**Data Mining**

Mohammad Khattab

1

We have a scoring data set with 10 Attributes , taking in consideration that the class label is Payment Method which has four Categories :

1. Bank Transfer
2. Website Account
3. Credit Card
4. Monthly Billing

And we want to apply this dataset into 4 classification methods :

* 1. Naïve Bayes
  2. Rule-Based
  3. Tree Decision
  4. Lazy

**Note : All attributes are Imbalance.**

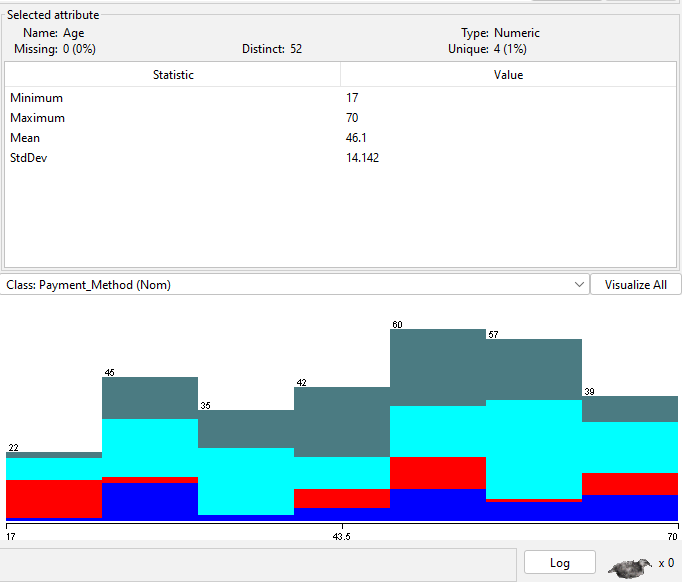
2

Graphical user interface

Description automatically generated with medium confidence

Figure (a) shows gender visualization.

**We noticed that Gender values are closer to be balance but at the end there aren’t equalized so they are not balance.**

****

Figure(b) shows age visualization .

**We notice the imbalance between age values clearly.**

3

1.NaiveBayes

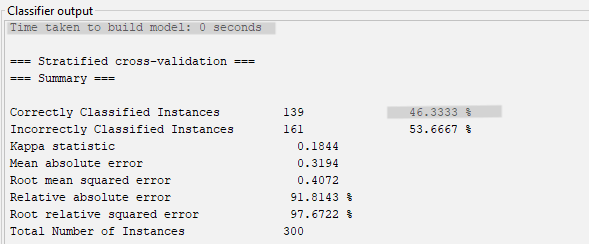


Figure (1.1) shows time needed to construct naïve model and accuracy with 10-Folds Cross-Validation.

**We noticed that 10-Folds Cross-Validation construct model with 0 sec which is error in the dataset maybe ,and with low accuracy estimated to 46.3333% and that’s mean naïve is not a good model for this dataset .**

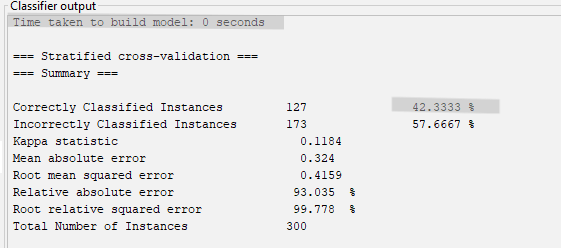


Figure (1.2) shows time needed to construct naïve model and accuracy with 2-Folds Cross-Validation.

**We noticed that 2-Folds Cross-Validation construct model with 0 sec which is error in the dataset ,and with lower accuracy than 10-Folds estimated to 42.3333% which means it’s also not recommended to be used in this dataset.**

**So, we noticed that naïve bayes is not a good model for building this data set.**

4

**Graphical user interface

Description automatically generated with medium confidence**

Figure(1.3) shows weighted average to the naïve model with 10-Folds Cross-Validation.

TP Rate = 0.463 and that’s mean the model predict 0.463 with certain class label and it predict it true.

FP Rate = 0.282 and that’s mean the model predict 0.282 with certain class label and it predict it false.

**And that’s lead us with the low prediction accuracy that naïve model has.**

Precision calculated by the equation below :

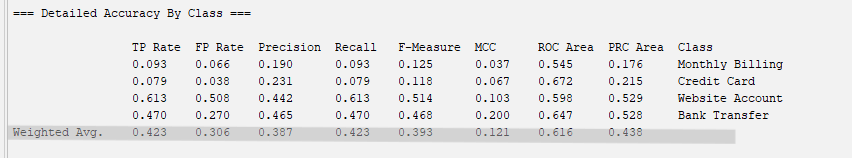
Precision = TP/(TP+FP) 🡺 0.463 / (0.463 + 0.282 ) = 0.453 (**low precision**).

Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.463 / (0.463 + 0 ) 🡪 0.463

F-Measure calculated by the equation below :

F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.453\*0.463)/(0.453+0.463) = 0.438



Figure(1.4) shows weighted average to the naïve model with 2-Folds Cross-Validation.

TP Rate = 0.423 and that’s mean the model predict 0.423 with certain class label and it predict it true.

FP Rate = 0.306 and that’s mean the model predict 0.306 with certain class label and it predict it false.

Precision calculated by the equation below :

Precision = TP/(TP+FP) 🡺 0.423 / (0.423 + 0.306 ) = 0.387 (**low precision**).

Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.423 / (0.423 + 0 ) 🡪 0.423

F-Measure calculated by the equation below :

F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.387\*0.423)/(0.387+0.423) = 0.393

5

**Then 2-Folds Cross-Validation has less accuracy than 10-Folds Cross-Validation and less Precision , Recall , F-Measure .**

Confusion Matrix for naïve with 10-Folds Cross-Validation Extract from Weka :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |
| TP | 5 | 3 | 18 | 17 |
| FN | 0 | 7 | 17 | 14 |
| FP | 3 | 7 | 73 | 36 |
| TN | 3 | 5 | 38 | 53 |

**We always try to reduce values of FN , FP because these values are the ones who the model predict them wrong , so we notice that prediction goes wrong extremmly .**

**So, we infer the low accuracy with naive 10-Folds Cross-Validtaion.**

Confusion Matrix for naïve with 2-Folds Cross-Validation Extract from Weka:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |
| TP | 4 | 2 | 26 | 11 |
| FN | 2 | 3 | 22 | 11 |
| FP | 10 | 4 | 73 | 32 |
| TN | 5 | 4 | 44 | 47 |

**In average 2-Folds Cross-Validation less accuracy than naïve 10-Folds Cross-Validation in predict class labels.**

6



Figure(1.5) shows naïve 10-Folds Cross-Validation visualization.



Figure(1.6) shows naïve 2-Folds Cross-Validation visualization.

7

2. Rule-Based

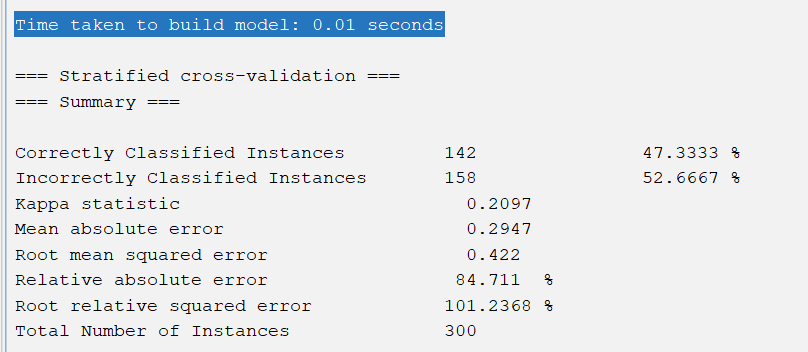


Figure (2.1) shows time needed to construct Rule model and accuracy with 10-Folds Cross-Validation.

**We noticed that 10-Folds Cross-Validation construct model with 0.01 sec which is really good and fast ,but with low accuracy estimated to 47.3333% and that’s mean rules is not a good model for this dataset .**

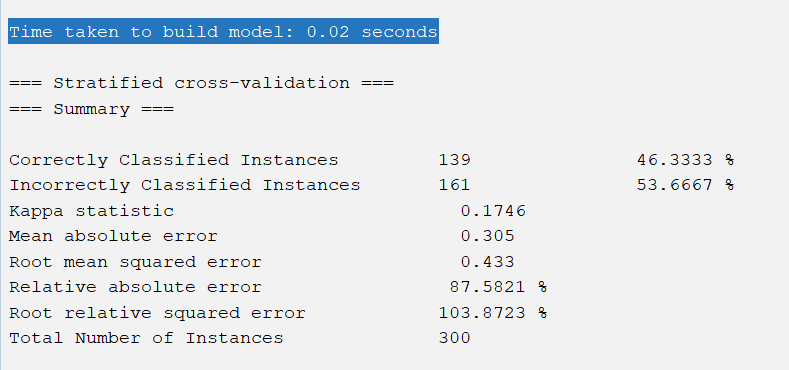
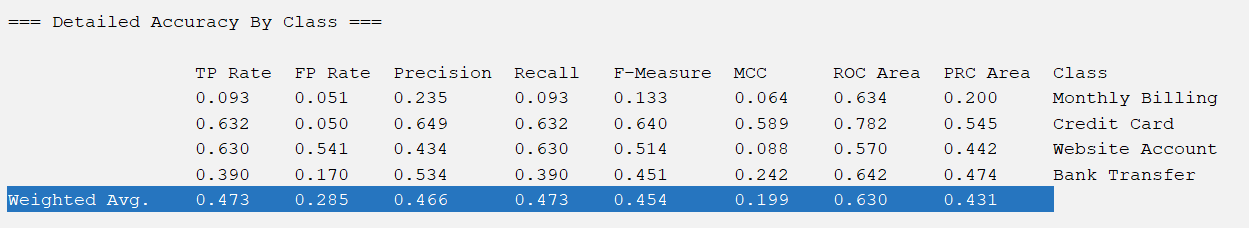
****

Figure (2.2) shows time needed to construct Rules model and accuracy with 2-Folds Cross-Validation.

8

**We noticed that 2-Folds Cross-Validation construct model with 0.02 sec which is really good and fast ,but with lower accuracy than 2-Folds estimated to 46.3333% which means it’s also not recommended to be used in this dataset.**

**So, we noticed that Rule-Based is not a good model for building this data set.**

****

Figure(2.3) shows weighted average to the Rule-Based model with 10-Folds Cross-Validation.

TP Rate = 0.473 and that’s mean the model predict 0.473 with certain class label and it predict it true.

FP Rate = 0.285 and that’s mean the model predict 0.285 with certain class label and it predict it false.

**And that’s lead us with the low prediction accuracy that rule model has.**

Precision calculated by the equation below :

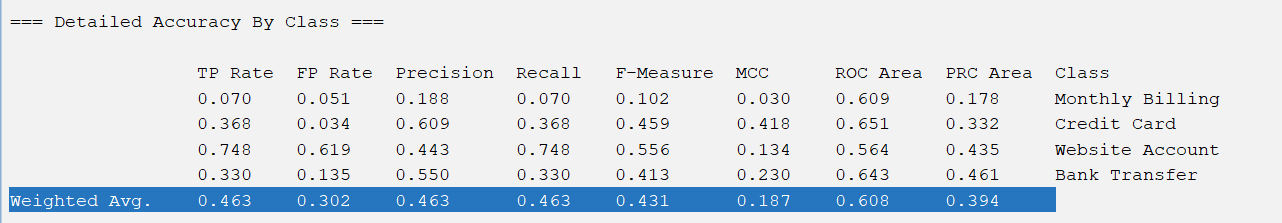
Precision = TP/(TP+FP) 🡺 0.473 / (0.473 + 0.285 ) = 0.466 (**low precision**).

Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.473 / (0.473 + 0 ) 🡪 0.473

F-Measure calculated by the equation below :

F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.466\*0.473)/(0.466+0.473) = 0.454



Figure(2.4) shows weighted average to the Rule-Based model with 2-Folds Cross-Validation.

9

TP Rate = 0.463 and that’s mean the model predict 0.463 with certain class label and it predict it true.

FP Rate = 0.302 and that’s mean the model predict 0.302 with certain class label and id predict it false.

Precision calculated by the equation below :

Precision = TP/(TP+FP) 🡺 0.463 / (0.463 + 0.302 ) = 0.463 (**low precision**).

Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.463 / (0.463 + 0 ) 🡪 0.463

F-Measure calculated by the equation below :

F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.463\*0.463)/(0.463+0.463) = 0.431

**Then Rule-Based 2-Folds Cross-Validation has less accuracy than 10-Folds Cross-Validation and less Precision , Recall , F-Measure .**

Confusion Matrix for Rule-Based 10-Folds Cross-Validation Extract from Weka :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |
| TP | 4 | 1 | 34 | 4 |
| FN | 1 | 24 | 10 | 3 |
| FP | 10 | 7 | 75 | 24 |
| TN | 2 | 5 | 54 | 39 |

**We always try to reduce values of FN , FP because these values are the ones who the model predict them wrong , so we notice that prediction goes wrong extremmly .**

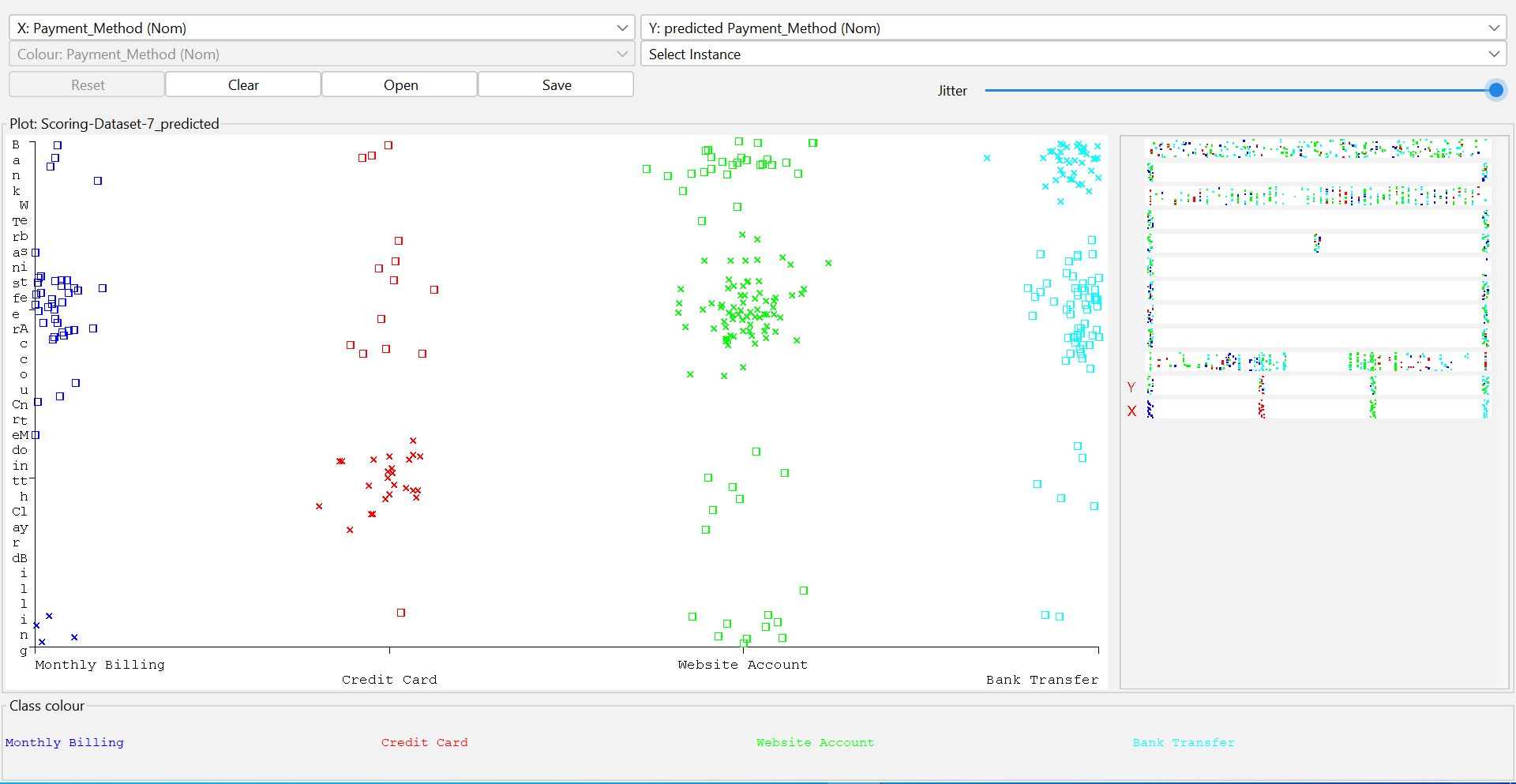
**So, we infer the low accuracy with rules 10-Folds Cross-Validtaion.**

Confusion Matrix for rules with 2-Folds Cross-Validation Extract from Weka:

