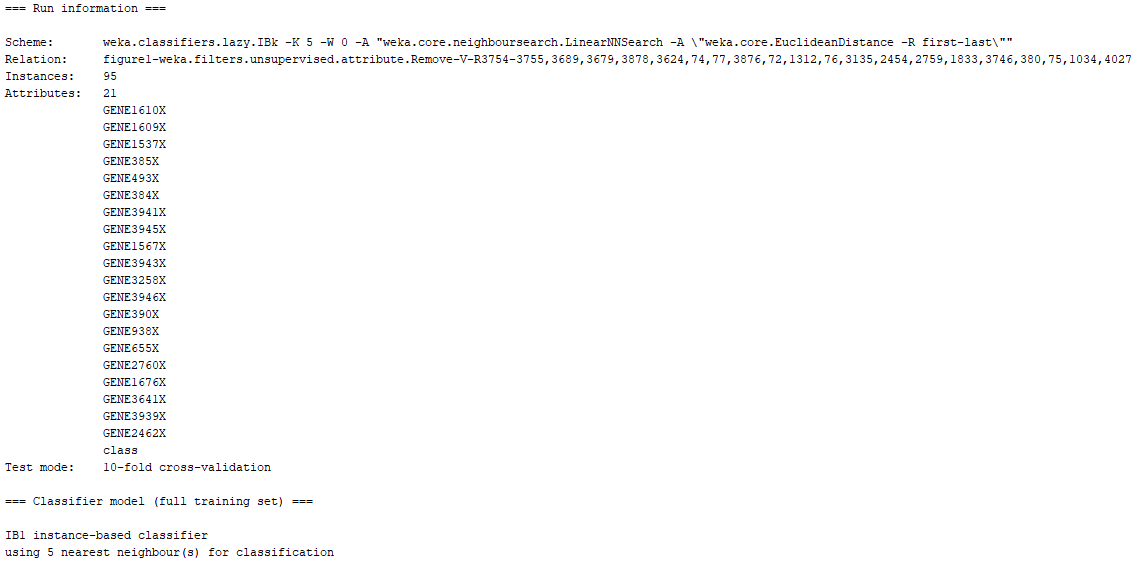
***5-Nearest Neighbor Chi Square with all Datasets Results:***

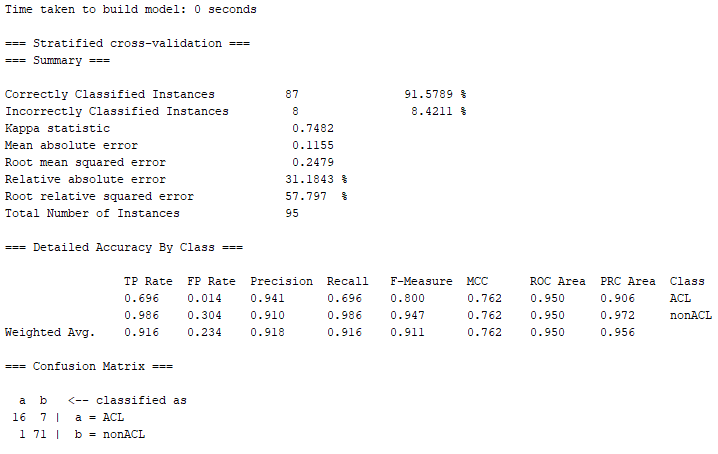
Next, the 5-Nearest Neighbor classifier will be trained using the dataset and feature sizes of the Chi Square feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the 5-Nearest Neighbor technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the chi square datasets for the 5-Nearest Neighbor classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| 5-Nearest Neighbor | 5 | 0.069 | 0.261 | 0.887 |
| 5-Nearest Neighbor | 6 | 0.042 | 0.304 | 0.918 |
| 5-Nearest Neighbor | 7 | 0.069 | 0.261 | 0.927 |
| 5-Nearest Neighbor | 8 | 0.069 | 0.217 | 0.934 |
| 5-Nearest Neighbor | 9 | 0.069 | 0.261 | 0.938 |
| 5-Nearest Neighbor | 10 | 0.069 | 0.304 | 0.927 |
| 5-Nearest Neighbor | 20 | 0.014 | 0.304 | 0.956 |
| 5-Nearest Neighbor | 50 | 0.028 | 0.217 | 0.952 |
| 5-Nearest Neighbor | 100 | 0.028 | 0.130 | 0.918 |
| 5-Nearest Neighbor | 200 | 0.028 | 0.130 | 0.943 |

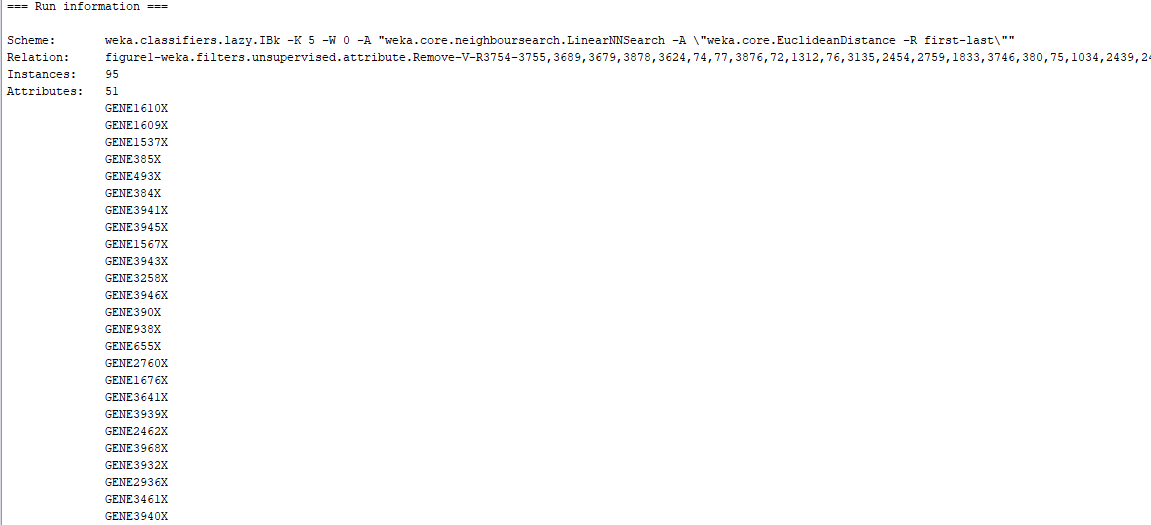
Table 8: 5-Nearest Neighbor Classifier with all Datasets for Chi Square Performance Metrics

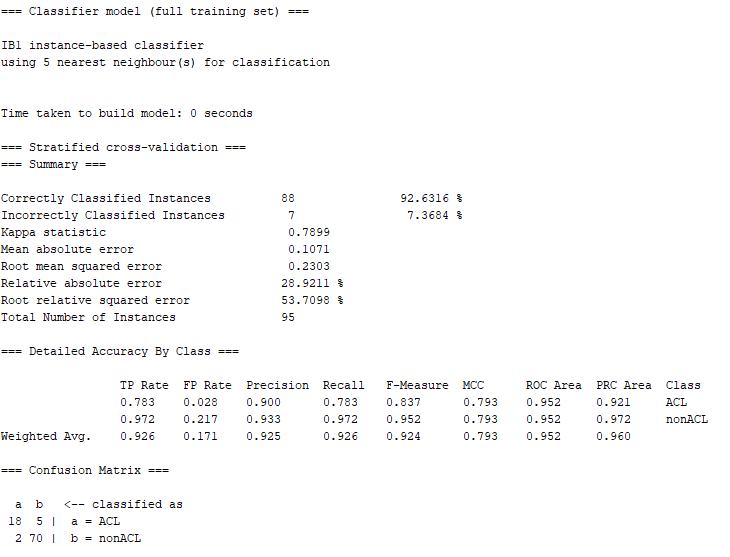
Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed decently with an AUC of 0.887. On the other side of the spectrum, when the model was trained with 200 the AUC did improve greatly only delivering a value of 0.943. It is necessary to note that there are no patterns from the FPR and FNR values from the table above. From an FNR perspective, the values remained constant throughout the different feature sizes. These values fluctuated varying in the size of features of the dataset. Nevertheless, most of the number repeated through each dataset. It is important to note as well that the FNR value which was the highest contributed to the best AUC value for the model. For the FPR value, it can be noted that t also continues to produce constant repetitive numbers. Nevertheless, one of the numbers that does not repeat in the series contributed to the highest AUC attained by the model. It is important to note that this was also the smallest value for FPR at 0.014. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC had very accurate but far from the best value it could get. As the dimensionality of the dataset increased, the AUC value also increased until reaching a value 0.956 when the model was trained with 20 features. After the number of features increased from 20, then the value of AUC decreased as more dimensionality was added decreasing the classification performance of the model. It is necessary to note that as the number of features in the dataset increased and the AUC value increased, when the dataset reach a number of features of 10, the increase in the AUC value broke and in this articular case, the AUC value decreased. Furthermore, based on the results from the table above, it can be concluded that when the model was trained using a feature dimensionality of 20, the classifiers had the best AUC and FPR performance metrics when compared to all other models, Although the FNR was not the best. Finally, when we compare the 20-feature model with the base model, it can be easily noted that the 20 feature possesses better performance metrics results when compared with its counterpart. The pictures of the top 3 models for the Chi Square feature ranking technique which included a feature dimensionality of 20, 50, and 200 can be observed below:



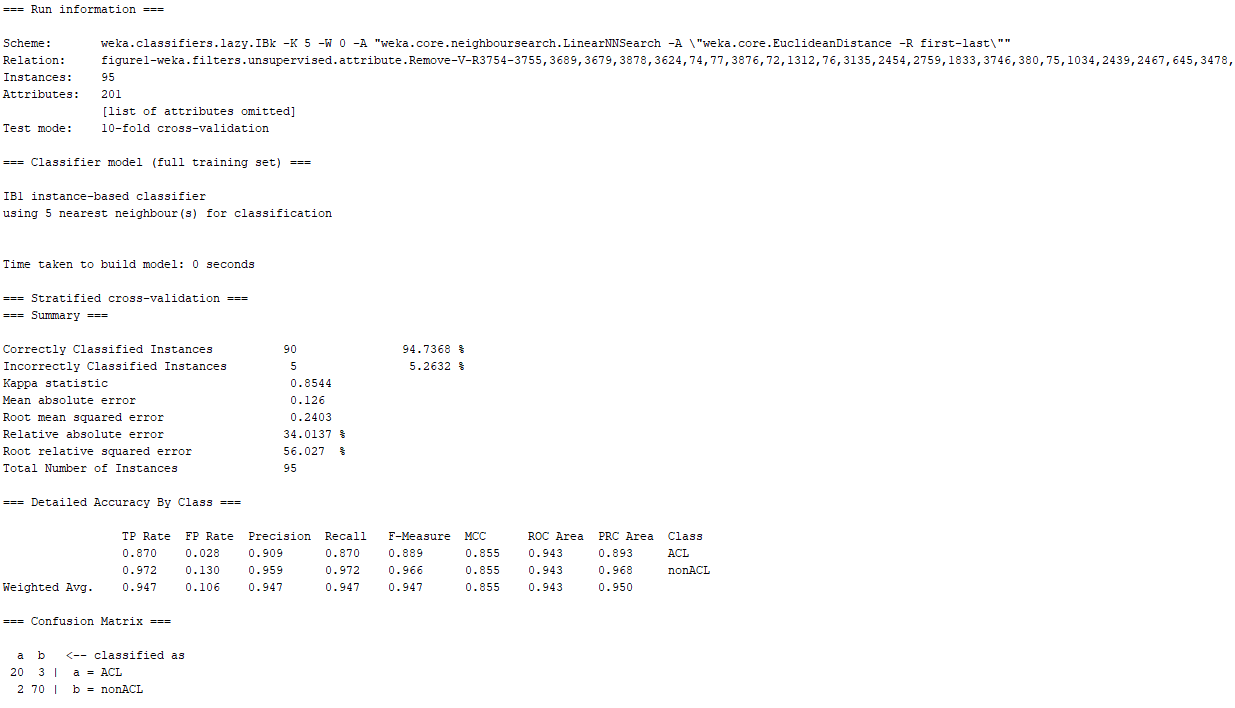


Picture 21: 5-Nearest Neighbor Classifier with 20 Feature Dataset – Chi Square





Picture 22: 5-Nearest Neighbor Classifier with 50 Feature Dataset – Chi Square



Picture 23: 5-Nearest Neighbor Classifier with 200 Feature Dataset – Chi Square

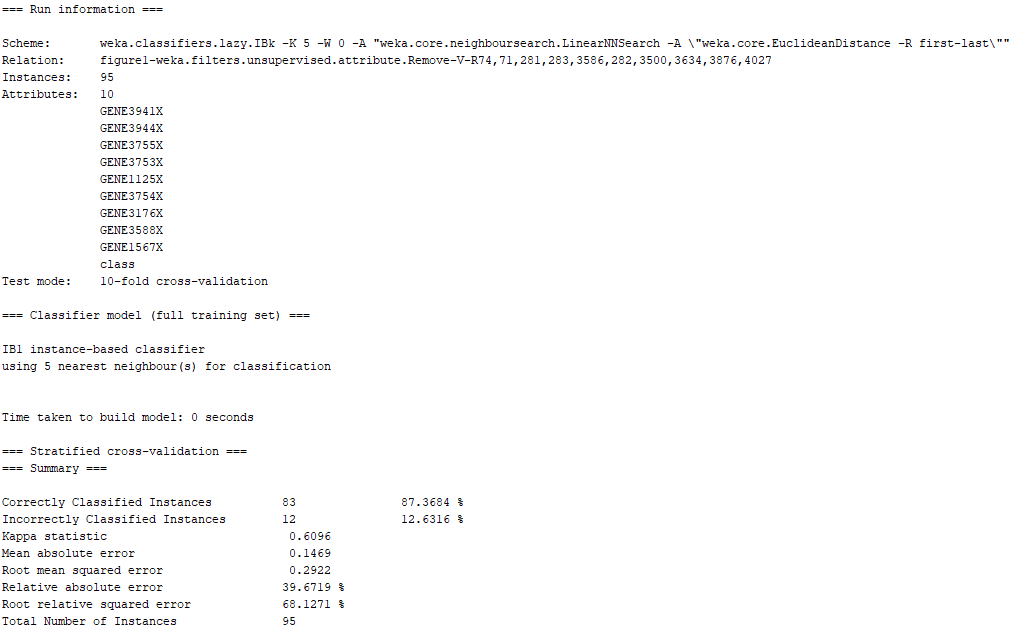
***5-Nearest Neighbor Gain Ratio with all Datasets Results:***

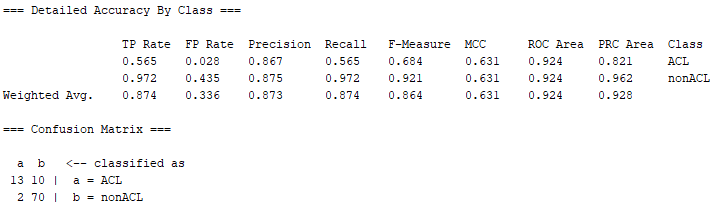
Next, the 5-Nearest Neighbor classifier will be trained using the dataset and feature sizes of the Gain Ratio feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the 5-Nearest Neighbor technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the Gain Ratio datasets for the 5-Nearest Neighbor classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| 5-Nearest Neighbor | 5 | 0.014 | 0.652 | 0.825 |
| 5-Nearest Neighbor | 6 | 0.014 | 0.696 | 0.853 |
| 5-Nearest Neighbor | 7 | 0.014 | 0.696 | 0.811 |
| 5-Nearest Neighbor | 8 | 0 | 0.696 | 0.812 |
| 5-Nearest Neighbor | 9 | 0.028 | 0.435 | 0.924 |
| 5-Nearest Neighbor | 10 | 0.042 | 0.348 | 0.922 |
| 5-Nearest Neighbor | 20 | 0.014 | 0.348 | 0.906 |
| 5-Nearest Neighbor | 50 | 0.014 | 0.217 | 0.912 |
| 5-Nearest Neighbor | 100 | 0.042 | 0.174 | 0.942 |
| 5-Nearest Neighbor | 200 | 0.056 | 0.130 | 0.920 |

