Z Number: 23226341

Name: Carlos Gross-Martinez

Professor: Dr. Taghi M. Khoshgoftaar

Course: CAP6778 – Advanced Data Mining & Machine Learning

Assignment 5: Random Under-sampling

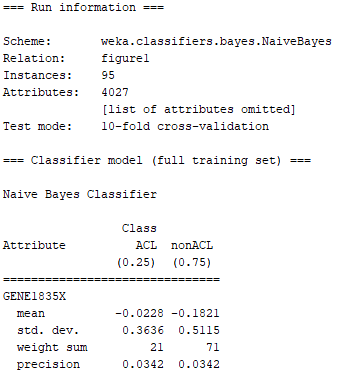
Dataset Analysis

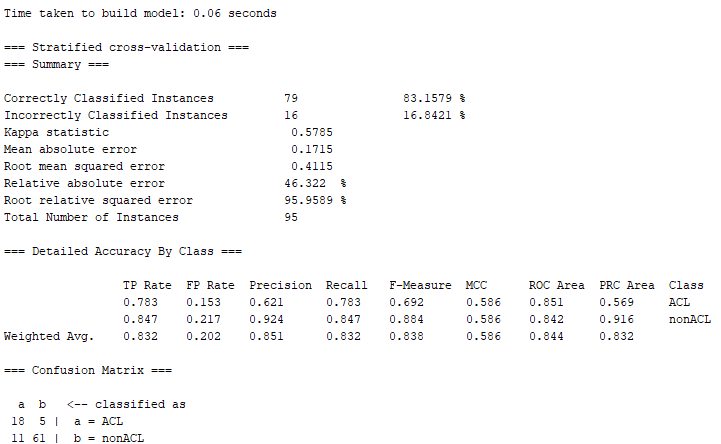
The dataset used in this assignment consisted of 95 instances/samples and 4027 attributes. From the 4027 attributes 4026 were input attributes to the model while 1 attribute was the class label. Furthermore, out of the 95 samples provided in the dataset, 23 were from the minority class labeled “ACL” 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”. Therefore, the class imbalance ratio of the dataset can be determined to be 76:24 where 76% reflects the majority class while 24% reflects the minority class.

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 with the full dataset, an evaluation of their results will be conducted with emphasis on the false positive rate, false negative rate, and area under ROC curve performance metrics. The evaluation will be conducted Individually for each model, and then, both models will be compared. 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. Once Finally, the results from both classifiers will be compared.

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





Picture 1: Naïve bayes Classifier with Full Dataset

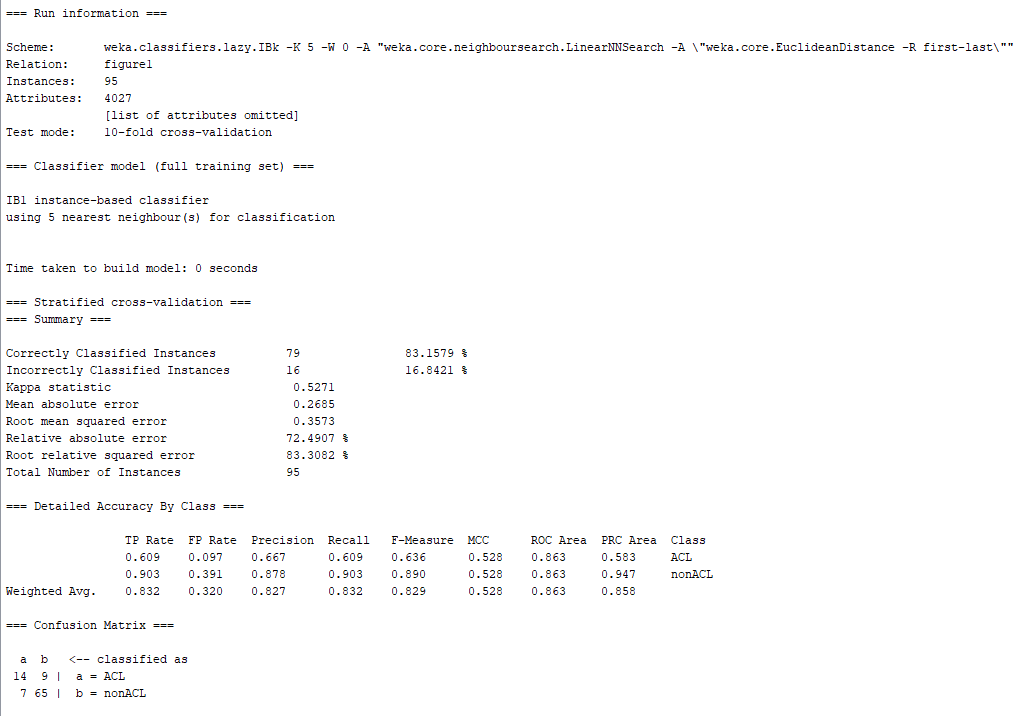
Based on the information from the picture above which displays the Naïve Bayes classifier training information, we can quickly determine the FPR, FNR, and AUC information. This data can be observed in table 1 below:

|  |  |  |  |
| --- | --- | --- | --- |
| 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. Based on this information, it can be concluded that the model is biased and greatly favors the proper classification of the majority class. This can be determined because the FPR value, which reflects the misclassification of the majority class, is reduced by 30% from the value of the FNR of 0.217. It is important to note that the overall performance of the classifier under the ROC value was satisfactory at 0.844. From these observations, it can be concluded that this classifier is not properly balanced nor robust since it favors greatly favors the classification of the majority class in the dataset.

