DATA MINING ANALYSIS OF BIRDS BONES AND LIVING HABITATS

ICT616
Data Resource Management

A Report by –
Rajesh Jyothi - 33669079
Candida Jeremiah - 32865458

A Report on Bird and Bone Data Mining Analysis

Abstract

Data mining methods are largely implemented in a broad spectrum of fields: finance, banking, retail sales, manufacturing, health care and marketing for analysing available data and extracting information and knowledge to support decision-making. This report presents the results of the dataset on six ecological bird groups provided by Dr D. Liu for a data mining research project implemented at Beijing Museum of Natural History, revealing the high potential of data mining applications (Feyyad, 1996).

Introduction

Industries today operate in a very complex and highly competitive environment. The main challenge being analysing their performance, to identify their uniqueness and to build a strategy for further development and future actions. Large datasets are collected from different data generating source like electronic gadgets, mobile phones, shopping, transport etc. Lying within those data sets are patterns that are indicators of customers interests, habits and behaviours. Data mining is a process used by companies to locate and interpret those patterns and turn raw data into useful information to make better-informed decisions and to better serve customers. Feyyad, defined Data Mining as a process of identifying the hidden trends or patterns of interest or set of representations using classification or regression (prediction) models like Decision trees, Random Forest or KNN etc. (Feyyad, 1996). The objective of this report is to find the data models that are opting to the data set and provide better accuracy by using the available patterns in the data.

Data mining uses two techniques for analysing data, classification and prediction. In this report, we will discuss the classification technique using the 'Bird and Bone' dataset. There are many kinds of birds: pigeons, ducks, ostriches and penguins where some birds fly, and some don't, but instead, they run fast. Similarly, some swim underwater while others wade in a shallow pool. According to their living environments and living habits, birds are classified into different ecological groups. In the following report, we will discuss the implementation of data mining techniques and methods, mainly focusing on revealing the high potential of data mining application. The objective of this report is to conduct a project to find a data model that provides accurate results by using the available patterns in the data to predict the eight ecological groups based on the length and width of the bird's bone (Dr D. Liu, 2018).

The Research Methodology

The initiated data mining project is implemented following the CRISP-DMDM (Cross-Industry Standard Process for Data Mining) model, widely used by researchers in the field during the last ten years. It is a cyclic approach, including six main phases – Business Understanding, Data Understanding, Data Preparation, Data Modelling, Evaluation and Deployment. The data is processed in loops between the stages, resulting from the complex non-linear nature of the data mining process to achieve consistent and reliable results. The software tool used for this analysis is the open-source software, Rapid Miner, offering a range of classification methods for data mining.

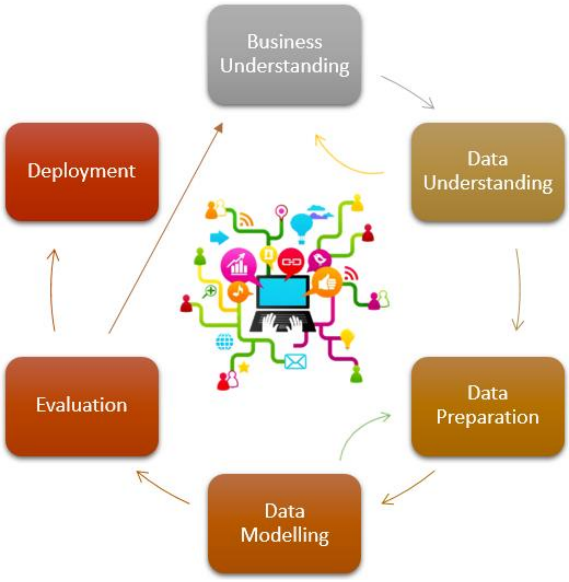


Figure 1: CRISP-DMDM for Data Mining process

Business Understanding

Business understanding is one of the critical phases of the Data mining Process. During the Business Understanding Phase, a review is performed to understand the Business requirements and its issues that have been dealt in the past by the application of data mining techniques and methods. Based on the outcomes of the performed research, the goal and objectives for the report are formulated. In this report, we are working on Birds bone dataset and trying to classify the birds to different ecological groups based on bone length and diameter.

Birds can be classified into eight ecological groups, based on their living habits and environment, but in this report based on data availability, we have focused on six ecological groups. They are:

- Swimming Birds (SW)
- Wading Birds (W)
- Terrestrial Birds (T)
- Raptors (R)
- Scansorial Birds (S)
- Singing Birds (SO)

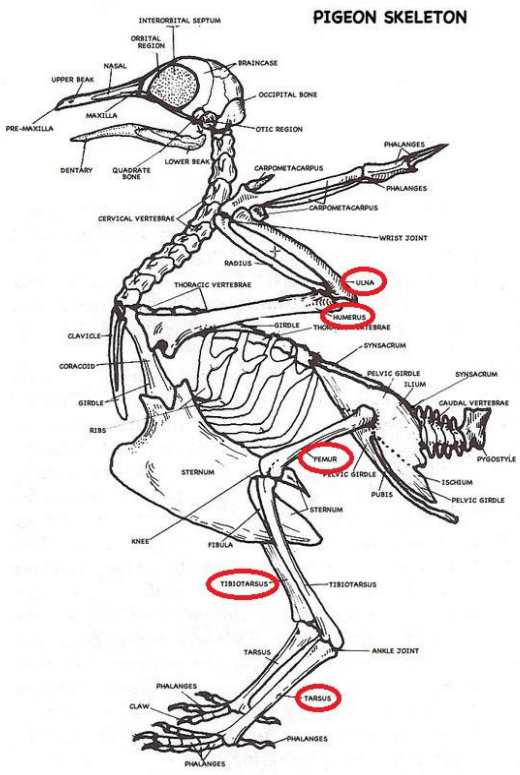


Figure 2: Pigeon Skeleton with different bones marked

The five specific bones of the wing and leg helped to categorise the birds into six groups while taking into consideration that if birds can run, have strong and long legs and birds that fly have strong wings. Humerus and Ulna are wing bones whereas the Femur, Tibiotarsus and Tarsometatarsus are leg bones. *Figure 2:* shows the selected bone positions in birds.

