

**ISM 6136 – Datamining/Predictive Analytics**

**Class Assignment 4**

**5 points**

**TASK: Performing classification using Naïve Bayes Classifier in XLMiner**

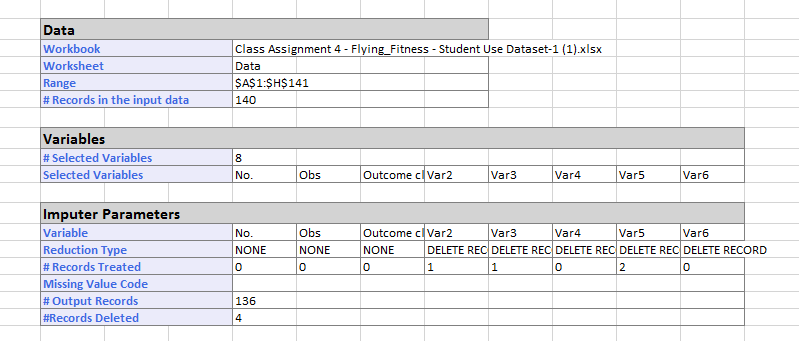
**Part 1 – Answers highlighted in yellow**

1. Dataset provided has test results on the ﬂying ﬁtness tests for 40 pilots. There are six other categorical variables (var2 - var 6) indicative of the

performance of the pilots on various physical and psychological tests. Predict the outcome class of the pilots in the ‘new data’ sheet of the dataset using the Naïve Bayes Classifier.

1. Follow the datamining steps below:
2. Understand the problem and purpose of data mining task
3. Obtain the dataset for analysis
4. Explore, clean and preprocess data
   1. Cleanup any column that is not a predictor – The first column (No.) and second column (Obs.) are deleted as they are not useful in our prediction
   2. Perform ‘Missing Data Handling’ *(Delete any record that is corrupt)*

As seen in the snapshot below, four records were deleted as they were corrupt.



1. Reduction of data dimension (if needed)

Not required in this case.

1. Partition data

Two partitions were done – 60/40 and 75/25

1. Choose the data mining techniques/algorithms – Classify, Naïve Bayes and build **at least 2 models**

Data mining technique- Supervised Classification- Naives Bayyes

Model #1- 60/40

Model #2- 75/25

1. Interpret the results and depending on the model selection criteria choose the **best model**

|  |  |  |
| --- | --- | --- |
| **Criteria & Explanation** | **Model #1** | **Model #2** |
| % error – Lower the better.  Training Vs Validation within models – Not great difference.  Validation Vs Validation between model #1 and #2 – **Model 1 has lower error rate and hence better** | Training – 15.85%  Validation – 12.96% | Training – 13.73%  Validation – 17.65% |
| % Accuracy – Higher the better.  Training Vs Validation within models – Not great difference.  Validation Vs Validation between model #1 and #2 – **Model 1 has higher accuracy and hence better** | Training – 84.14%  Validation – 87.04% | Training – 86.27%  Validation – 82.35% |
| Precision – higher the better.  Training Vs Validation within models – Not great difference.  Validation Vs Validation between model #1 and #2 – **Both are on tie – No decision for this criterion.** | Training – 0.95  Validation – 1 | Training – 0.96  Validation – 1 |
| Sensitivity - higher the better.  Training Vs Validation within models – Not great difference.  Validation Vs Validation between model #1 and #2 – **Model 1 has higher sensitivity and hence better** | Training – 0.61  Validation – 0.5 | Training – 0.64  Validation – 0.33 |
| F1- higher the better.  Training Vs Validation within models – Not great difference.  Validation Vs Validation between model #1 and #2 – **Model 1 has higher F1 and hence better** | Training – 0.75  Validation – 0.67 | Training – 0.77  Validation – 0.5 |
| Lift Chart (Validation)– greater the area between lift curve and baseline, better the model.  Based on the scale, Model 1 shows larger area, hence model 1 is better. |  |  |
| ROC (Validation) – Closer the curve is to the top-left corner, better the model.  Model#2 is showing better performance |  |  |
| AUC (Validation) – higher the better. Model #2 is winning here. | 0.94 | 0.97 |

The best model selected based on the above is Model #1 based on the lower overall error rate. Only ROC and AUC validation suggests model #2 is better. Based on my understanding from <https://www.researchgate.net/post/Understanding_AUC_of_ROC_sensitivity_and_specificity_values>, ROC and AUC only suggest that a model is performing better on average, which is why we should look at other factors, like F1 and sensitivity to choose the best model. Hence I decided that Model #1 is better.