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



Picture 2: 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) |
| 5-Nearest Neighbor | 0.097 | 0.391 | 0.863 |

Table 2: 5-Nearest Neighbor Classifier with Full Dataset Performance Metrics

From table 2 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. Based on this information, it can be concluded that the model is biased and overwhelmingly favors the proper classification of the majority class. This can be determined because the FPR value, which reflects the misclassification of the majority class, is reduced by 75% from the value of the FNR of 0.391. It is important to note that the overall performance of the classifier under the ROC value was satisfactory at 0.863. Nevertheless, this model cannot be considered a balanced classifier since it greatly favors the proper classification of the majority class in the dataset.

With the training of models completed and noted, it is now possible to conduct an analysis on both classifiers’ performance metrics. The summarize performance metrics for both classifiers trained with the full dataset can be observed in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.153 | 0.217 | 0.844 |
| 5-Nearest Neighbor | 0.097 | 0.391 | 0.863 |

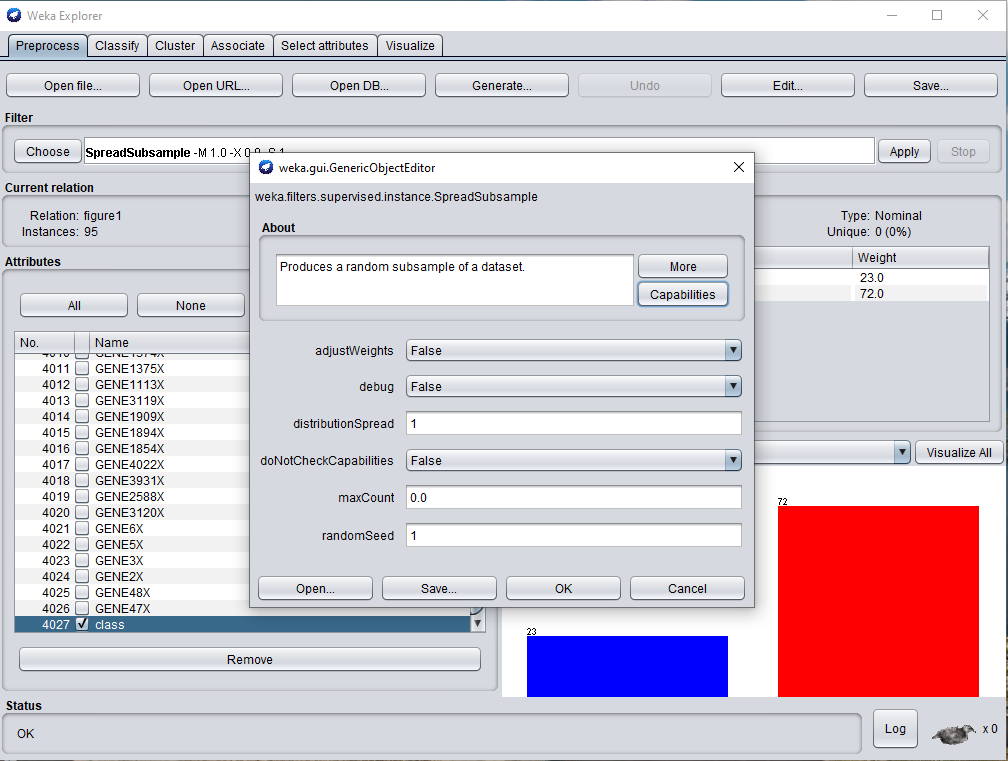
Table 3: Naïve Bayes & 5-Nearest Neighbor Classifiers with Full Dataset Performance Metrics

Based on the information of the table above, it can be easily concluded that the naïve bayes model had a higher FPR rate when compared to the 5-nearest neighbors. This means that naïve bayes model misclassified more data samples from the majority class when compare to its counterpart. Moreover, it can also be observed that the naïve bayes model had a lower FNR rate when compared to the 5-nearest neighbor. This means that 5-nearest neighbor misclassified more data samples from the minority class when compared to its counterpart. Furthermore, it can also be noted that the ROC value for the naïve bayes classifier was lower when compared to the 5-neares neighbor. Therefore, based on the ROC metric it can be concluded that 5-nearest neighbor had better performance results when compare to its counterpart. Nevertheless, it is essential to make mention that the 5-nearest neighbor model had a higher misclassification rate of the minority class and greatly favored the proper classification of the majority class. Hence, the 5-nearest neighbor is not a balance model nor robust as the minority class in the dataset which is the most important class contains an extremely high misclassification under this model. On the other hand, the naïve baes model had a higher misclassification of the majority class, but performed better at classifying the minority class in the dataset. Therefore, it can be concluded that between the two models, naïve bayes has a more balance classification of both classes and hence is a more robust model when compare to its counterpart.