Data Understanding

In the data understanding phase, the dataset should be well understood. Here, in this dataset, there are 420 observations and 12 attributes (the length and diameter of each of the five bones, the ecological type and row id). To further understand the data we used bar graphs to find the proportion of the bird type and find the missing values, scatter plots to see the distribution of the data and correlation matrix to see how well the dataset is correlated.

Proportion of data

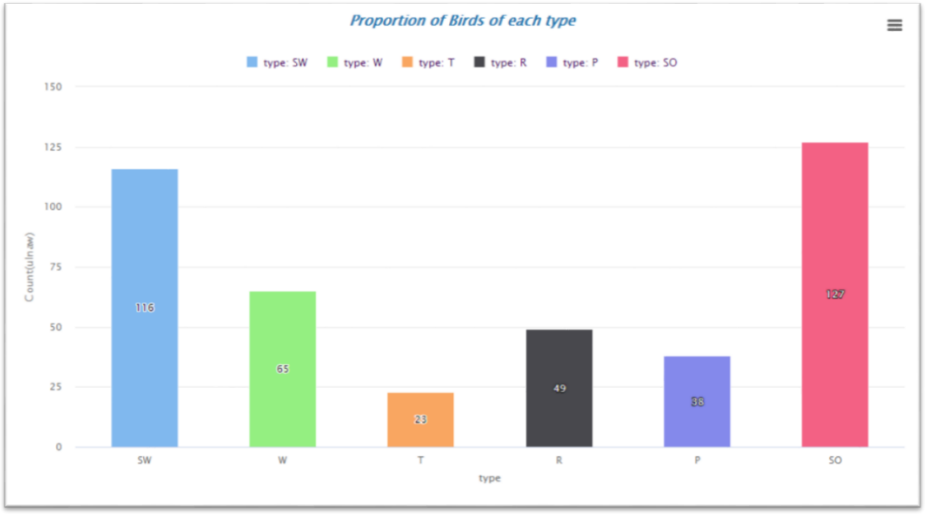


Figure 3: The horizontal graph above provides the proportion of data based on the six ecological groups of birds

The above bar graph shows the proportion of the birds where swimming and singing birds are relatively high, 116 swimming birds and 127 singing birds, i.e. 27.6% and 30.2% respectively. The rest of the contribution if from the Wading, Terrestrial, Raptors and Scansorial birds which are comparatively lower.

Finding Missing Values

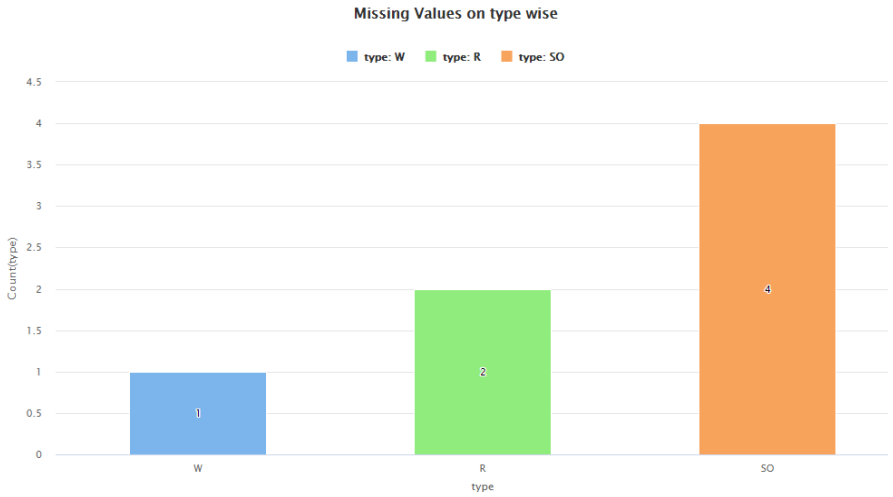


Figure 3: Missing Values

A few missing values were identified within three ecological groups, one from Wading, two from Raptor and four from Singing bird types. The process of finding the missing values can be seen in Appendix A.

Scatter Plots

Consistent (=Good) vs Non-Consistent (=Poor) View: Clustering is the process of partitioning data into homogeneous groups. While there are various potential criteria's in defining a good view but for a class structure a good view should be at least consistent with that class structure as demonstrated in the below two graphs "scatter plot for length and width of Femur" & "scatter plot for length and width of Tibiotarsus".