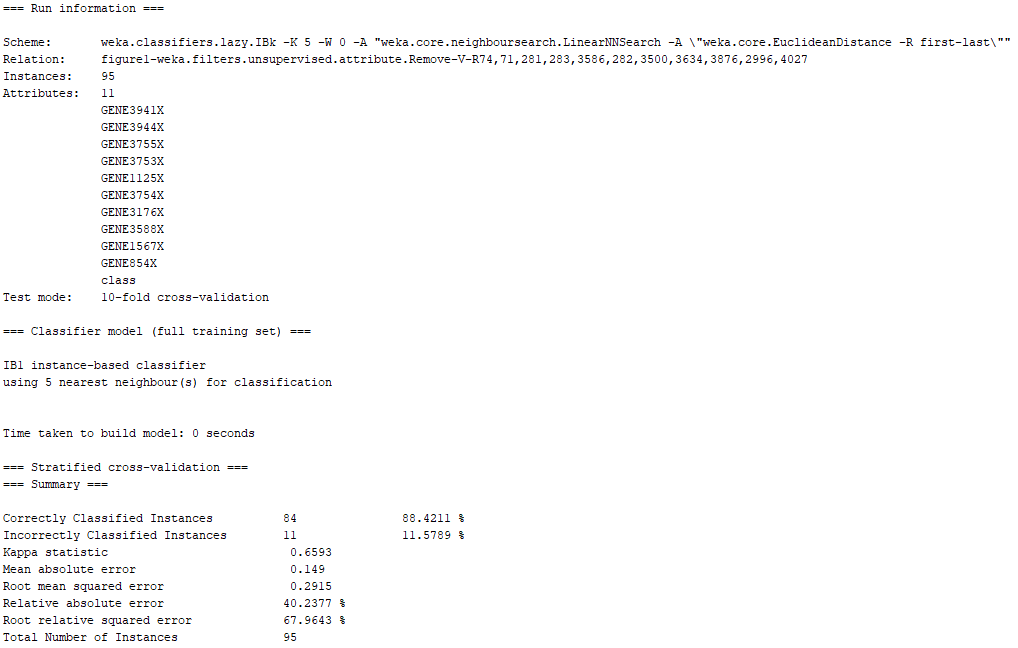
Table 8: 5-Nearest Neighbor Classifier with all Datasets for Gain Ratio Performance Metrics

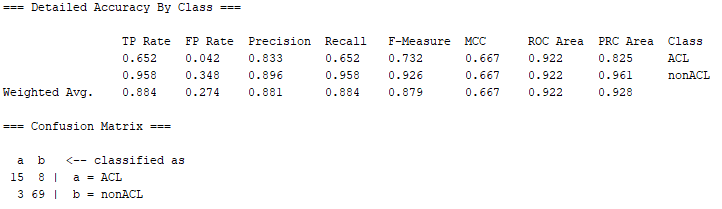
Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed decently with an AUC of 0.853. On the other side of the spectrum, when the model was trained with 200 the AUC did improve greatly only delivering a value of 0.920. It is necessary to note that there are no patterns from the FPR and FNR values from the table above. From an FNR perspective, the values remained constant throughout the different feature sizes. These values fluctuated varying in the size of features of the dataset. Nevertheless, most of the number repeated through each dataset. It is important to note as well that the FNR value of 0.174 contributed to the best AUC value for the model. For the FPR value, it can be noted that t also continues to produce constant repetitive numbers. Nevertheless, the number that contributed to the highest AUC attained by the model was 0.042. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC had a not very accurate value but it was far from the best value it could get. As the dimensionality of the dataset increased, the AUC fluctuated between lower and higher values until it reached 0.942 when the model was trained with 100 features. Furthermore, based on the results from the table above, it can be concluded that when the model was trained using a feature dimensionality of 100, the classifiers had the best AUC performance metrics when compared to all other models, Although the FNR and FPR were not the best. The pictures of the top 3 models for the gain ratio feature ranking technique which included a feature dimensionality of 9, 10, and 100 can be observed below:



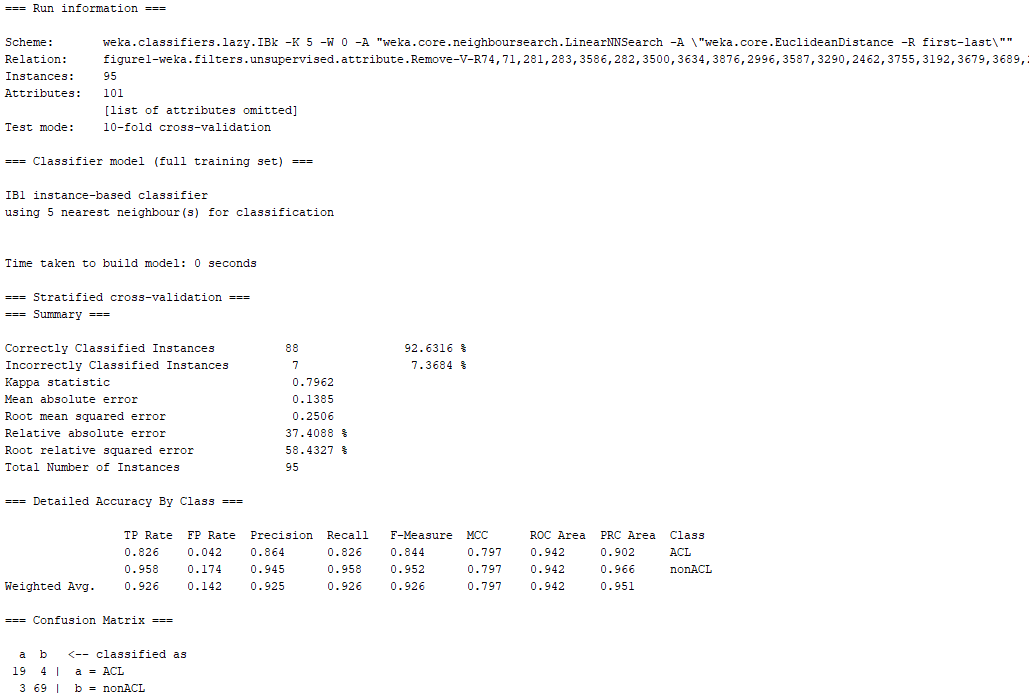


Picture 24: 5-Nearest Neighbor Classifier with 9 Feature Dataset – Gain Ratio





Picture 25: 5-Nearest Neighbor Classifier with 10 Feature Dataset – Gain Ratio



Picture 26: 5-Nearest Neighbor Classifier with 100 Feature Dataset – Gain Ratio

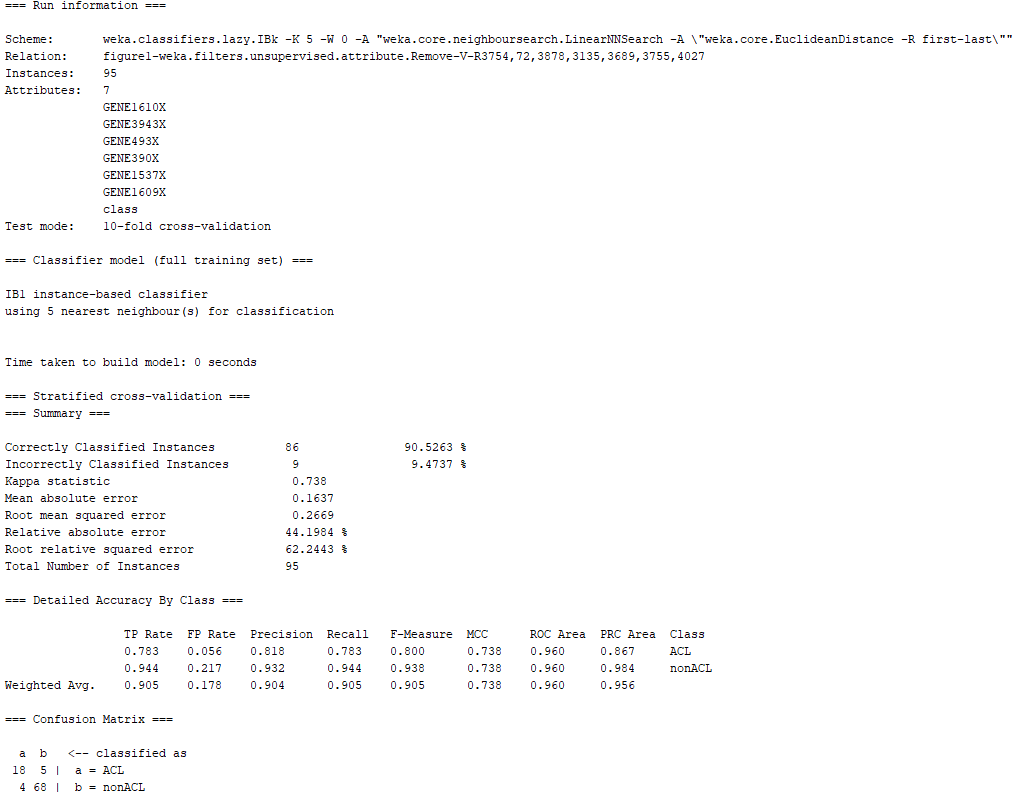
***5-Nearest Neighbor Information Gain with all Datasets Results:***