Part 2: Random Under-sampling

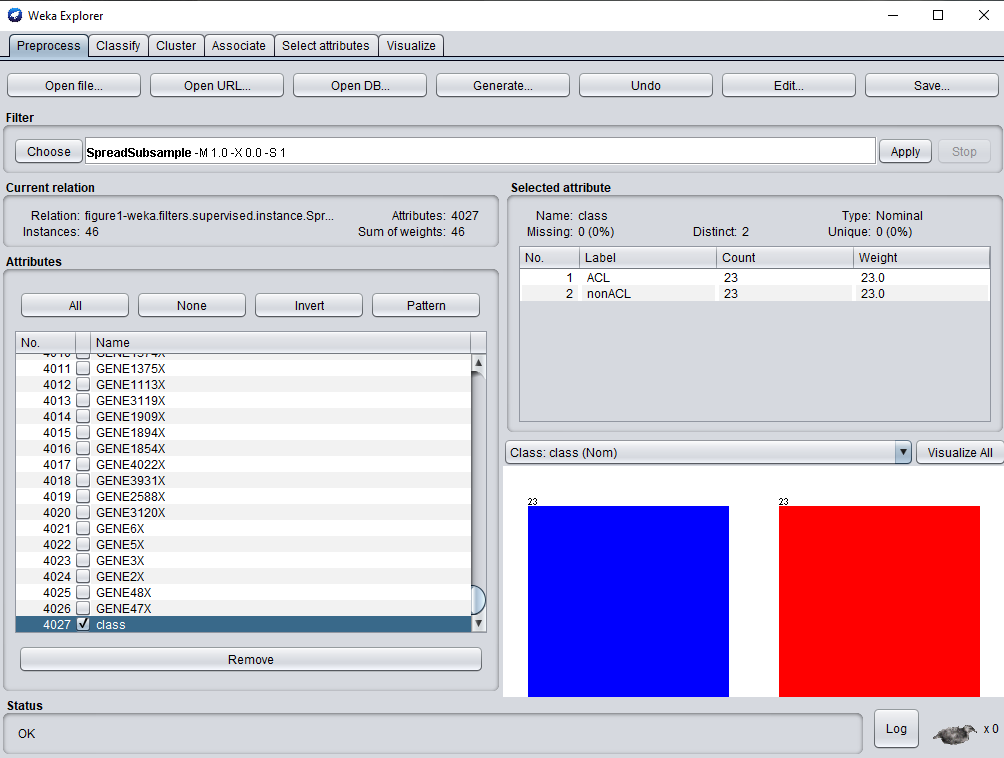
In this segment of the assignment, one filter will be applied to the full dataset in order to implement a random under-sampling technique and create a more balance subset of the full dataset. The random under-sampling technique is applied as a strategy to accomplish a more balance training dataset which can help to build a more robust model where the bias towards the proper classification of the majority class in the dataset can be greatly reduced when building the models. The filter to conduct the random under-sampling is provided in the Weka tool and is known as “SpreadSubsample”. This will be used to create subsets of the dataset by reducing the number of samples of the majority class to the ratios specified in the in the assignment. The ratios of dataset will include 50:50 and 65:35 splits. Currently the dataset contains a distribution ratio of 76:24, where 76% of the samples in the dataset represent the majority class while 24% represents the minority class. On the split of 50:50 ratio, the under-sampling technique will be applied in order to reduce the number of samples of the majority class randomly to match the number of samples in the minority class. Therefore, after the under-sampling technique has been applied to the full dataset, there will be a total of 46 samples in the dataset. 23 from the minority class and 23 from the majority class. On the other hand, on the 65:35 split, after the under-sampling technique has been applied to the full dataset, there will be a total of 66 samples in the dataset. 23 from the minority class representing 35% of the dataset while 43 samples will be from the majority class representing 65% of the dataset.

The first split carried out to train the models will be the 50:50. In order to accomplish the 50:50 split, the “distributionSpread” value for the “SpreadSubsample” filter will be set to 1 in order to obtain a uniform distribution of the dataset. A picture with the setting configured in Weka can be observed below:



Picture 3: Distribution Spread Configuration in Weka for 50:50 Split

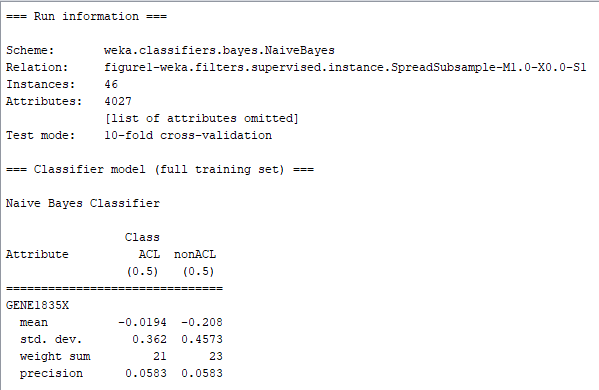
With the configuration completed and the filter applied to the dataset the final dataset with a 50:50 split after the random under-sampling technique has been applied can be observed below:

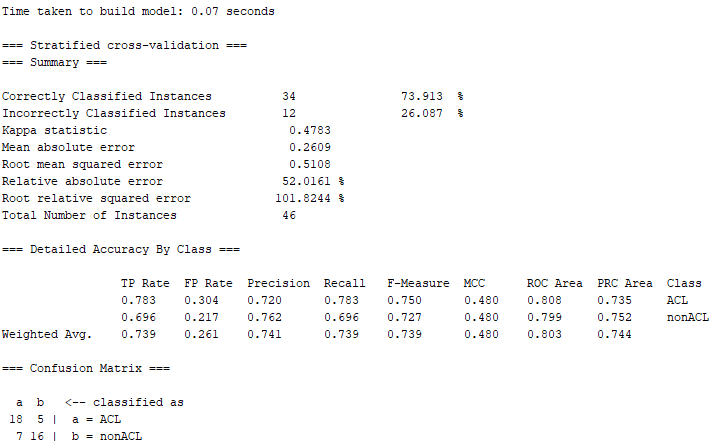


Picture 4: Dataset After Random Under-Sampling Technique

***Naïve Bayes 50:50 Dataset Training Results:***

The first model to be trained with the 50:50 split dataset will be the naïve bayes. The results from training the model with new dataset can be observed below:





Picture 5: Naïve bayes Classifier with 50:50 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 4 below:

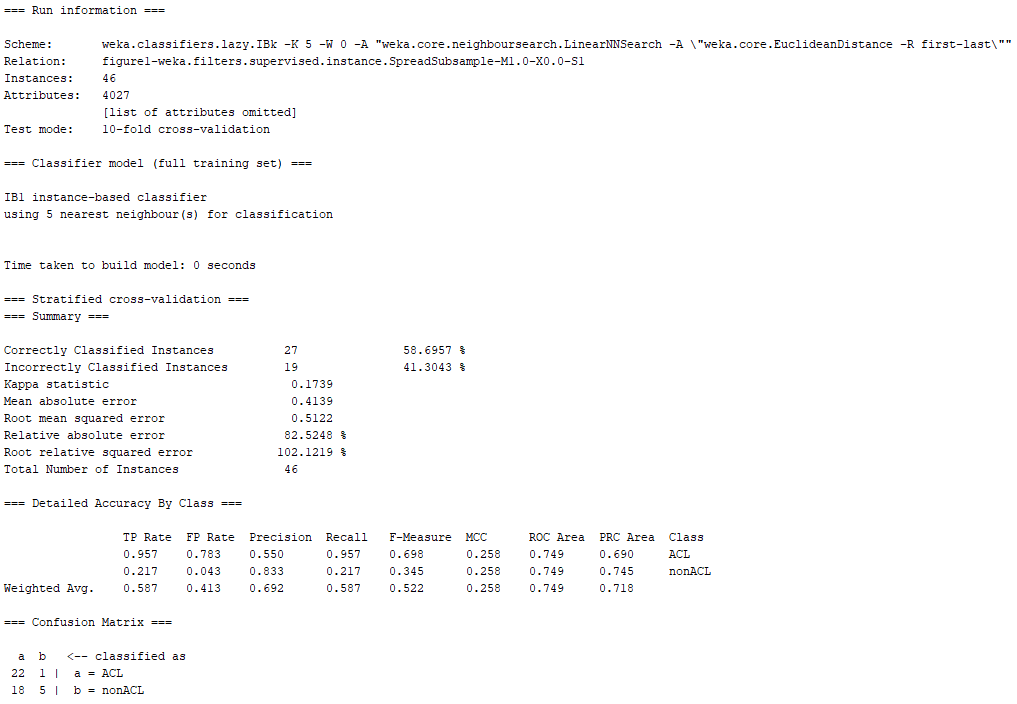
|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.304 | 0.217 | 0.803 |

Table 4: Naïve Bayes Classifier with 50:50 Dataset Performance Metrics

From table 4 above it can be noted that this classifier performed better at classifying the primary (positive) class when compared to the secondary class since it had a lower false negative rate of 0.217, while the secondary (negative) class had a higher misclassification rate 0.304. Based on this information, it can be concluded that the model slightly favors the proper classification of the minority class. This can be determined because the FPR value, which reflects the misclassification of the majority class, is increased by 29% from the value of the FNR of 0.217. It is important to note that the overall performance of the classifier under the ROC value was not the best at 0.803. From these observations, it can be concluded that this classifier slightly favors the proper classification of the minority class of the dataset.

***5-Nearest Neighbor 50:50 Dataset Training Results:***

The next model to be trained with the 50:50 dataset is the 5-nearest neighbor. The results from training the classifier can be observed below:



Picture 6: 5-Nearest Neighbor Classifier with 50:50 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) |
| 5-Nearest Neighbor | 0.783 | 0.043 | 0.749 |

Table 5: 5-Nearest Neighbor Classifier with 50:50 Dataset Performance Metrics

From table 5 above it can be noted that this classifier performed better at classifying the primary (positive) class when compared to the secondary class since it had a lower false negative rate of 0.043, while the secondary (negative) class had a higher misclassification rate 0.783. Based on this information, it can be concluded that the model is biased and overwhelmingly favors the proper classification of the minority class. This can be determined because the FPR value, which reflects the misclassification of the majority class, is increased by 95% from the value of the FNR of 0.043. It is important to note that the overall performance of the classifier under the ROC value was not very satisfactory at only 0.749. Nevertheless, this model cannot be considered a balanced classifier since it monumentally favors the proper classification of the minority class in the dataset.