1. Deploy **best model** on the new data and paste a screen shot of your prediction results

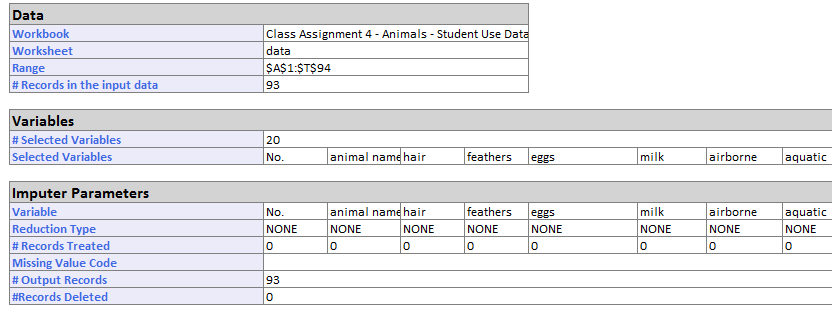
|  |  |  |  |
| --- | --- | --- | --- |
| **Record ID** | **Prediction: Outcome class type** | **PostProb: 0** | **PostProb: 1** |
| **Record 1** | 0 | 0.868580637 | 0.131419363 |
| **Record 2** | 0 | 1 | 0 |
| **Record 3** | 0 | 0.586398496 | 0.413601504 |
| **Record 4** | 0 | 1 | 0 |
| **Record 5** | 0 | 1 | 0 |
| **Record 6** | 0 | 0.586398496 | 0.413601504 |
| **Record 7** | 0 | 1 | 0 |
| **Record 8** | 0 | 0.530899993 | 0.469100007 |
| **Record 9** | 1 | 0 | 1 |
| **Record 10** | 1 | 0.27626 | 0.72374 |
| **Record 11** | 0 | 0.530899993 | 0.469100007 |
| **Record 12** | 1 | 0.147089429 | 0.852910571 |
| **Record 13** | 1 | 0 | 1 |
| **Record 14** | 0 | 0.840657458 | 0.159342542 |

1. Submit the **Excel workbook and this word document with explanation for steps c) through h).**

**Part 2**

1. Dataset provided has been taken from Zoo database that was created by Richard Forsyth in 1990 to illustrate its PC-Beagle program (rule finding program). It contains a list of animals in rows and their associated attributes described in 17 distinct qualitative variables (columns): hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs, tail, domestic, catsize. Based on these qualitative variables the animals are classified into outcome class type 1 through 7. Predict the class type of the animals in the ‘new data’ sheet using Naïve Bayes Classifier.
2. Follow the datamining steps below:
3. Understand the problem and purpose of data mining task
4. Obtain the dataset for analysis
5. Explore, clean and preprocess data
   1. Cleanup any column that is not a predictor – Column 1 (No.) and Column 2(Animal Name) are not predictors and hence deleted
   2. Perform ‘Missing Data Handling’ *(Delete any record that is corrupt)*

*No records deleted – no corrupt data- snapshot provided below.*



1. Reduction of data dimension (if needed)

Performed PCA for 10 dimensions for Model #2 and Model #3.

1. Partition data

Model #1(all dimensions)- 60/40

Model #2(10 dimensions)- 60/40

Model #3(10 dimensions)- 75/25

1. Choose the data mining techniques/algorithms – Classify, Naïve Bayes and build **at least 2 models**

Data mining technique- Supervised Classification- Naives Bayyes

3 models built are as follows:

Model #1(all dimensions)- 60/40

Model #2(10 dimensions)- 60/40

Model #3(10 dimensions)- 75/25

1. Interpret the results and depending on the model selection criteria choose the **best model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria & Explanation** | **Model #1** | **Model #2** | **Model #3** |
| % error – Lower the better.  Training Vs Validation within models – Not great difference.  Validation Vs Validation between model #1 and #2 – **Model 1 has lower error rate and hence better** | Training – 0%  Validation – 2.70% | Training – 7.14%  Validation – 10.81% | Training – 8.57%  Validation – 13.04% |
| % Accuracy – Higher the better.  Training Vs Validation within models – Not great difference.  Validation Vs Validation between model #1 and #2 – **Model 1 has higher accuracy and hence better** | Training – 100%  Validation – 97.3% | Training – 92.86%  Validation – 89.19% | Training – 91.43%  Validation – 86.96% |
| Diagonal – higher the better. A higher diagonal for the same partition indicates less errors.  Model 1 and 2 are of the same partition (60/40) and hence can be compared. **Model 1 indicates better accuracy.** | Training – 56  Validation – 36 | Training – 52  Validation – 33 | Training – 64  Validation – 20 |
| No-Diagonal – lower the better. A low diagonal for the same partition indicates less errors.  Model 1 and 2 are of the same partition (60/40) and hence can be compared. **Model 1 indicates better accuracy.** | Training – 0  Validation – 1 | Training – 4  Validation – 4 | Training – 6  Validation – 3 |
| Precision – higher the better. | Data not available | | |
| Sensitivity - higher the better. |
| F1- higher the better. |
| Lift Chart (Validation) | Non-binary- Not applicable | | |
| ROC (Validation) |
| AUC (Validation) |

**From the above results, it can be said that model 1 is our best model – highest accuracy, less errors.**

1. Deploy **best model** on the new data and present your prediction results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| leopard | **Record 1** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| chicken | **Record 2** | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| starfish | **Record 3** | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| penguin | **Record 4** | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| kiwi | **Record 5** | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| chub | **Record 6** | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| antelope | **Record 7** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| gnat | **Record 8** | 6 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

Unable to classify starfish because the algorithm did not find anything similar to the characteristics of a ‘starfish’. The rest were predicted accurately.

For example, penguin, chicken and kiwi are class 2 – bird

Gnat is an insect- 6

Antelope and leopard – 1 – mammal.

1. Submit the **Excel workbook and this word document with explanation for steps c) through h).**