Next, the 5-Nearest Neighbor classifier will be trained using the dataset and feature sizes of the Information Gain feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the 5-Nearest Neighbor technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the Information Gain datasets for the 5-Nearest Neighbor classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| 5-Nearest Neighbor | 5 | 0.069 | 0.174 | 0.927 |
| 5-Nearest Neighbor | 6 | 0.056 | 0.217 | 0.960 |
| 5-Nearest Neighbor | 7 | 0.028 | 0.348 | 0.917 |
| 5-Nearest Neighbor | 8 | 0.028 | 0.304 | 0.929 |
| 5-Nearest Neighbor | 9 | 0.042 | 0.261 | 0.924 |
| 5-Nearest Neighbor | 10 | 0.042 | 0.304 | 0.939 |
| 5-Nearest Neighbor | 20 | 0.083 | 0.130 | 0.930 |
| 5-Nearest Neighbor | 50 | 0.056 | 0.087 | 0.956 |
| 5-Nearest Neighbor | 100 | 0.097 | 0.087 | 0.938 |
| 5-Nearest Neighbor | 200 | 0.056 | 0.130 | 0.944 |

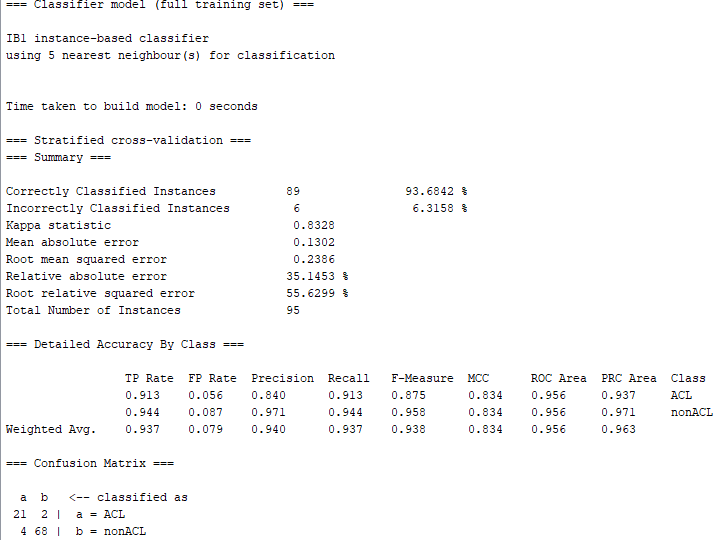
Table 8: 5-Nearest Neighbor Classifier with all Datasets for Information Gain Performance Metrics

Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed decently with an AUC of 0.927. On the other side of the spectrum, when the model was trained with 200 the AUC did improve greatly only delivering a value of 0.944. It is necessary to note that there are no patterns from the FPR and FNR values from the table above. From an FNR perspective, the values remained constant throughout the different feature sizes. These values fluctuated varying in the size of features of the dataset. Nevertheless, most of the number repeated through each dataset. It is important to note as well that the FNR value of 0.217 contributed to the best AUC value for the model. For the FPR value, it can be noted that t also continues to produce constant repetitive numbers. Nevertheless, the number that contributed to the highest AUC attained by the model was 0.056. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC had a not very accurate value but it was far from the best value it could get. As the dimensionality of the dataset increased, the AUC fluctuated between lower and higher values until it reached 0.960 when the model was trained with 6 features. Furthermore, based on the results from the table above, it can be concluded that when the model was trained using a feature dimensionality of 6, the classifiers had the best AUC performance metrics when compared to all other models, Although the FNR and FPR were not the best. The pictures of the top 3 models for the information gain feature ranking technique which included a feature dimensionality of 6, 50, and 200 can be observed below:

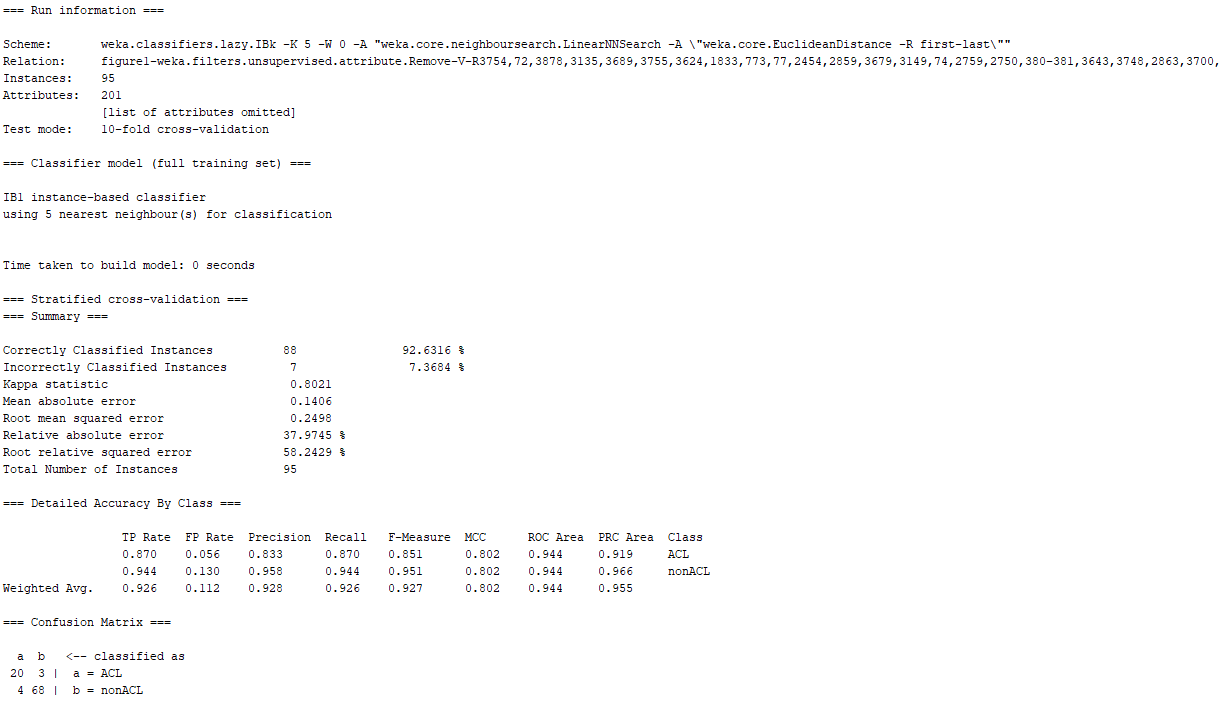


Picture 27: 5-Nearest Neighbor Classifier with 6 Feature Dataset – Information Gain





Picture 28: 5-Nearest Neighbor Classifier with 50 Feature Dataset – Information Gain



Picture 29: 5-Nearest Neighbor Classifier with 200 Feature Dataset – Information Gain