With the training of models completed and noted, it is now possible to conduct an analysis on both classifiers’ performance metrics. The summarize performance metrics for both classifiers trained with the 50:50 dataset can be observed in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.304 | 0.217 | 0.803 |
| 5-Nearest Neighbor | 0.783 | 0.043 | 0.749 |

Table 6: Naïve Bayes & 5-Nearest Neighbor Classifiers with 50:50 Dataset Performance Metrics

Based on the information of the table above, it can be easily concluded that the naïve bayes model had a lower FPR rate when compared to the 5-nearest neighbors. This means that naïve bayes model classified correctly more data samples from the majority class when compare to its counterpart. Moreover, it can also be observed that the naïve bayes model had a higher FNR rate when compared to the 5-nearest neighbor. This means that 5-nearest neighbor classified correctly more data samples from the minority class when compared to its counterpart. Furthermore, it can also be noted that the ROC value for the naïve bayes classifier was higher when compared to the 5-neares neighbor. Therefore, based on the ROC metric it can be concluded that naïve bayes had better performance results when compare to its counterpart. Nevertheless, it is essential to make mention that the 5-nearest neighbor model had a higher misclassification rate of the majority class and greatly favored the proper classification of the minority class. Hence, the 5-nearest neighbor is not a balance model nor robust as the majority class in the dataset contains an extremely high misclassification under this model. On the other hand, the naïve bayes model had a higher misclassification of the majority class, but performed better at classifying the minority class in the dataset. Therefore, it can be concluded that between the two models, naïve bayes has a more balance classification of both classes and hence is a more robust model when compare to its counterpart.

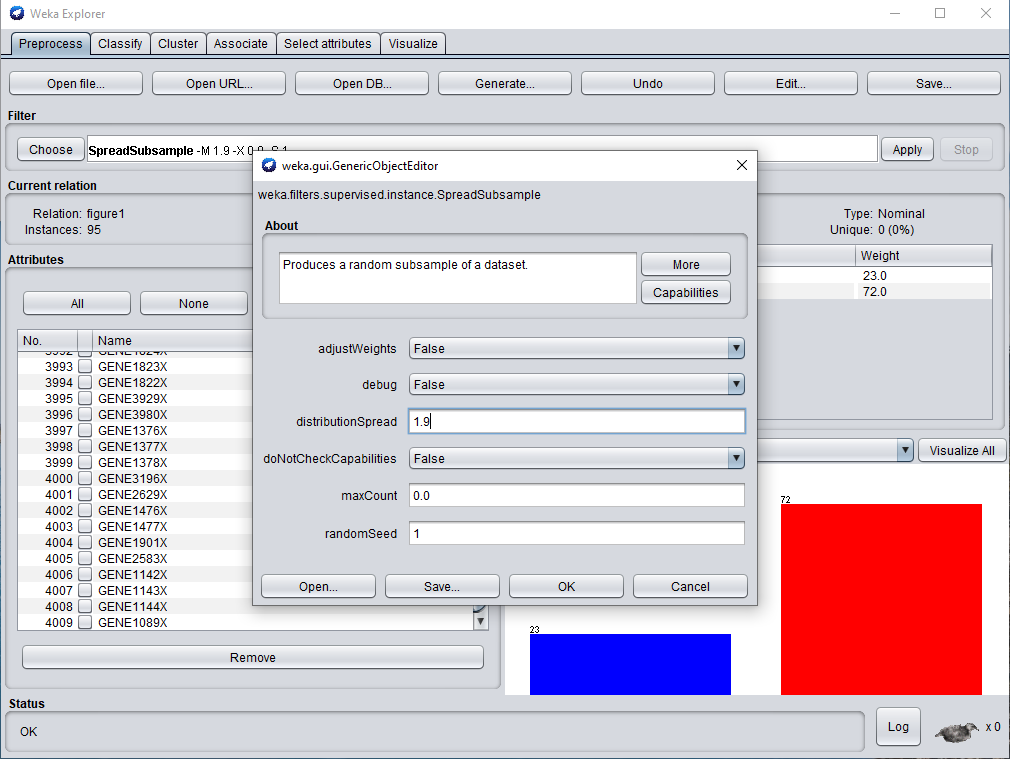
With this metrics in hand, it is now possible to compared the performance between the models that were trained with the full dataset and the 50:50 dataset. The summarized performance of all classifiers can be observed in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Dataset | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | Not Sampled | 0.153 | 0.217 | 0.844 |
| 5-Nearest Neighbor | Not Sampled | 0.097 | 0.391 | 0.863 |
| Naïve Bayes | 50:50 | 0.304 | 0.217 | 0.803 |
| 5-Nearest Neighbor | 50:50 | 0.783 | 0.043 | 0.749 |

Table 7: Naïve Bayes & 5-Nearest Neighbor Classifiers with 50:50 and Full Dataset Performance Metrics