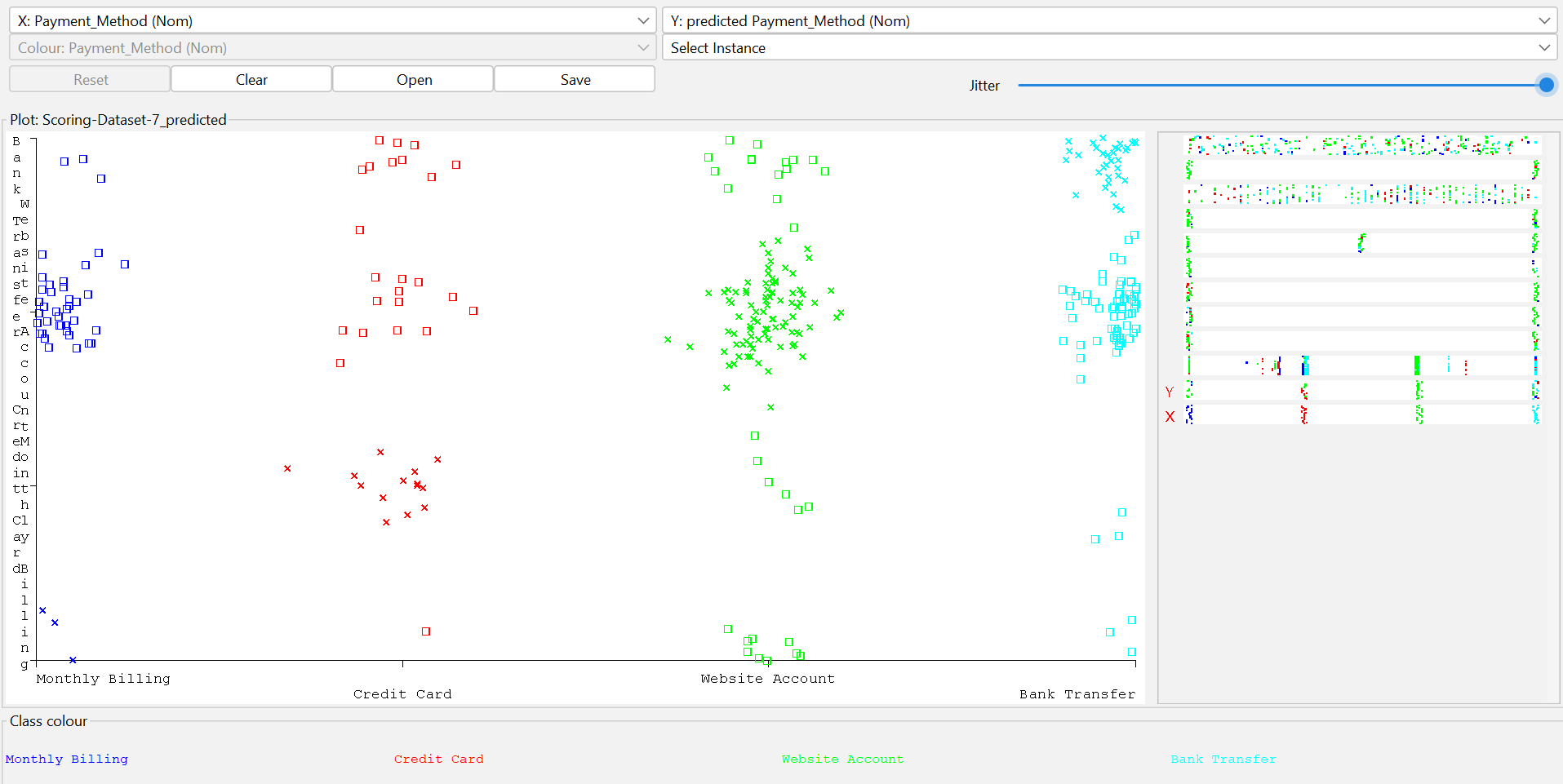
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |
| TP | 3 | 0 | 37 | 3 |
| FN | 1 | 14 | 14 | 9 |
| FP | 9 | 6 | 89 | 15 |
| TN | 3 | 3 | 61 | 33 |

10

**In average Rule-Based 2-Folds Cross-Validation has less accuracy than Rule 10-Folds Cross-Validation in predict class labels.**



Figure(2.5) shows Rule 10-Folds Cross-Validation visualization.

****

Figure(2.6) shows Rule 2-Folds Cross-Validation visualization.

11

**3.** Tree Decision

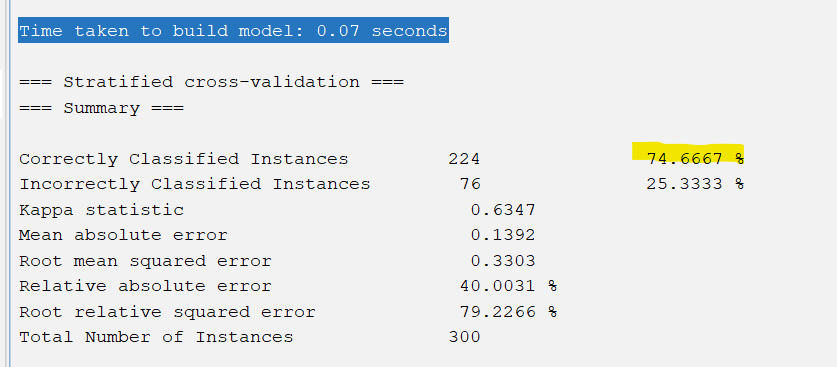
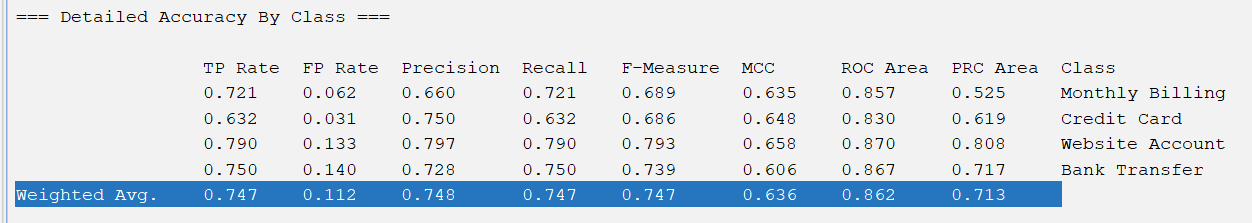


Figure (3.1) shows time needed to construct tree model and accuracy with 10-Folds Cross-Validation.