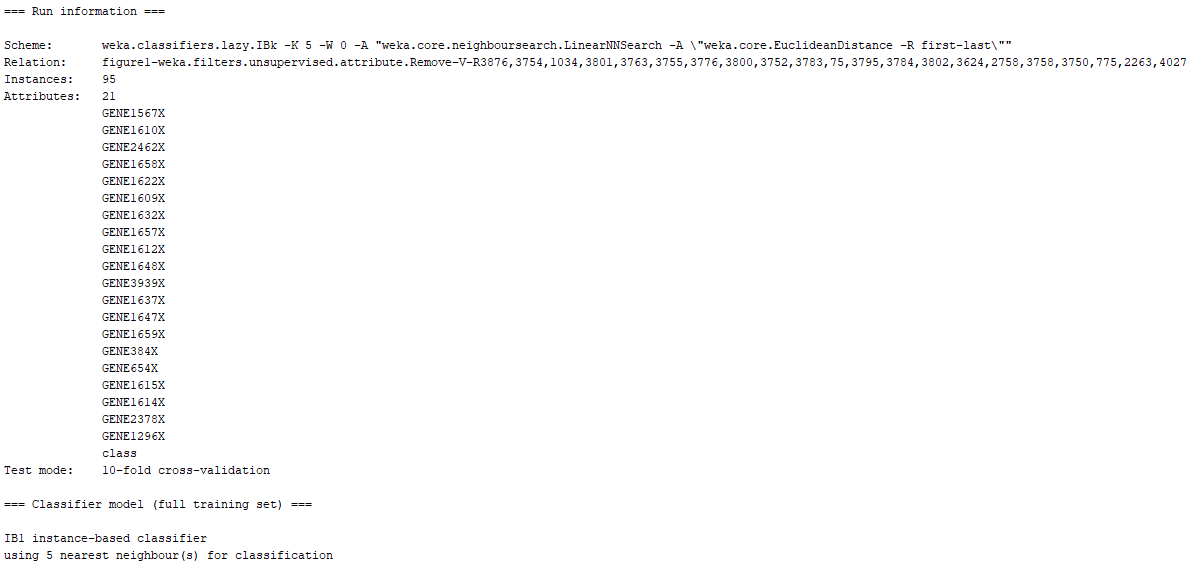
***5-Nearest Neighbor ReliefF with all Datasets Results:***

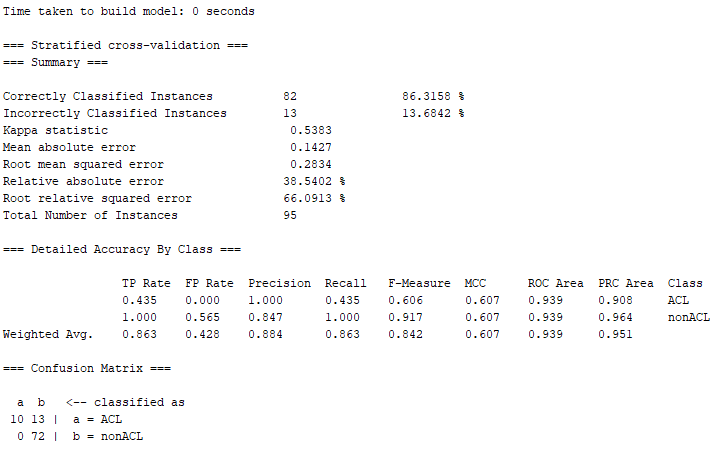
Next, the 5-Nearest Neighbor classifier will be trained using the dataset and feature sizes of the ReliefF feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the 5-Nearest Neighbor technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the ReliefF datasets for the 5-Nearest Neighbor classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| 5-Nearest Neighbor | 5 | 0.028 | 0.609 | 0.803 |
| 5-Nearest Neighbor | 6 | 0.014 | 0.522 | 0.788 |
| 5-Nearest Neighbor | 7 | 0.042 | 0.609 | 0.829 |
| 5-Nearest Neighbor | 8 | 0.042 | 0.565 | 0.851 |
| 5-Nearest Neighbor | 9 | 0.028 | 0.565 | 0.866 |
| 5-Nearest Neighbor | 10 | 0.014 | 0.609 | 0.899 |
| 5-Nearest Neighbor | 20 | 0 | 0.565 | 0.939 |
| 5-Nearest Neighbor | 50 | 0.014 | 0.391 | 0.899 |
| 5-Nearest Neighbor | 100 | 0 | 0.391 | 0.934 |
| 5-Nearest Neighbor | 200 | 0.014 | 0.348 | 0.908 |

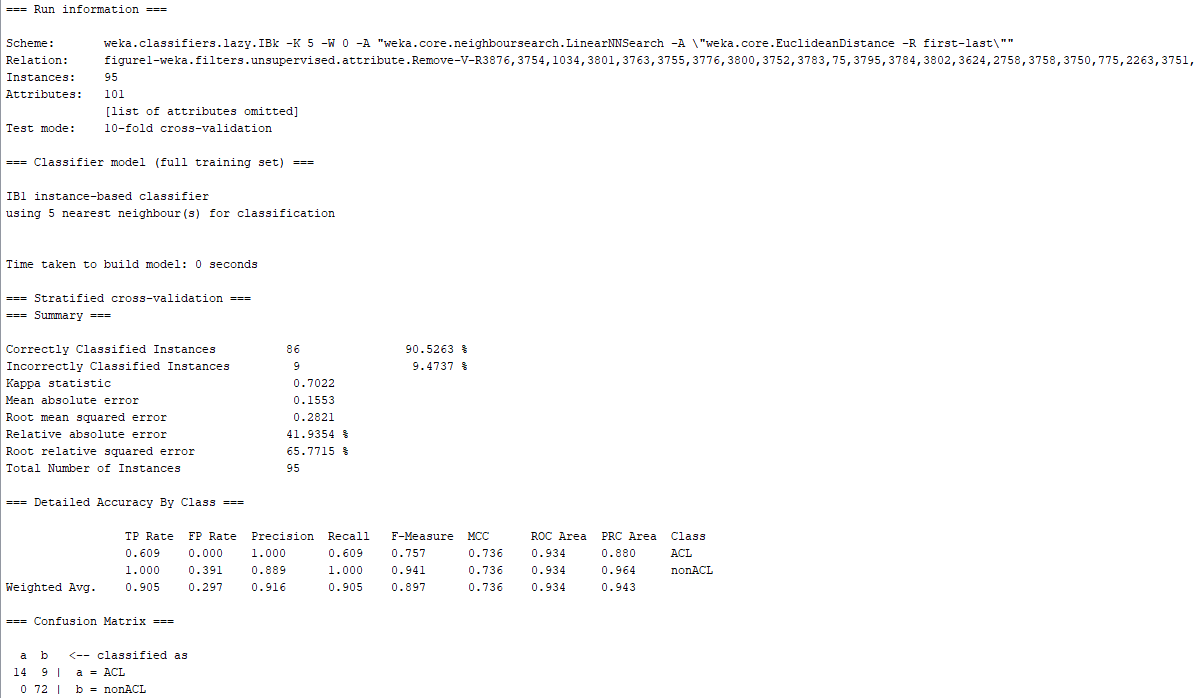
Table 9: 5-Nearest Neighbor Classifier with all Datasets for ReliefF Performance Metrics

Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed decently with an AUC of 0.803. On the other side of the spectrum, when the model was trained with 200 the AUC did improve greatly only delivering a value of 0.908. It is necessary to note that there are no patterns from the FPR and FNR values from the table above. From an FNR perspective, the values remained constant throughout the different feature sizes. These values fluctuated varying in the size of features of the dataset. Nevertheless, most of the number repeated through each dataset. It is important to note as well that the FNR value of 0.565 contributed to the best AUC value for the model. For the FPR value, it can be noted that t also continues to produce constant repetitive numbers. Nevertheless, the number that contributed to the highest AUC attained by the model was 0. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC had a not very accurate value but it was far from the best value it could get. As the dimensionality of the dataset increased, the AUC gradually increased until it reached 0.939 when the model was trained with 20 features. Furthermore, based on the results from the table above, it can be concluded that when the model was trained using a feature dimensionality of 20, the classifiers had the best AUC performance metrics when compared to all other models, Although the FNR was not the best. It is also essential to make mentioned that these models highly favored the classification of the majority class of the dataset even that the misclassification of the minority class was unprecedented. Furthermore, the AUC value provide for this model at is best is greatly misleading since when the highest AUC was obtained, no misclassifications occurred form the majority class while the minority class had a misclassification of almost half of all its samples in the dataset. Hence, even though that the model was indicating through a positive AUC value that the classifier was robust, in reality the classifier was not taking into consideration the proper classification of the minority class. The pictures of the top 3 models for the ReliefF feature ranking technique which included a feature dimensionality of 20, 100, and 200 can be observed below:

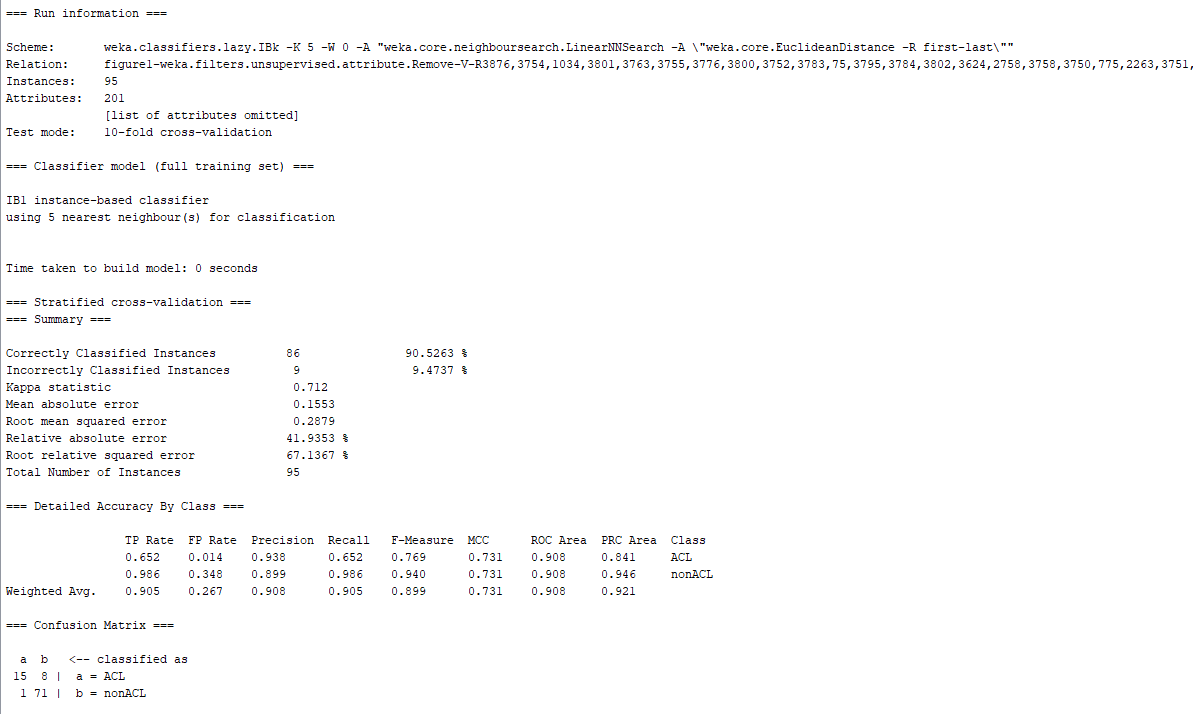




Picture 30: 5-Nearest Neighbor Classifier with 20 Feature Dataset – ReliefF



Picture 31: 5-Nearest Neighbor Classifier with 100 Feature Dataset – ReliefF



Picture 32: 5-Nearest Neighbor Classifier with 200 Feature Dataset – ReliefF

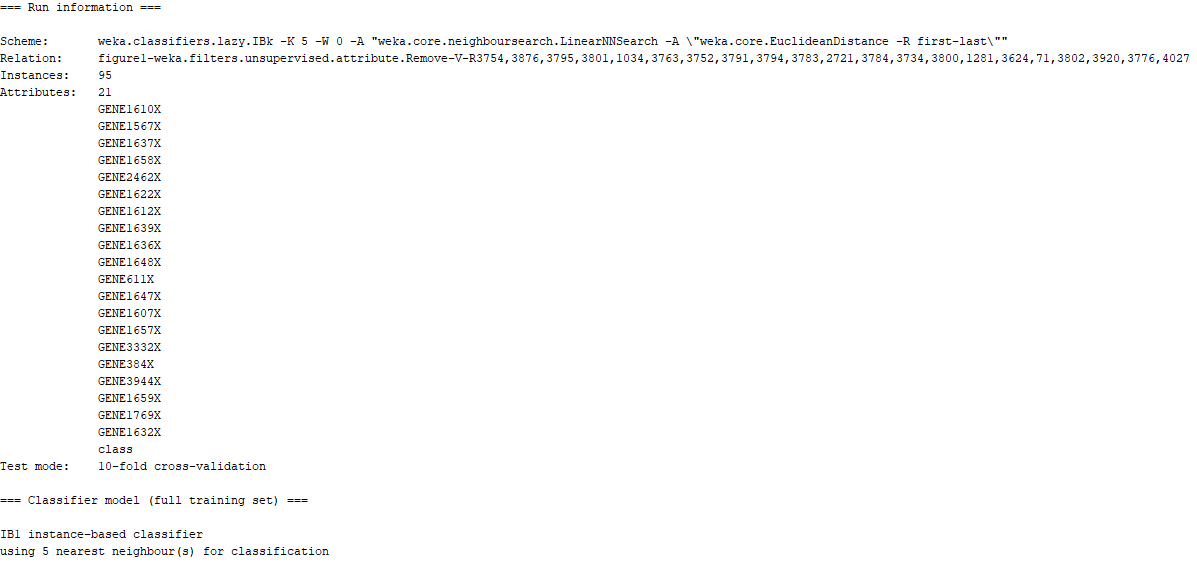
***5-Nearest Neighbor ReliefW with all Datasets Results:***

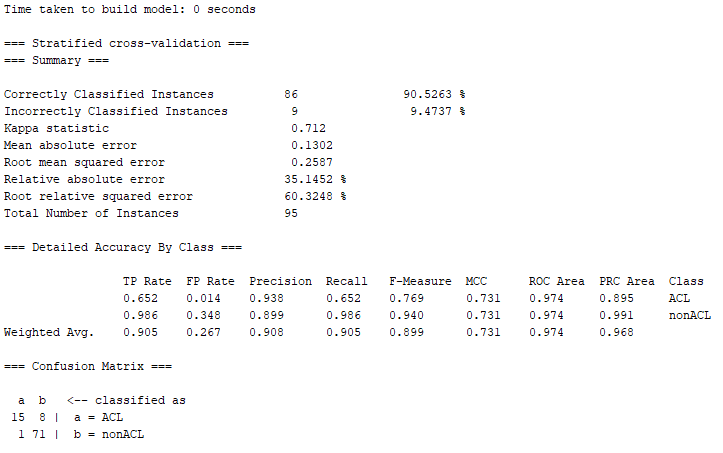
Next, the 5-Nearest Neighbor classifier will be trained using the dataset and feature sizes of the ReliefW feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the 5-Nearest Neighbor technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the ReliefW datasets for the 5-Nearest Neighbor classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| 5-Nearest Neighbor | 5 | 0.056 | 0.609 | 0.836 |
| 5-Nearest Neighbor | 6 | 0.056 | 0.652 | 0.824 |
| 5-Nearest Neighbor | 7 | 0.042 | 0.739 | 0.816 |
| 5-Nearest Neighbor | 8 | 0.042 | 0.565 | 0.844 |
| 5-Nearest Neighbor | 9 | 0.014 | 0.478 | 0.853 |
| 5-Nearest Neighbor | 10 | 0.028 | 0.565 | 0.899 |
| 5-Nearest Neighbor | 20 | 0.014 | 0.348 | 0.974 |
| 5-Nearest Neighbor | 50 | 0 | 0.348 | 0.979 |
| 5-Nearest Neighbor | 100 | 0.028 | 0.261 | 0.939 |
| 5-Nearest Neighbor | 200 | 0.056 | 0.261 | 0.927 |

Table 9: 5-Nearest Neighbor Classifier with all Datasets for ReliefW Performance Metrics