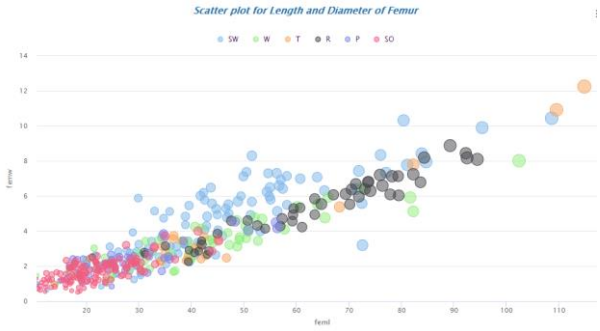


Figure 4: Scatter plot Femur

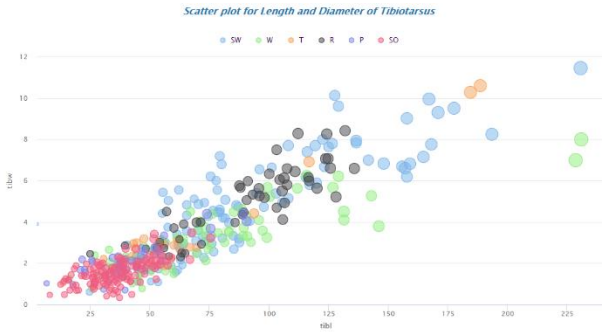


Figure 5: Scatter plot Tibiotarsus

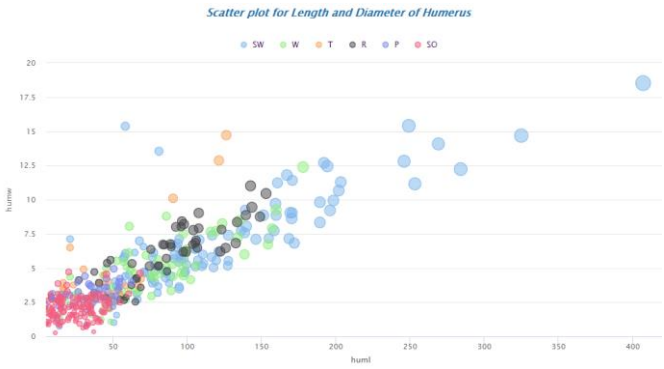


Figure 6: Scatter plot of Humerus

Both graphs portray six clusters representing six groups of birds as mentioned on the legend, and the ten attributes are the length and width of the five bones used to classify the birds. The scatterplot measurements used are, width and length of the respective bones to classify the group of the bird it falls into. We can see that most data points are located close to the class centres showing linearity and resulting in a consistent view.

In contrast, the below two scatterplots "scatter plot for length and width of Ulna" and "scatter plot for length and width of Tarsometatarsus", the classes are separated and not clustered together, resulting in a poor consistency rating.

Outliers

An outlier is a data point that differs significantly from other observations. It only could mean an observation coming from a different group or subset, and does not have any undesirable implications. Sometimes an outlier, on the other hand, could be that it is the beginning of a new group within the same dataset. For instance, looking at in the graphs below we see a lot of plot that deviates from the linearity of the graph, and it could mean few different things, like an error in the data, missing values or even the beginning of a new species or variety of bird.

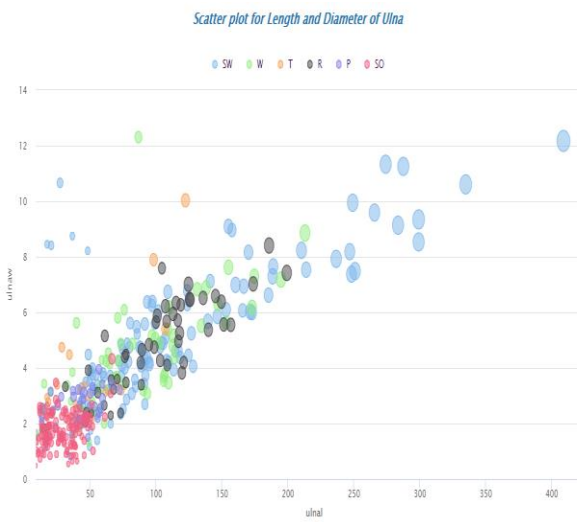


Figure 7: Scatter Plot for Ulna

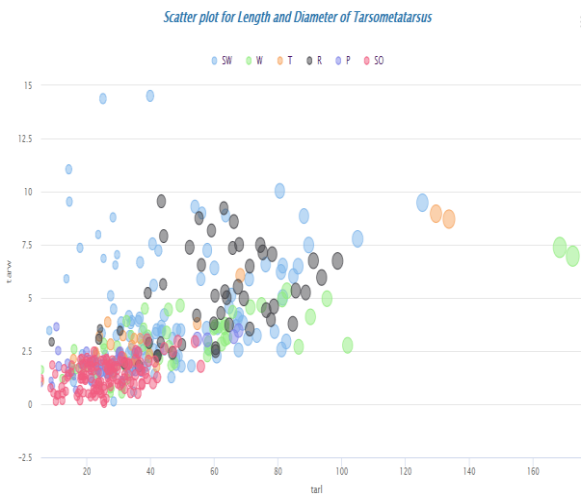


Figure 8: Scatter plot for Tarsometatarsus

Any data point that falls outside the data set's inner fences is classified as a minor **outlier**, while one that falls outside the outer fences is classified as a major **outlier.**

Correlation Matrix

Another method of viewing a dataset is the correlation matrix which is a table showing correlation coefficients between sets of variables. In the correlation matrix, the value of each cell has a range of 0 to 1. For example, the attribute 'humw' and 'huml' has a high correlation of value '0.917', whereas the attribute 'tarw' and 'tarl' has a lesser correlation as shown in figure 6 and 8.

Finding the correlation between the Explanatory Variables

Attribut...	huml	humw	ulnal	ulnaw	feml	femw	tibl	tibw	tarl	tarw
huml	1	0.917	0.976	0.887	0.771	0.850	0.827	0.869	0.695	0.751
humw	0.917	1	0.901	0.958	0.871	0.936	0.820	0.901	0.676	0.892
ulnal	0.976	0.901	1	0.872	0.759	0.838	0.762	0.819	0.654	0.738
ulnaw	0.887	0.958	0.872	1	0.826	0.898	0.797	0.870	0.643	0.871
feml	0.771	0.871	0.759	0.826	1	0.945	0.860	0.904	0.832	0.869
femw	0.850	0.936	0.838	0.898	0.945	1	0.868	0.960	0.779	0.902
tibl	0.827	0.820	0.762	0.797	0.860	0.868	1	0.929	0.922	0.737
tibw	0.869	0.901	0.819	0.870	0.904	0.960	0.929	1	0.826	0.855
tarl	0.695	0.676	0.654	0.643	0.832	0.779	0.922	0.826	1	0.606
tarw	0.751	0.892	0.738	0.871	0.869	0.902	0.737	0.855	0.606	1

Figure 9: Correlation Matrix

Data Pre-processing

The Data Pre-processing Phase is an essential step in the data mining process as the phrase "garbage in, garbage out" is particularly applicable to data mining projects. The steps included in data pre-processing are eliminating missing values by finding the mean measurements instead of removing those values. By finding the mean value, missing values were eliminated. When attributes are numeric, as this dataset, an arithmetic

measure of central tendencies, such as **mean, median** or **mode** is an acceptable replacement for missing values.