**We noticed that Tree 10-Folds Cross-Validation construct model with 0.07 sec which is not good and slow ,but with high accuracy estimated to 74.6667% and that’s mean Tree Decision is a good model for this dataset if the time is not important.**

****

Figure(3.2) shows weighted average to the Tree model with 10-Folds Cross-Validation.

TP Rate = 0.747 and that’s mean the model predict 0.747 with certain class label and it predict it true.

FP Rate = 0.112 and that’s mean the model predict 0.112 with certain class label and it predict it false.

**And that’s lead us with the high prediction accuracy that tree model has.**

12

Precision calculated by the equation below :

Precision = TP/(TP+FP) 🡺 0.747 / (0.747 + 0.112 ) = 0.748 (**high precision**).

Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.747 / (0.747 + 0 ) 🡪 0.747

F-Measure calculated by the equation below :

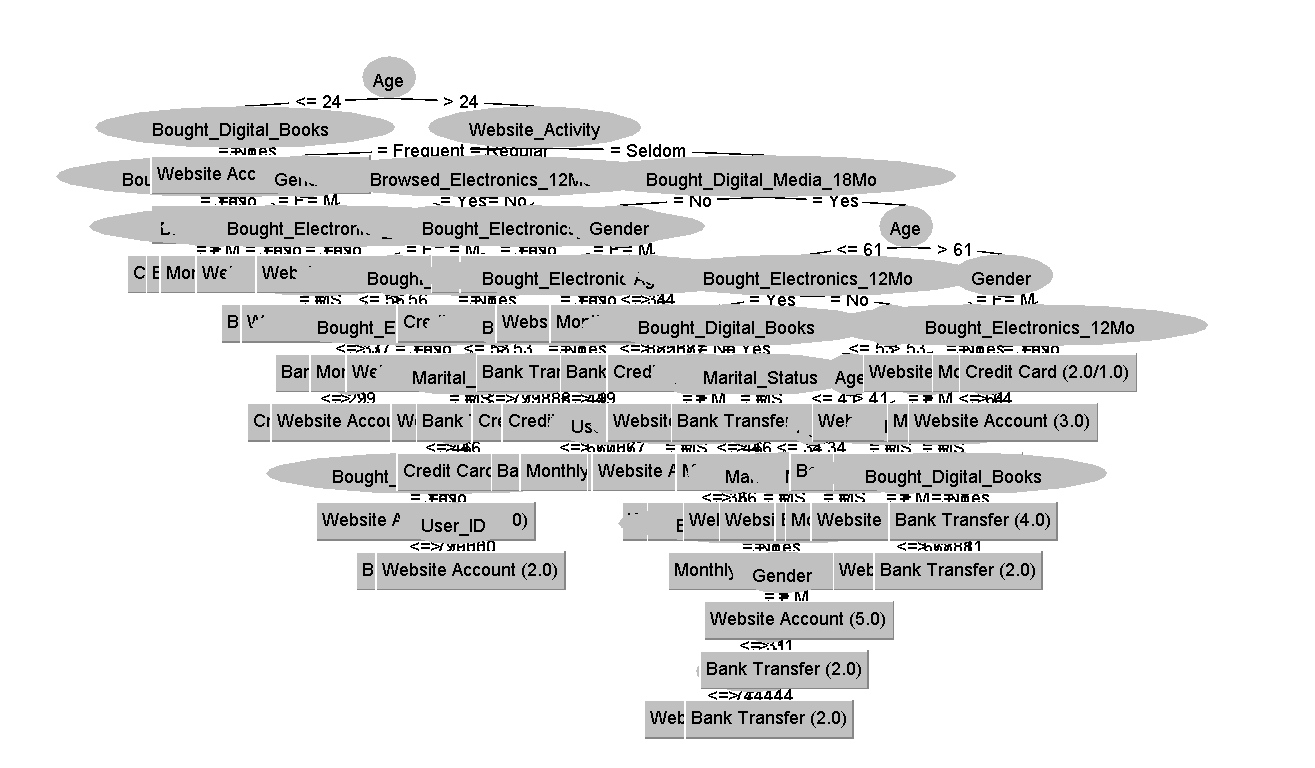
F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.748\*0.747)/(0.748+0.747) = 0.747

Confusion Matrix for Tree 10-Folds Cross-Validation Extract from Weka :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |
| TP | 31 | 0 | 6 | 6 |
| FN | 5 | 24 | 3 | 6 |
| FP | 7 | 2 | 94 | 16 |
| TN | 4 | 6 | 15 | 75 |

**We always try to reduce values of FN , FP because these values are the ones who the model predict them wrong , so we notice that prediction goes wrong extremmly .**

13

****

Figure(2.3) shows Tree 10-Folds Cross-Validation visualization.

4-lazy

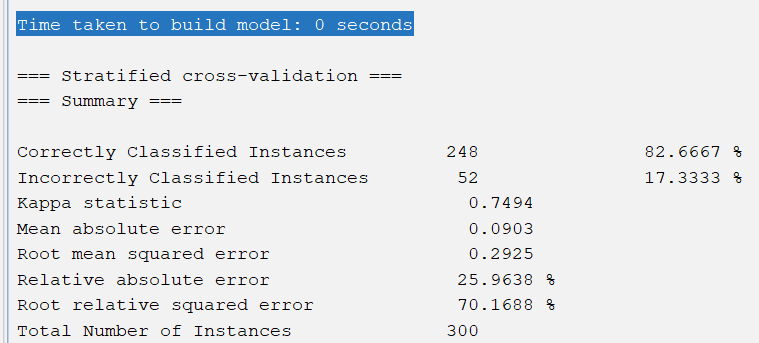


Figure (4.1) shows time needed to construct lazy model and accuracy with 10-Folds Cross-Validation.