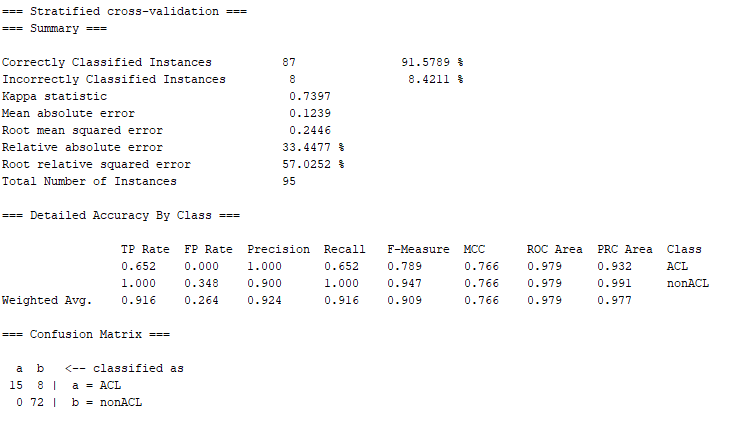
Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed decently with an AUC of 0.836. On the other side of the spectrum, when the model was trained with 200 the AUC did improve greatly only delivering a value of 0.927. It is necessary to note that there are no patterns from the FPR and FNR values from the table above. From an FNR perspective, the values remained constant throughout the different feature sizes. These values fluctuated varying in the size of features of the dataset. Nevertheless, most of the number repeated through each dataset. It is important to note as well that the FNR value of 0.348 contributed to the best AUC value for the model. For the FPR value, it can be noted that t also continues to produce constant repetitive numbers. Nevertheless, the number that contributed to the highest AUC attained by the model was 0. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC had a not very accurate value but it was far from the best value it could get. As the dimensionality of the dataset increased, the AUC gradually increased until it reached 0.979 when the model was trained with 50 features. Furthermore, based on the results from the table above, it can be concluded that when the model was trained using a feature dimensionality of 50, the classifiers had the best AUC performance metrics when compared to all other models, although the FNR was not the best. It is also essential to make mentioned that these models highly favored the classification of the majority class of the dataset even though that the misclassification of the minority class was unprecedented. Furthermore, the AUC value provide for this model at its best is greatly misleading since when the highest AUC was obtained, no misclassifications occurred form the majority class while the minority class had a misclassification of a great number of its samples in the dataset. Hence, even though that the model was indicating through a positive AUC value that the classifier was robust, in reality the classifier was not taking into consideration the proper classification of the minority class. The pictures of the top 3 models for the ReliefW feature ranking technique which included a feature dimensionality of 20, 50, and 100 can be observed below:



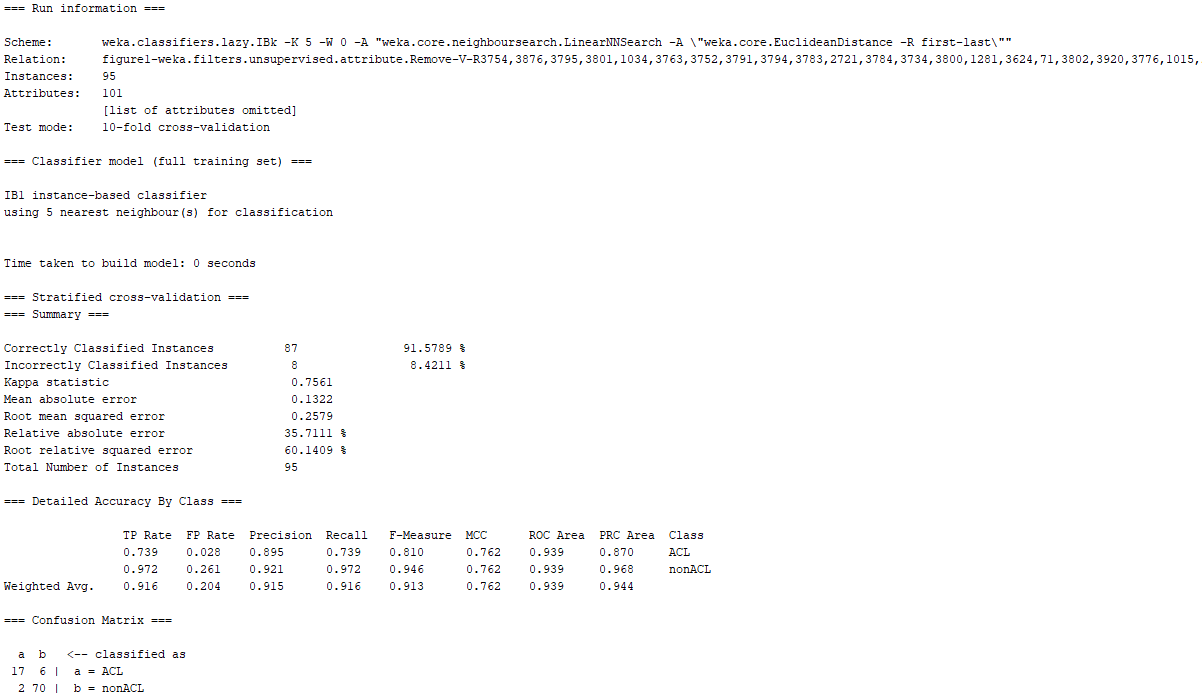


Picture 33: 5-Nearest Neighbor Classifier with 20 Feature Dataset – ReliefW





Picture 34: 5-Nearest Neighbor Classifier with 50 Feature Dataset – ReliefW



Picture 35: 5-Nearest Neighbor Classifier with 100 Feature Dataset – ReliefW

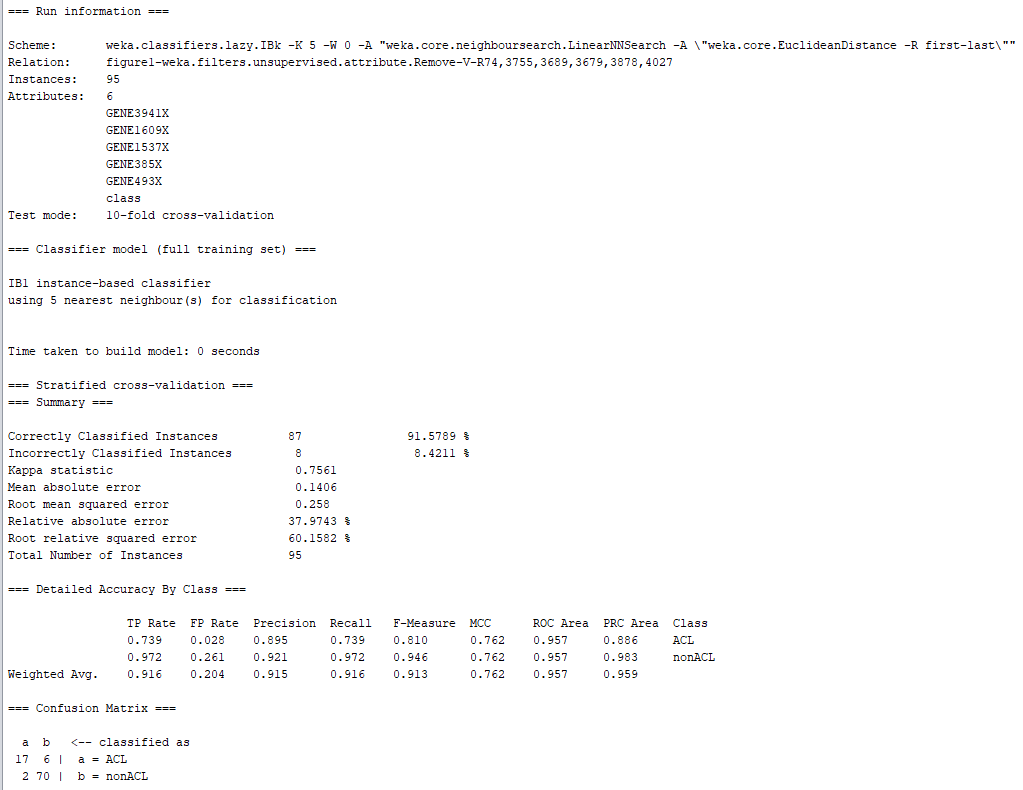
***5-Nearest Neighbor Symmetric Uncertainty with all Datasets Results:***