Therefore, to handle these missing values Rapid Miner uses the operator 'Replace Missing Values", and by selecting the attribute, the tool calculates the mean value of all the variables of that attribute and fills in the missing field. Mean substitution is considered an inferior approach but superior to listwise deletion, therefore, as data miners, we must be responsible for thinking about each change we make in our data, and whether or not we threaten the integrity of our data by making that change (North, 2012).

Data Modelling

On the complete understanding of the dataset, several models such as Naïve bays, Decision tree, Random Forest and KNN were tried and tested, and best models were selected. To train and test the models, Random Sampling and cross-validation methods are used.

Random Sampling and Cross-Validation

In finding the right model for the bird dataset, methods such as Random sampling and cross-validation were used to the train and test the model. The training set was used to train the model, and the test set used to validate data it has never seen before using the classic approach of sampling the data into 70%-30%.

Cross-validation is a method for modelling according to the predictive ability of the models. In cross-validation, the data is split into several partitions to train multiple algorithms on the partitions to improve the robustness of the model by holding out data from the training process. The data is split into 3, 5, 10 or any K number of splits. In this manner, all parts of the dataset are tested alternatively (Shao, 1993). To fit the models 10-fold cross-validation is used. The results obtained from the cross-validation method produced much higher accuracy results which proved a better methodology for training the data.

The Evaluation of Results

Decision Tree

Decision trees are powerful and popular tools for classification. A decision tree is a tree-like structure, which starts from root attributes and ends with leaf nodes. It performs different tests with the data by using more than one attributes in test leaves. The algorithms describe the relationship between attributes and the relative importance of attributes. The advantages of decision trees are that they represent rules which could easily be understood and interpreted by users, do not require complex data preparation, and perform well for numerical and categorical variables (Bhargava et al., 2013). For this dataset, the 'Humerus length' was taken and the decision of separating the other length and width of the bones into smaller branches in a similar manner until it reached the fifth level where it classified the birds into ecological groups. The process of obtaining the results can be seen in Appendix B, C & D

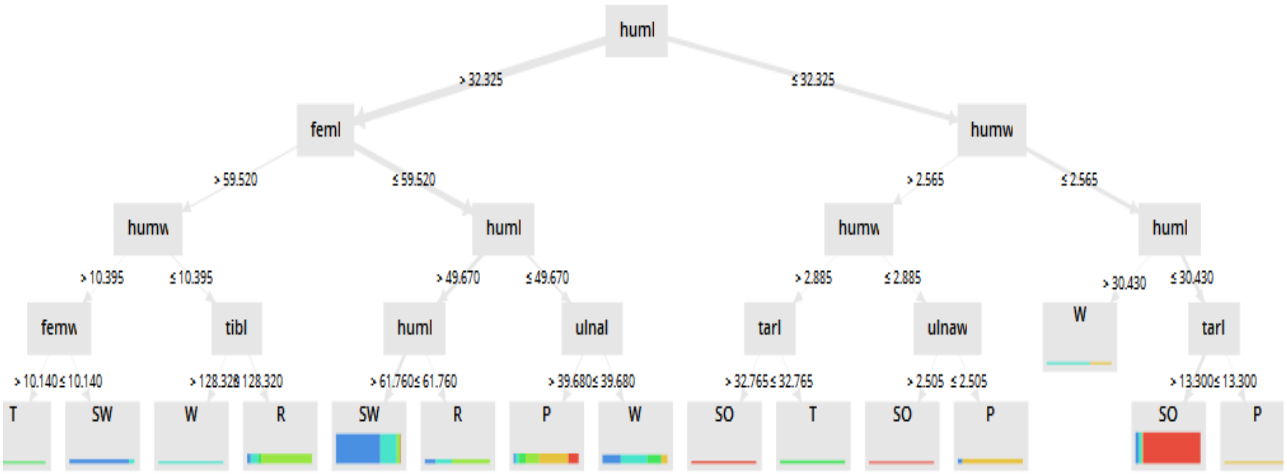


Figure 10: Decision Tree for Cross-Validation.

Figure 11 & 12 shows the results for Decision Tree using Random sampling and Cross-Validation in a Confusion matrix.

accuracy: 64.29%

	true SW	true W	true T	true R	true P	true SO	class precision
pred. SW	28	7	1	4	0	0	70.00%
pred. W	1	6	0	0	1	0	75.00%
pred. T	1	0	4	0	1	1	57.14%
pred. R	1	7	1	8	1	0	44.44%
pred. P	0	1	0	2	3	0	50.00%
pred. SO	3	2	3	0	7	32	68.09%
class recall	82.35%	26.09%	44.44%	57.14%	23.08%	96.97%	

Figure 11: Results for Decision Tree using Random sampling

accuracy: 68.10% +/- 6.37% (micro average: 68.10%)

	true SW	true W	true T	true R	true P	true SO	class precision
pred. SW	88	34	3	11	1	4	62.41%
pred. W	11	13	2	1	0	1	46.43%
pred. T	3	0	10	1	2	1	58.82%
pred. R	5	10	2	34	4	0	61.82%
pred. P	3	2	1	2	21	2	67.74%
pred. SO	6	6	5	1	10	120	81.08%
class recall	75.86%	20.00%	43.48%	68.00%	55.26%	93.75%	

Figure 12: Results for Decision Tree using Cross-Validation

From above (figure 12) from a total of 116 swimming birds, 88 were classified as true swimming birds and rest 28 were miss classified as other categorised as other bird types. Similarly, out of 65 Wading birds, 13 were classified as wading birds and 42 were miss classified. And among these misclassified wading birds, 34 were

miss classified as swimming birds; this is because the length and width data points of Wading bird and Swimming birds are closely distributed.