14

**Time Taken to build model will be always 0 in any lazy k-Fold Cross-Validation because this model doesn’t construct model , and with the most accuracy estimated to 82.6667% and that’s mean lazy is the best model for this dataset .**

Graphical user interface, text, application

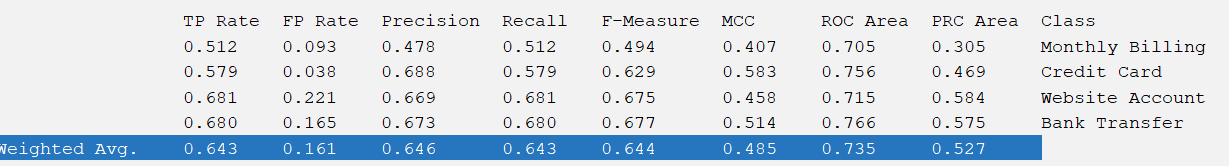
Description automatically generated

Figure (4.2) shows time needed to construct lazy model and accuracy with 6-Folds Cross-Validation.

**6-Folds has also high accuracy estimated to 80% ,but less accuracy than lazy 10-folds .**

15

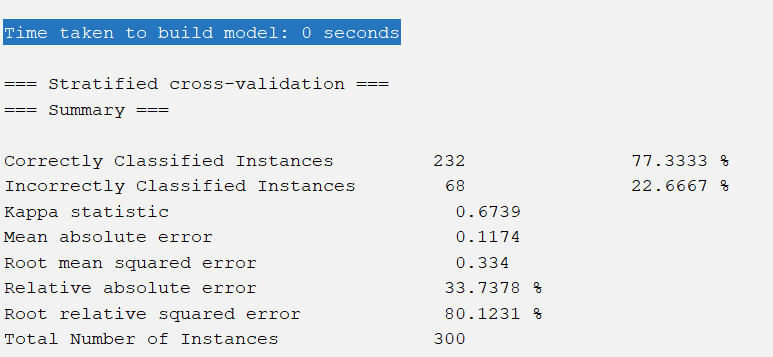


Figure (4.3) shows time needed to construct lazy model and accuracy with 4-Folds Cross-Validation.

**4-Folds has also high accuracy estimated to 77.3333% but less accuracy than lazy 10-folds and 6-folds .**

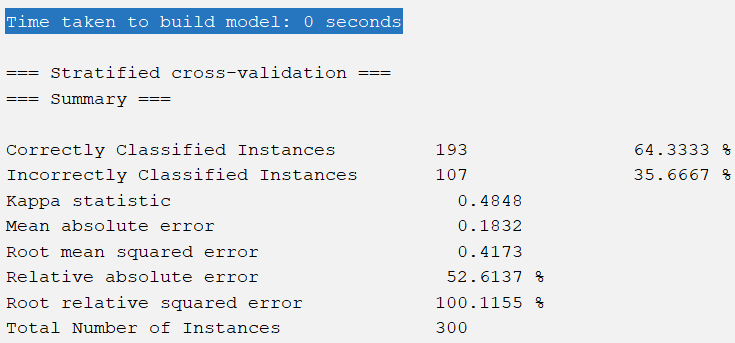
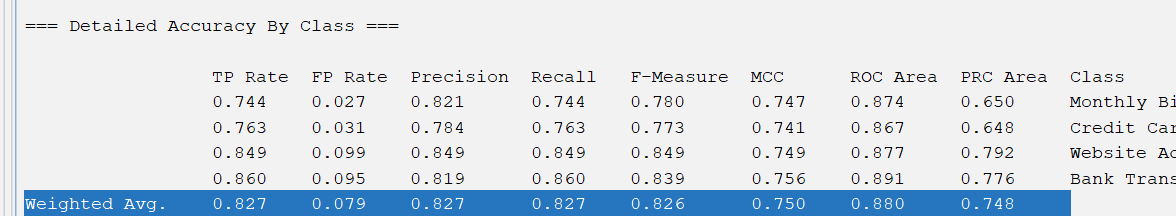


Figure (4.4) shows time needed to construct lazy model and accuracy with 2-Folds Cross-Validation.

**2-Folds has accuracy estimated to 64.3333 ,but less accuracy than lazy 10-folds and 6-folds and 4-folds.**

**So, we noticed that lazy 10-Folds is the best model for this data set.**

16



Figure(4.5) shows weighted average to the lazy model with 10-Folds Cross-Validation.

TP Rate = 0.827 and that’s mean the model predict 0.827 with certain class label and it predict it true.

FP Rate = 0.079 and that’s mean the model predict 0.079 with certain class label and it predict it false.

**And that’s lead us with the high prediction accuracy that lazy model has.**

Precision calculated by the equation below :

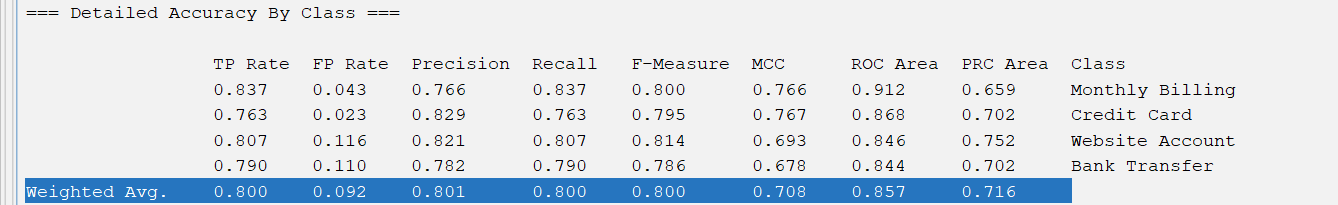
Precision = TP/(TP+FP) 🡺 0.827 / (0.827 + 0.079 ) = 0.906 (**high precision**).

Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.827 / (0.827 + 0 ) 🡪 0.827

F-Measure calculated by the equation below :

F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.906\*0.827)/(0.906+0.827) = 0.864



Figure(4.6) shows weighted average to the lazy model with 6-Folds Cross-Validation.

TP Rate = 0.800 and that’s mean the model predict 0.800 with certain class label and it predict it true.