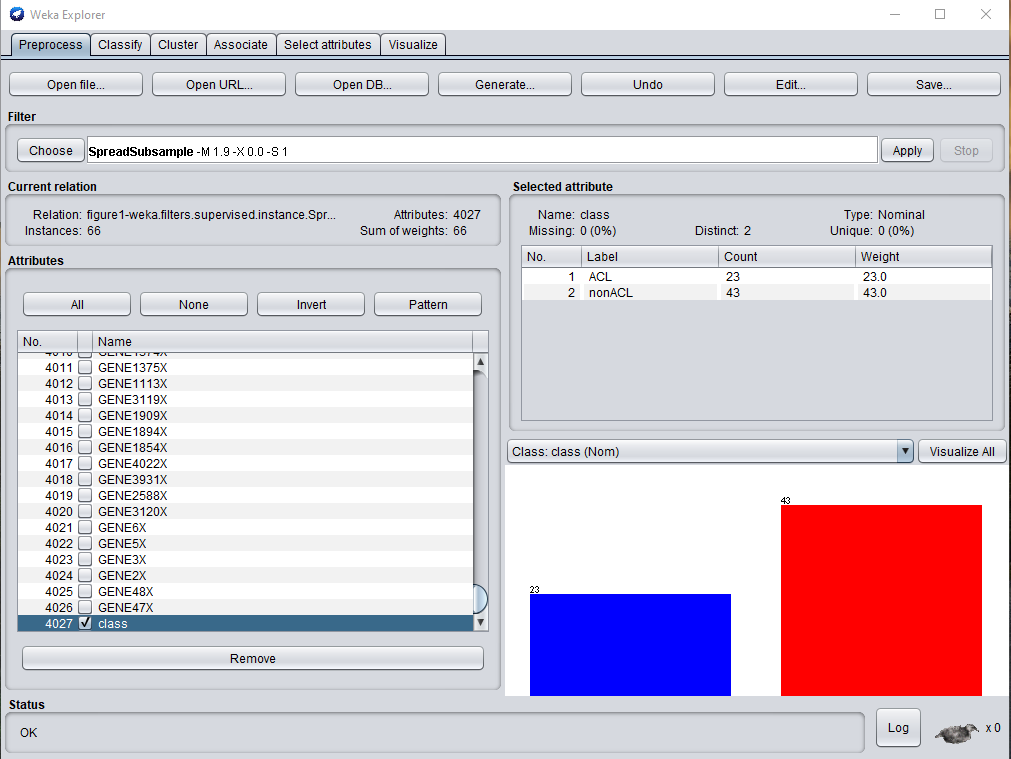
When comparing the naïve bayes models, it can be quickly determined that the FPR value greatly increased when training the model with the 50:50 dataset. The FNR value on the other hand, remained constant between the two naïve models while the ROC value decreased slightly in the model trained using the 50:50 split. Therefore, it can be concluded that the naïve bayes classifier trained with the 50:50 dataset decreased the proper classification of the majority (negative) class and this resulted in a slightly lower ROC value in the model. When comparing the 5-nearest neighbor models, it can be noted the FPR value increased immensely for the classifier trained with the 50:50 split. Additionally, the FNR metric was also affected drastically by greatly reducing the rate in the model trained with the 50:50 split. Due to these big changes in value from the FPR and FNR, the ROC metric of the model trained with the 50:50 was impacted heavily when compare to the model trained with the full dataset. From these results, it can be noted that when the models were trained using the full dataset, the end models performed better at classifying the majority class. Nevertheless, when the models were built with the 50:50 split, the classifiers greatly increased the misclassification of the majority class while keeping constant or greatly reducing the classification of the primary class. Nevertheless, in either instance, a full dataset or 50:50 split, the classifiers were always biased towards favoring the proper classification of one class over the other one. It is essential to note that based on the results from the table, the model which maintain the most balanced classification of both classes will be the naïve bayes classifier which was trained with the full dataset. Nonetheless, it is important to make mention as well that naïve bayes model with the full dataset slightly favors the proper classification of the majority class in the dataset.

The second split carried out to train the models will be the 65:35. In order to accomplish the 65:35 split, the “distributionSpread” value for the “SpreadSubsample” filter will be set to 1.9 in order to obtain the dataset with 23 samples representing the minority class while 43 samples will represent the majority class. A picture with the setting configured in Weka can be observed below:



Picture 7: Distribution Spread Configuration in Weka for 65:35 Split

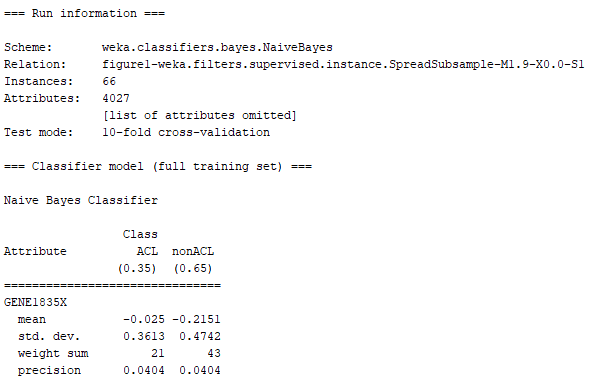
With the configuration completed and the filter applied to the dataset the final dataset with a 65:35 split after the random under-sampling technique has been applied can be observed below:

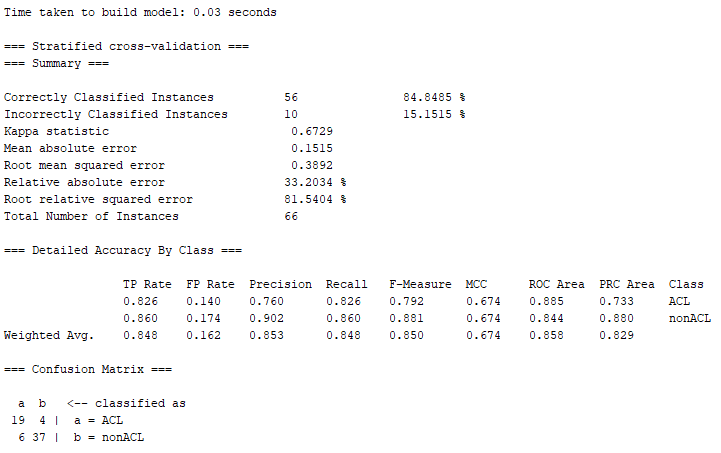


Picture 8: Dataset After Random Under-Sampling Technique

***Naïve Bayes 65:35 Dataset Training Results:***

The first model to be trained with the 65:35 split dataset will be the naïve bayes. The results from training the model with new dataset can be observed below:





Picture 9: Naïve bayes Classifier with 65:35 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 4 below:

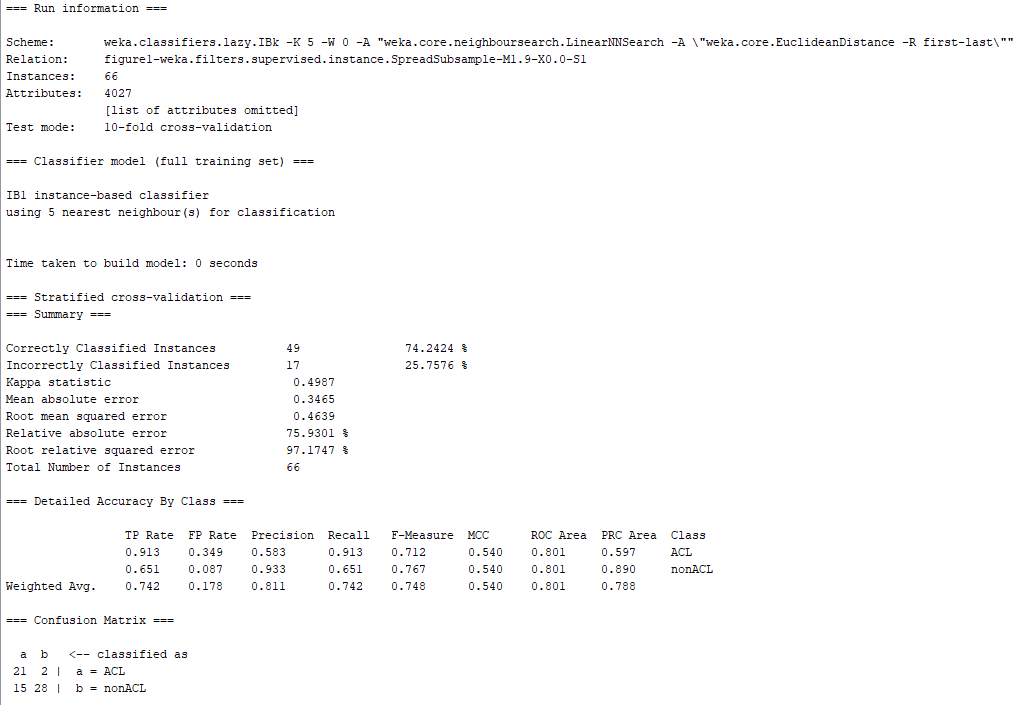
|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.140 | 0.174 | 0.858 |

Table 8: Naïve Bayes Classifier with 65:35 Dataset Performance Metrics

From table 8 above it can be noted that this classifier performed slightly better at classifying the secondary (negative) class when compared to the primary class since it had a lower false positive rate of 0.140, while the primary (positive) class had a higher misclassification rate 0.174. Based on this information, it can be concluded that the model slightly favors the proper classification of the majority class. This can be determined because the FPR value, which reflects the misclassification of the majority class, is decreased by 20% from the value of the FNR of 0.174. It is important to note that the overall performance of the classifier under the ROC value was not the best at 0.858. From these observations, it can be concluded that this classifier slightly favors the proper classification of the majority class of the dataset.