The results obtained showed that Decision Trees do not work well for numerical attributes and not flexible when it comes to classifying new samples.

Random Forest

In the random forest approach, a large number of decision trees are created, and every observation fed into each decision tree. The most common outcome for each observation provided the final output. When a new observation is fed into all the trees a majority vote is taken for each classification. Therefore the results obtained were far better than the decision tree as seen below in figure 12 & 13. To obtain these results 20 trees were considered where each length or width of a bone was fed into each of the trees to achieve an outcome that classified the birds into its correct ecological groups. The process of obtaining the results can be seen in Appendix H, I & J

accuracy: 70.75%

	true SW	true W	true T	true R	true P	true SO	class precision
pred. SW	69	25	3	5	4	3	63.30%
pred. W	4	8	0	5	4	0	38.10%
pred. T	0	0	8	0	0	4	66.67%
pred. R	4	4	2	26	1	0	70.27%
pred. P	2	0	1	0	14	5	63.64%
pred. SO	3	5	0	0	2	83	89.25%
class recall	84.15%	19.05%	57.14%	72.22%	56.00%	87.37%	

Figure 13: Results for Random Forest Random Sampling

accuracy: 76.43% +/- 5.66% (micro average: 76.43%)

	true SW	true W	true T	true R	true P	true SO	class precision
pred. SW	100	29	3	10	0	1	69.93%
pred. W	5	19	1	0	3	0	67.86%
pred. T	1	0	12	0	0	0	92.31%
pred. R	3	6	2	37	3	1	71.15%
pred. P	0	3	3	3	28	1	73.68%
pred. SO	7	8	2	0	4	125	85.62%
class recall	86.21%	29.23%	52.17%	74.00%	73.68%	97.66%	

Figure 14: Results for Random Forest Cross-Validation

From the above figure 14, we can observe that Random forest using cross-validation gives better results than the Decision trees. Here, Swimming birds are classified well by Random forest model with an accuracy of 86.21%. Noticeably, singing birds classified with 97.66% accuracy.

The k-Nearest Neighbour (k-NN)

The k-Nearest Neighbour (KNN) is a method used for classifying data points based on its closest neighbour point. The algorithm used is the majority vote of its K nearest neighbours where K is initially assigned a minimum value and thereby tested by increasing the K value (Cunningham, P., & Delany, S. J, 2007). For this

method, we used K=2 value and the results obtained are as shown below. Refer to Appendix E, F & G for Rapid Miner steps used to establish these results.

accuracy: 78.57%

	true SW	true W	true T	true R	true P	true SO	class precision
pred. SW	65	9	0	1	0	1	85.53%
pred. W	9	24	1	0	4	1	61.54%
pred. T	2	2	9	1	1	4	47.37%
pred. R	1	2	4	32	3	1	74.42%
pred. P	2	3	0	2	13	0	65.00%
pred. SO	3	2	0	0	4	88	90.72%
class recall	79.27%	57.14%	64.29%	88.89%	52.00%	92.63%	

Figure 15: KNN Random Sampling

accuracy: 84.64% +/- 6.98% (micro average: 84.69%)

	true SW	true W	true T	true R	true P	true SO	class precision
pred. SW	69	5	0	2	0	1	89.61%
pred. W	6	33	1	3	0	0	76.74%
pred. T	1	0	11	3	3	0	61.11%
pred. R	1	1	2	24	2	1	77.42%
pred. P	1	1	0	4	19	0	76.00%
pred. SO	4	2	0	0	1	93	93.00%
class recall	84.15%	78.57%	78.57%	66.67%	76.00%	97.89%	

Figure 16: KNN Cross-Validation

The various K values used for this process is recorded in the table below. For K=1 & 2 the results obtained were good as it could classify the group of birds accurately, but from the K value 3, the results decreased because the length and width of the birds bone are observed to be closely grouped as seen in the scatter plots. Therefore, for the larger K values, the model seems to be biased in predicting the results.

K- Value	Random sampling	Cross Validation
1	78.57	83.37
2	78.57	83.37
3	72.11	79.29
4	70.75	77.57
5	71.77	75.51
6	68.71	77.2
7	69.05	74.13
8	67.69	72.75
9	65.99	71.72
10	66.33	70.36

Figure 17: KNN model accuracies by changing the k value using Random Sampling and Cross-Validation

The final results of cross-validation and random sampling, provided in the table below;

Model	Random Sampling	Cross-Validation
Decision Tree	64.29	68.10
Random Forest	70.75	76.43
KNN (K = 2)	78.57	84.64

Table 1: Accuracy of the models compared in Rapid miner

Table 1 shows that cross-validation is the better approach in fitting the model for the birds bone data set than Random sampling. In the model used, KNN gave better results than Random forest and Decision tree.

Comparing Rapid Miner and R program results

The below tables show the accuracy of the models in R and Rapid Miner.

Models used	R Using Cross-Validation	Rapid Miner	
		Random Sampling	Cross-Validation
KNN	84.13	78.57	84.64
Random Forest	78.57	70.75	76.43
Decision Tree	59.52	64.29	68.10

Table 2: Accuracy of the models compared with R and Rapid Miner

Table 2 compares the R program results with rapid miner results in classifying the birds to different ecological groups. The results from R and Rapid Miner were almost similar for KNN model; this shows that KNN is consistent in both the platforms in classifying the birds to ecological groups. Whereas the Decision Tree and Random Forest shows a small variation inaccuracy, this may be due to cross-validation sampling method technique on data set. Overall, KNN performed well in the classification of birds to different ecological groups based on bone length and diameter

Conclusion

The results achieved by applying selected data mining techniques and methods for classification on the Bird and Bone dataset reveal that KNN did classify birds 84.64% accurately into different ecological groups. Decision tree did not work well for this data set, as the data set is more numerical and Decision tree failed to classify the new data points. Random Forest gave noticeably good results than Decision tree in classify Birds based on ecological groups. It is observed that cross-validation gave better results than random sampling. When comparing the results with the results of the R program, KNN gave similar results (in both Rapid miner and R program), whereas the Decision tree showed some variation. It is rightful to consider these initial steps as best practice for data mining processes used for data mining projects. As a result of this, the objective of the report established by using the CRISP-DM framework to find the data model that provided accurate results.