FP Rate = 0.092 and that’s mean the model predict 0.092 with certain class label and it predict it false.

**And that’s lead us with the high prediction accuracy that lazy model has.**

Precision calculated by the equation below :

Precision = TP/(TP+FP) 🡺 0.800 / (0.800 + 0.092 ) = 0.892 (**high precision**).

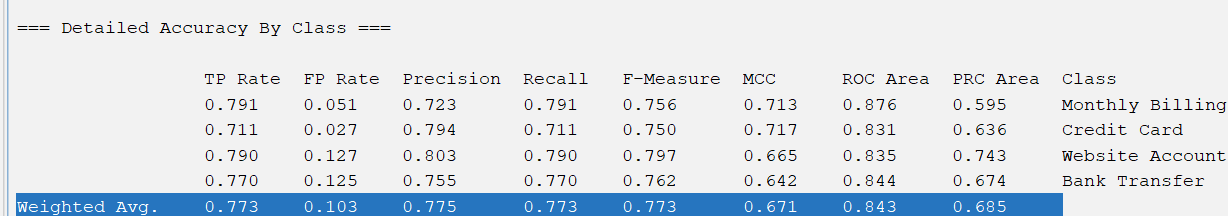
Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.800 / (0.800 + 0 ) 🡪 0.800

F-Measure calculated by the equation below :

F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.892\*0.800)/(0.892+0.800) = 0.843

17



Figure(4.7) shows weighted average to the lazy model with 4-Folds Cross-Validation.

TP Rate = 0.773 and that’s mean the model predict 0.773 with certain class label and it predict it true.

FP Rate = 0.103 and that’s mean the model predict 0.103 with certain class label and it predict it false.

Precision calculated by the equation below :

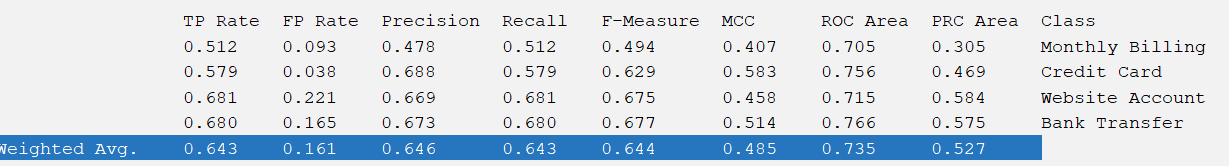
Precision = TP/(TP+FP) 🡺 0.773 / (0.773 + 0.103 ) = 0.876 (**high precision**).

Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.773 / (0.773 + 0 ) 🡪 0.773

F-Measure calculated by the equation below :

F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.876\*0.773)/(0.876+0.773) = 0.821



Figure(4.8) shows weighted average to the lazy model with 2-Folds Cross-Validation.

TP Rate = 0.643 and that’s mean the model predict 0.643 with certain class label and it predict it true.

FP Rate = 0.161 and that’s mean the model predict 0.161 with certain class label and it predict it false.

Precision calculated by the equation below :

Precision = TP/(TP+FP) 🡺 0.643 / (0.643 + 0.161 ) = 0.804 (**high precision**).

Recall calculated by the equation bellow :

Recall = TP / (TP + FN) 🡪 0.643 / (0.643 + 0 ) 🡪 0.643

F-Measure calculated by the equation below :

F-Measure = (2 \* precision \* recall ) / (precision + recall ) 🡪 (2\*0.804\*0.643)/(0.804+0.643) = 0.714

18

**Then 2-Folds Cross-Validation has less accuracy than 4-folds,6-folds and 10-Folds Cross-Validation and less Precision , Recall , F-Measure .**

Confusion Matrix for lazy with 10-Folds Cross-Validation Extract from Weka :

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |  |  |  |  |  |
| TP | **32** | **4** | **3** | **4** |  |  |  |  |  |
| FN | **5** | **29** | **3** | **1** |  |  |  |  |  |
| FP | **1** | **3** | **101** | **14** |  |  |  |  |  |
| TN | **1** | **1** | **12** | **86** |  |  |  |  |  |

**We always try to reduce values of FN , FP because these values are the ones who the model predict them wrong , so we notice that prediction goes right .**

**So, we infer the high accuracy with lazy 10-Folds Cross-Validtaion.**

Confusion Matrix for lazy with 6-Folds Cross-Validation Extract from Weka :

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |  |  |  |  |  |
| TP | **36** | **1** | **3** | **3** |  |  |  |  |  |
| FN | **5** | **29** | **3** | **1** |  |  |  |  |  |
| FP | **1** | **4** | **96** | **18** |  |  |  |  |  |
| TN | **5** | **1** | **15** | **79** |  |  |  |  |  |

**We always try to reduce values of FN , FP because these values are the ones who the model predict them wrong , so we notice that prediction goes right relativly .**

19

Confusion Matrix for lazy with 4-Folds Cross-Validation Extract from Weka :

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |  |  |  |  |  |
| TP | **34** | **1** | **3** | 5 |  |  |  |  |  |
| FN | **5** | **27** | 4 | 2 |  |  |  |  |  |
| FP | 2 | 5 | **94** | **18** |  |  |  |  |  |
| TN | 6 | **1** | **16** | **77** |  |  |  |  |  |

**We always try to reduce values of FN , FP because these values are the ones who the model predict them wrong , so we notice that prediction goes right .**

Confusion Matrix for lazy with 2-Folds Cross-Validation Extract from Weka :

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class Label | Monthly Billing | Credit Card | Website Acc | Bank Transfer |  |  |  |  |  |
| TP | 22 | 3 | 10 | 8 |  |  |  |  |  |
| FN | 4 | **22** | 9 | 3 |  |  |  |  |  |
| FP | 11 | 5 | 81 | 22 |  |  |  |  |  |
| TN | 9 | 2 | 21 | 68 |  |  |  |  |  |

**We always try to reduce values of FN , FP because these values are the ones who the model predict them wrong , so we notice that prediction goes wrong relativly .**

**So, we infer the Immediate accuracy with lazy 4-Folds Cross-Validtaion.**

**In average 2-Folds Cross-Validation less accuracy than lazy 4-folds,6-folds and 10-Folds Cross-Validation in predict class labels.**

**Lazy-10Folds > 6-Folds > 4-Folds >2-Folds .**

20



Figure(4.9) shows lazy 10-Folds Cross-Validation visualization.