Next, the 5-Nearest Neighbor classifier will be trained using the dataset and feature sizes of the Symmetric Uncertainty feature selection technique. The model will be trained 10 times, once with each dataset, and place the results in a table for comparison and evaluation. Moreover, the pictures from the top 3 models which deliver the best classification performance for the 5-Nearest Neighbor technique will be selected to add in this paper, although all the metrics from all classifiers results will be used to conduct the analysis. The table with the results for all the Symmetric Uncertainty datasets for the 5-Nearest Neighbor classifier can be observed below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Number of Features | False Positive Rate (FPR) | False Negative Rate (FNR) | Area Under ROC Curve (ROC) |
| 5-Nearest Neighbor | 5 | 0.028 | 0.261 | 0.957 |
| 5-Nearest Neighbor | 6 | 0.042 | 0.261 | 0.918 |
| 5-Nearest Neighbor | 7 | 0.042 | 0.217 | 0.938 |
| 5-Nearest Neighbor | 8 | 0.069 | 0.217 | 0.920 |
| 5-Nearest Neighbor | 9 | 0.042 | 0.348 | 0.934 |
| 5-Nearest Neighbor | 10 | 0.028 | 0.304 | 0.944 |
| 5-Nearest Neighbor | 20 | 0.056 | 0.217 | 0.941 |
| 5-Nearest Neighbor | 50 | 0.056 | 0.087 | 0.949 |
| 5-Nearest Neighbor | 100 | 0.069 | 0.087 | 0.954 |
| 5-Nearest Neighbor | 200 | 0.042 | 0.130 | 0.944 |

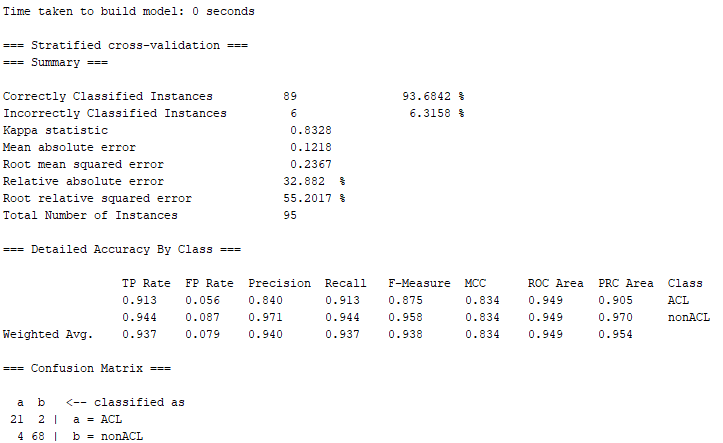
Table 10: 5-Nearest Neighbor Classifier with all Datasets for Symmetric Uncertainty Performance Metrics

Based on the information from the table above it can be noted that when the dataset contained the top 5 features, the model performed decently with an AUC of 0.957. On the other side of the spectrum, when the model was trained with 200 the AUC did improve greatly only delivering a value of 0.944. It is necessary to note that there are no patterns from the FPR and FNR values from the table above. From an FNR perspective, the values remained constant throughout the different feature sizes. These values fluctuated varying in the size of features of the dataset. Nevertheless, most of the number repeated through each dataset. It is important to note as well that the FNR value of 0.261 contributed to the best AUC value for the model. For the FPR value, it can be noted that it also continues to produce constant repetitive numbers. Nevertheless, the number that contributed to the highest AUC attained by the model was 0.028. Another curious observation was in regards to the AUC value. When the number of features was the lowest, the AUC was the highest at 0.957. As the dimensionality of the dataset increased, the AUC fluctuated through different values deviating from the highest number. Furthermore, based on the results from the table above, it can be concluded that when the model was trained using a feature dimensionality of 5, the classifiers had the best AUC performance metrics when compared to all other models, although the FNR was not the best. It is also essential to make mentioned that these models highly favored the classification of the majority class of the dataset even though that the misclassification of the minority class was very disproportional. Furthermore, the AUC value provide for this model at its best is greatly misleading since when the highest AUC was obtained, very few misclassifications occurred form the majority class while the minority class had a misclassification of a great number of its samples in the dataset. Hence, even though that the model was indicating through a positive AUC value that the classifier was robust, in reality the classifier was not taking into consideration the proper classification of the minority class. The pictures of the top 3 models for the Symmetric Uncertainty feature ranking technique which included a feature dimensionality of 5, 50, and 100 can be observed below:

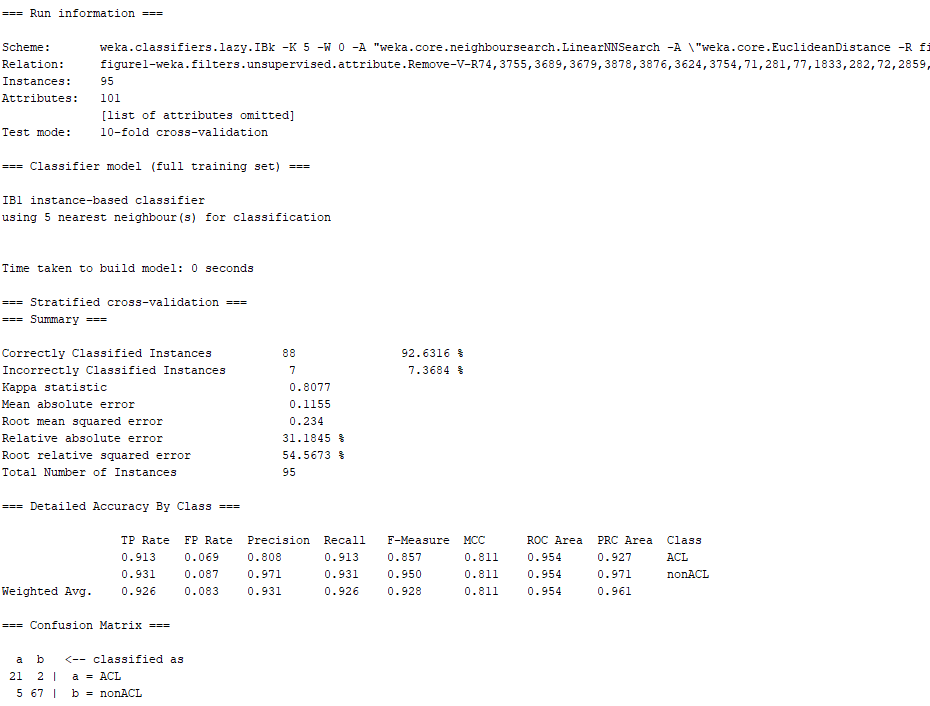


Picture 36: 5-Nearest Neighbor Classifier with 6 Feature Dataset – Symmetric Uncertainty





Picture 37: 5-Nearest Neighbor Classifier with 50 Feature Dataset – Symmetric Uncertainty



Picture 38: 5-Nearest Neighbor Classifier with 100 Feature Dataset – Symmetric Uncertainty

Based on these experiments it can be concluded that 5NN classifier did not perform as well as the Naïve Bayes since it did not matter which feature ranking technique was implemented, 5NN model always favored the proper classification of the majority class in the dataset. Although, the AUC for the KNN were high in magnitude, the truth is that also the FNR was extremely high while the FPR at some points were 0. This type of observation leads us to believe that it did not matter which feature ranking technique was implemented with the KNN, the model will always favor the majority class. On the other hand, the naïve bayes performed better at favoring the classification of the minority class with the lowest FNR between the two classifiers. Nevertheless, the naïve bayes favored the minority class to the point that different feature ranking techniques were misclassifying only one of the samples from the minority class while misclassing a bigger number for the majority class. Nevertheless, through the application of feature ranking techniques, it was possible to find a balance of FPR and FNR for specific feature ranking when implementing the naïve bayes model. From a feature selection stand point, it is essential to make note that the best performance of the classifiers either naïve or KNN, was attained when the number of features in the dataset was 20. With this size feature in the dataset regardless of the ranker, the models were able to provide the AUC value. Additionally, from a feature ranker stand point, it is necessary to make mention that for the naïve bayes, the chi square and symmetric uncertainty provided the best AUC for the model overall. In the same note, the worst feature rankers for naïve bayes were the two different types of relief. From a KNN stand point the best feature ranker was the symmetric uncertainty while the worst rankers were the two types of relief. Finally, it is important to emphasize that between the two models, the naïve bayes provided better performance with a balance value between the FNR and FPR which also provide a high AUC. With this information it can be concluded that naïve bayes produced a more balanced and robust classification between the classes of the dataset. On the other hand, the KNN model, although it had high values of AUC after being trained, the truth is that the misclassification of the minority was always greater regardless of the ranker used.