Reference

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, *3*(6).

Cunningham, P., & Delany, S. J. (2007). k-Nearest neighbour classifiers. *Multiple Classifier Systems*, *34*(8), 1-17. Retrieved from https://www.researchgate.net/profile/Sarah_Delany/publication/228686398_k-Nearest_neighbour_classifiers/links/0fcfd50d0c1d1f41ad000000/k-Nearest-neighbour-classifiers.pdf

Dr D. Liu, “Birds' Bones and Living Habits”, “Measurements of bones and ecological groups of birds”, Retrieved from https://www.kaggle.com/zhangjuefei/birds-bones-and-living-habits

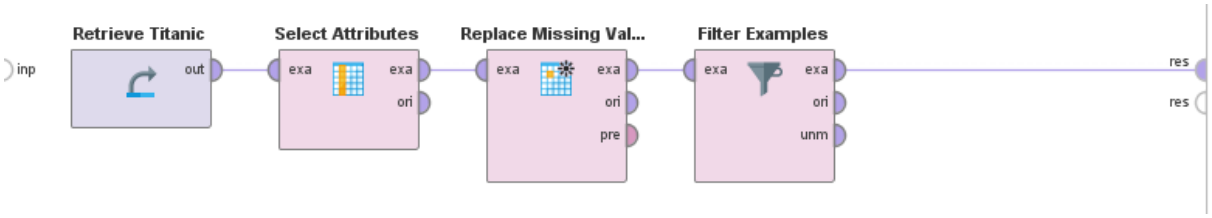
Feyyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert*, *11*(5), 20-25. Retrieved from https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=539013

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486-494. Retrieved from https://www.tandfonline.com/doi/pdf/10.1080/01621459.1993.10476299

Sips, M., Neubert, B., Lewis, J. P., & Hanrahan, P. (2009, June). Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum* (Vol. 28, No. 3, pp. 831-838). Oxford, UK: Blackwell Publishing Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-8659.2009.01467.x

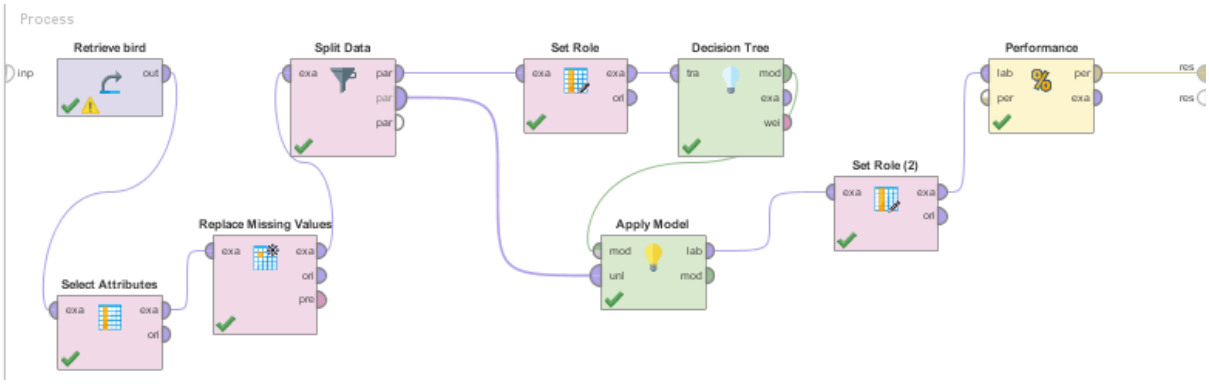
Appendices

Appendix A



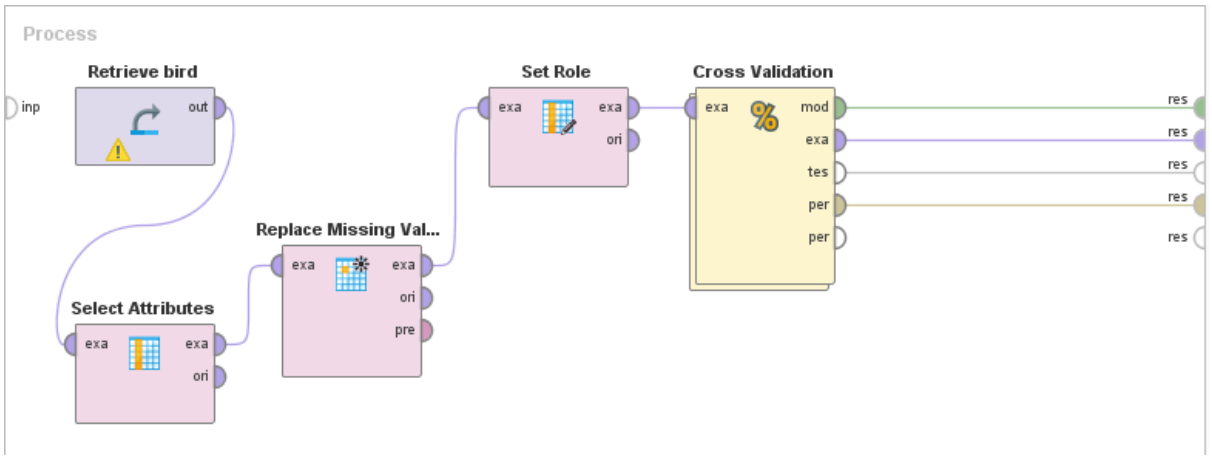
Appendix A: Process to find missing values