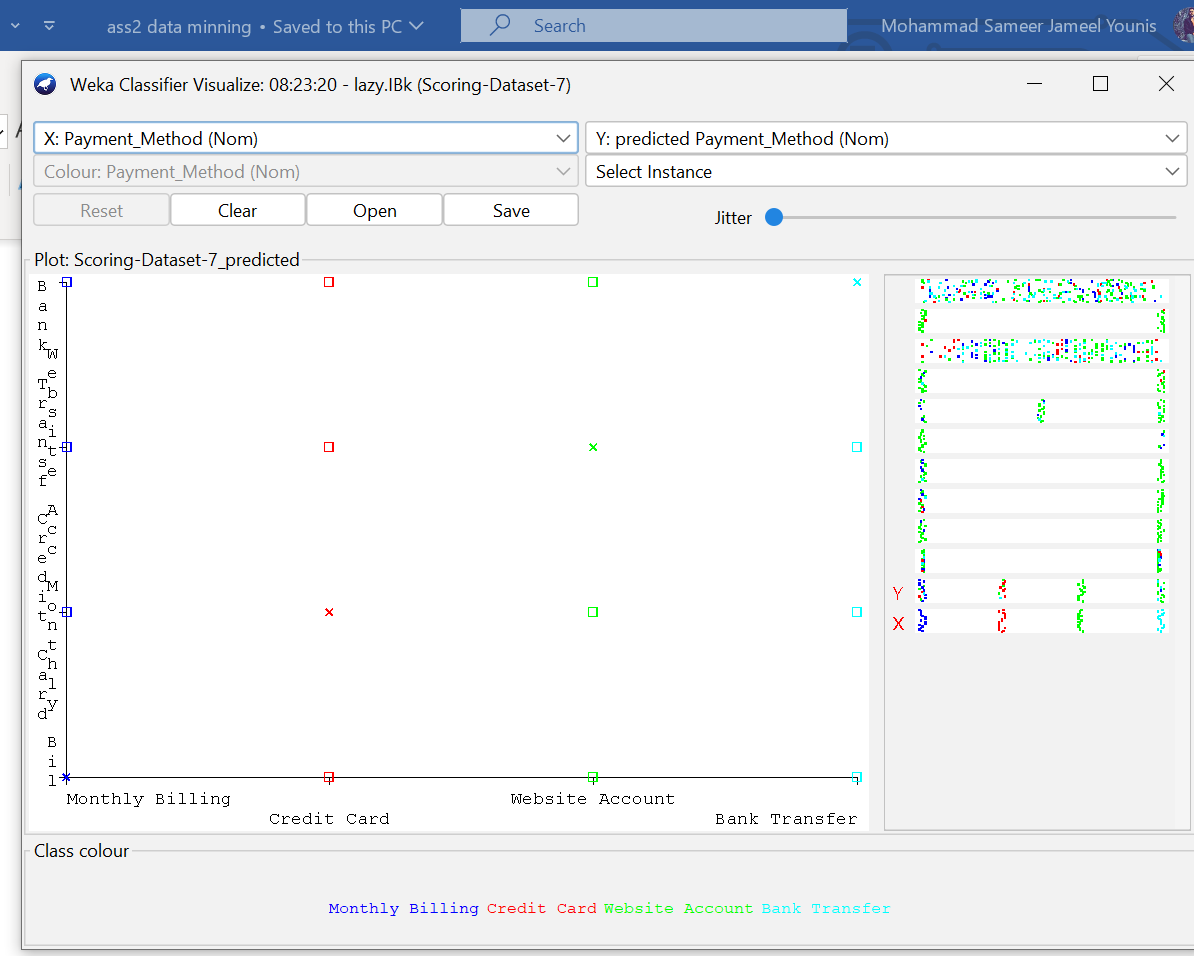
Figure(4.10) shows lazy 6-Folds Cross-Validation visualization.

21

Graphical user interface, chart

Description automatically generated

Figure(4.11) shows lazy 4-Folds Cross-Validation visualization.



Figure(4.12) shows lazy 2-Folds Cross-Validation visualization.

22

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Classifier | Accuracy | TP Rate | FP Rate | Precision | Recall | F-Measure | Class Label |
| NaiveBayes | 46% | 0.463 | 0.282 | 0.453 | 0.463 | 0.438 | Payment Method |
| Rule-Based | 47% | 0.473 | 0.285 | 0.466 | 0.473 | 0.454 | Payment Method |
| Tree | 74% | 0.747 | 0.112 | 0.748 | 0.747 | 0.747 | Payment Method |
| Lazy | 83% | 0.827 | 0.079 | 0.827 | 0.827 | 0.826 | Payment Method |

The above table shows comparing of the four classification methods using 10-Folds Cross-Validation.

**We noticed that naïve bayes classifier has the least Accuracy , TP rate , Precisions , Recall , F-Measure.**

**We noticed that Lazy classifier has the Most Accuracy , TP rate , Precisions , Recall , F-Measure.**

**Lazy > Tree Decision > Rule-Based > Naïve Bayes.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Naïve Bayes** | **Rule-Based** | **Tree Decision** | **Lazy** |
| **Time to construct model** | **0sec** | **0.1sec** | **0.7sec** | **0sec** |

The above table shows time to construct model to each of classifiers.

**We noticed that Lazy has 0 sec time to construct model and that’s because it doesn’t construct model .**

**And we noticed that the slowest classifier to build model is Tree decision with 0.07sec.**

**Lazy , Naïve Bayes > Rule-Based > Tree Decision.**

23