***5-Nearest Neighbor 65:35 Dataset Training Results:***

The next model to be trained with the 65:35 dataset is the 5-nearest neighbor. The results from training the classifier can be observed below:



Picture 10: 5-Nearest Neighbor Classifier with 65:35 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 9 below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| 5-Nearest Neighbor | 0.349 | 0.087 | 0.801 |

Table 9: 5-Nearest Neighbor Classifier with 65:35 Dataset Performance Metrics

From table 9 above it can be noted that this classifier performed better at classifying the primary (positive) class when compared to the secondary class since it had a lower false negative rate of 0.087, while the secondary (negative) class had a higher misclassification rate 0.349. Based on this information, it can be concluded that the model is biased and favors the proper classification of the minority class. This can be determined because the FPR value, which reflects the misclassification of the majority class, is increased by 75% from the value of the FNR of 0.087. It is important to note that the overall performance of the classifier under the ROC value was decent at only 0.801. Nevertheless, this model cannot be considered a balanced classifier since it greatly favors the proper classification of the minority class in the dataset.

With the training of models completed and noted, it is now possible to conduct an analysis on both classifiers’ performance metrics. The summarize performance metrics for both classifiers trained with the 65:35 dataset can be observed in the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | 0.140 | 0.174 | 0.858 |
| 5-Nearest Neighbor | 0.349 | 0.087 | 0.801 |

Table 10: Naïve Bayes & 5-Nearest Neighbor Classifiers with 65:35 Dataset Performance Metrics

Based on the information of the table above, it can be easily concluded that the naïve bayes model had a lower FPR rate when compared to the 5-nearest neighbors. This means that naïve bayes model classified correctly more data samples from the majority class when compare to its counterpart. Moreover, it can also be observed that the naïve bayes model had a higher FNR rate when compared to the 5-nearest neighbor. This means that 5-nearest neighbor classified correctly more data samples from the minority class when compared to its counterpart. Furthermore, it can also be noted that the ROC value for the naïve bayes classifier was higher when compared to the 5-neares neighbor. Therefore, based on the ROC metric it can be concluded that naïve bayes had better performance results when compare to its counterpart. Nevertheless, it is essential to make mention that the 5-nearest neighbor model had a higher misclassification rate of the majority class and greatly favored the proper classification of the minority class. Hence, the 5-nearest neighbor is not a balance model nor robust as the majority class in the dataset contains an extremely high misclassification under this model. On the other hand, the naïve bayes model had a higher misclassification of the minority class and performed better at classifying the majority class in the dataset. Therefore, it can be concluded that between the two models, naïve bayes has a more balance classification of both classes and hence is a more robust model when compare to its counterpart.

With this metrics in hand, it is now possible to compared the performance between the models that were trained with the full dataset and the 50:50 dataset. The summarized performance of all classifiers can be observed in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Dataset | False Positive Rate  (FPR) | False Negative Rate  (FNR) | Area Under ROC Curve  (ROC) |
| Naïve Bayes | Not Sampled | 0.153 | 0.217 | 0.844 |
| 5-Nearest Neighbor | Not Sampled | 0.097 | 0.391 | 0.863 |
| Naïve Bayes | 65:35 | 0.140 | 0.174 | 0.858 |
| 5-Nearest Neighbor | 65:35 | 0.349 | 0.087 | 0.801 |

Table 11: Naïve Bayes & 5-Nearest Neighbor Classifiers with 65:35 and Full Dataset Performance Metrics

When comparing the naïve bayes models, it can be quickly determined that the FPR value slightly decreased when training the model with the 65:35 dataset. Similarly, the FNR value also decreased slightly between the two naïve models while the ROC value increased slightly in the model trained using the 65:35 split. Therefore, it can be concluded that the naïve bayes classifier trained with the 65:35 dataset increased the proper classification of both classes and this resulted in a slightly higher ROC value in the model. When comparing the 5-nearest neighbor models, it can be noted the FPR value increased immensely for the classifier trained with the 65:35 split. Additionally, the FNR metric was also affected drastically by greatly reducing the rate in the model trained with the 65:35 split. Due to these big changes in value from the FPR and FNR, the ROC metric of the model trained with the 65:35 was impacted heavily when compare to the model trained with the full dataset by decreasing its value from 0.863 to 0.801 when the model was trained with the full dataset against the classifier with the 65:35 dataset. From these results, it can be noted that when the models were trained using the full dataset, the end models performed better at classifying the majority class. Nevertheless, when the models were built with the 65:35 split, the 5-nearest neighbor classifiers greatly increased the misclassification of the majority class while keeping greatly reducing the classification of the primary class. On the other hand, the naive bayes, slightly improved the classification of both classes with the 65:35 split dataset. It is essential to note that based on the results from the table, the model which maintain the most balanced classification of both classes was the naïve bayes classifier which was trained with the 65:35 dataset. This is because this model slightly decreased the misclassification of both classes, improving the overall classification of the dataset.

Based on these experiments, it can be concluded that the model which delivered the most balanced, less biased, and more robust classifier was the naïve bayes with 65:35 split. This is not only because it provided the highest ROC value between all models, but also because it performed the best at maintaining a balance proper classification of the two classes. Additionally, it can also be determined that the application of under-sampling technique to an unbalance dataset can greatly improve the robustness of a model by delivering a more balance classifier through its training phase. Moreover, based on the ratio split of the dataset when an under-sampling technique is applied, a model can be trained to favor the proper classification of one class over the other one. It is essential to note as well that between the two models used in the assignment, the 5-nearest neighbor model was able to deliver the most biased training of the model of the minority class when a trained with the under-sampling technique regardless of the split.