Appendix B



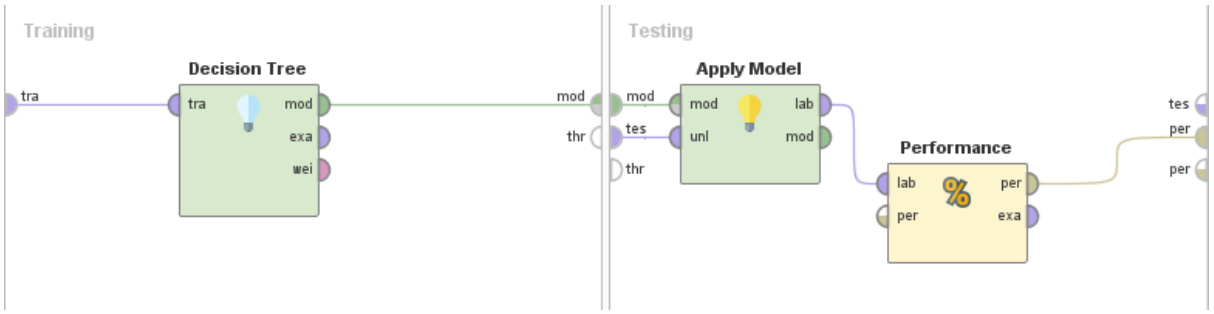
Appendix B: Process for Decision tree model using Random Sampling

Appendix C



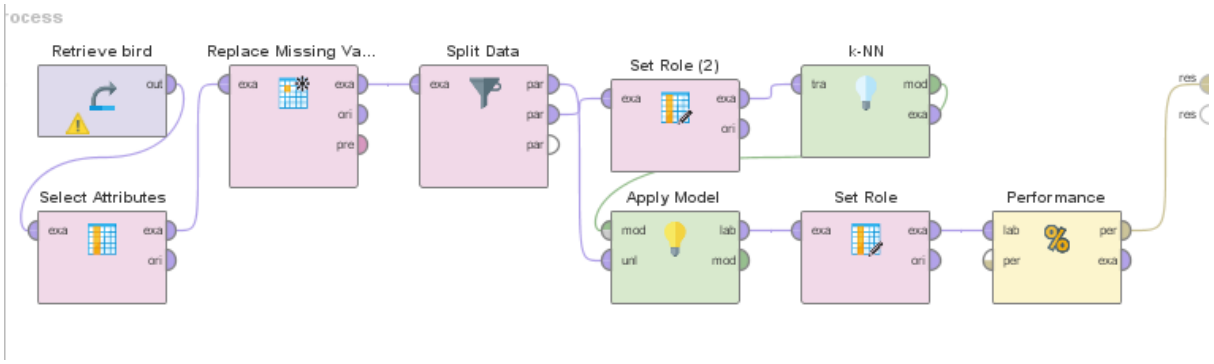
Appendix C: Process for Decision Tree model using Cross Validation – part A

Appendix D



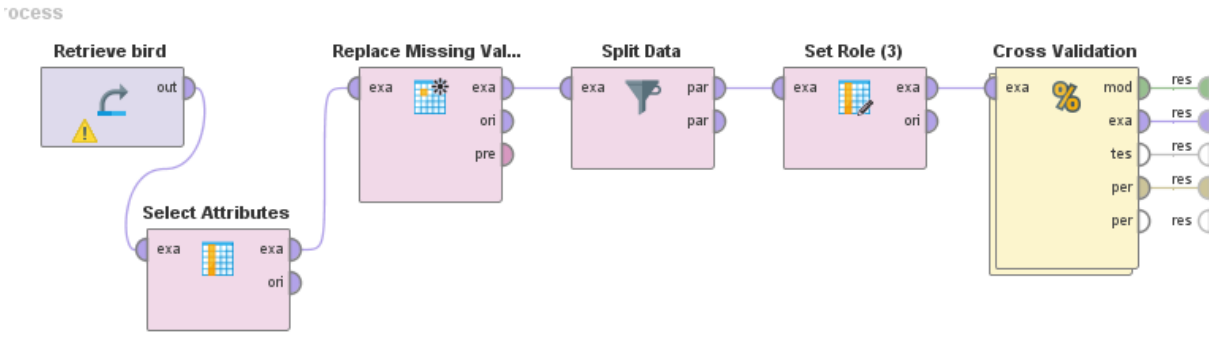
Appendix D: Process for Decision Tree Model using Cross Validation - part B

Appendix E



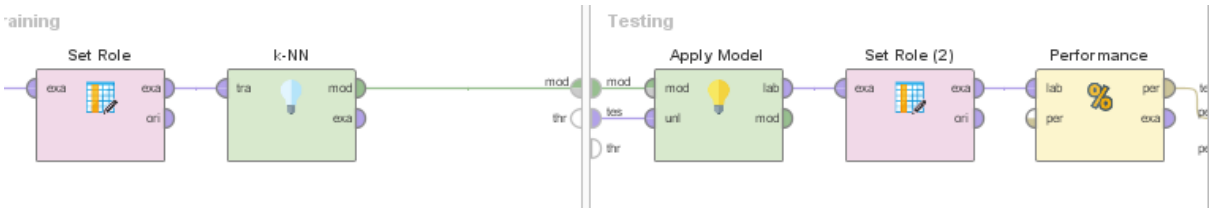
Appendix E: Process for KNN Model using Random Sampling

Appendix F



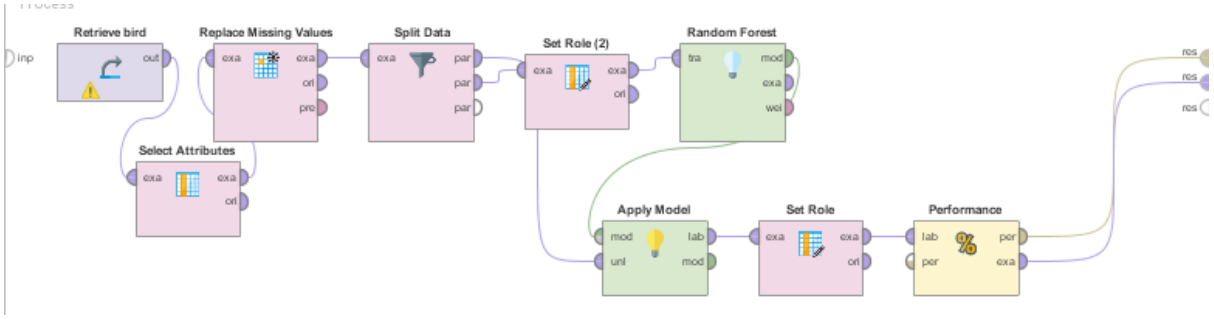
Appendix F: Process for Cross-Validation using KNN- Part A

Appendix G



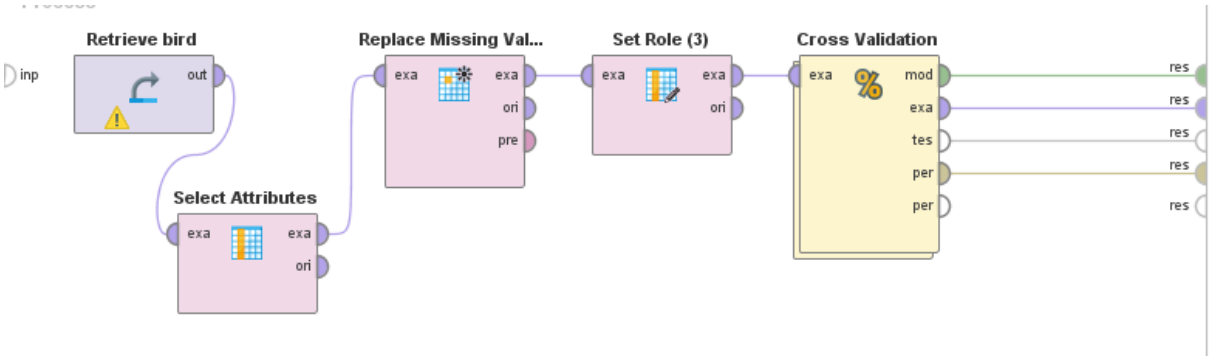
Appendix G: Process for Cross-Validation using KNN- Part B

Appendix H



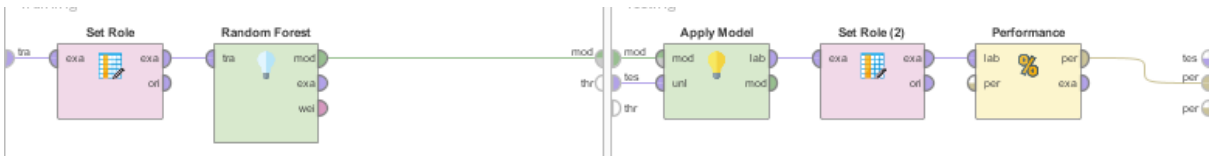
Appendix H: Process for Random Forest model using Random Sampling

Appendix I



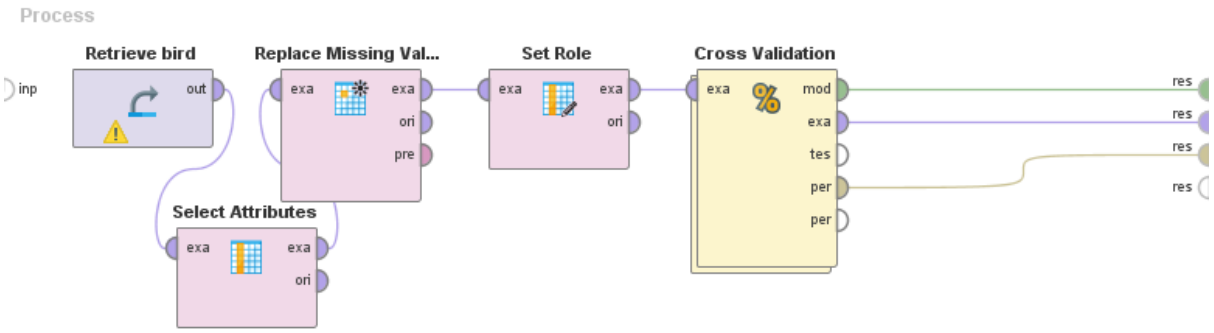
Appendix I: Process for Random forest using Cross-validation- Part A

Appendix J



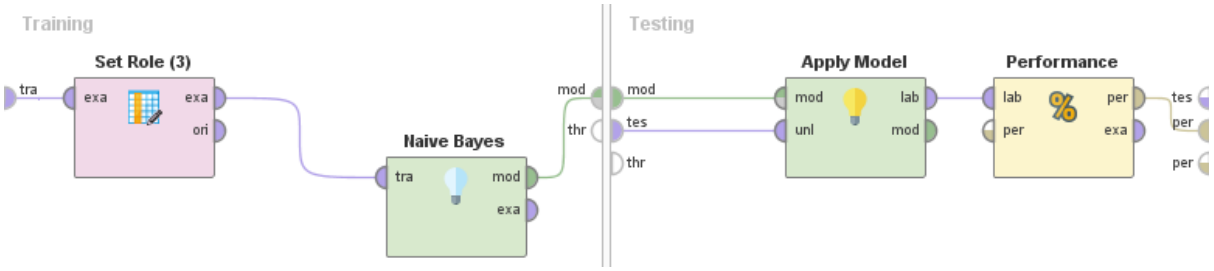
Appendix J: Process for Random Forest Using Cross-Validation - Part B

Appendix K



Appendix K: Process for Naive Bayes model using Cross-Validation – Part A

Appendix L



Appendix L: process for Naive Bayes model using Cross-validation - Part B

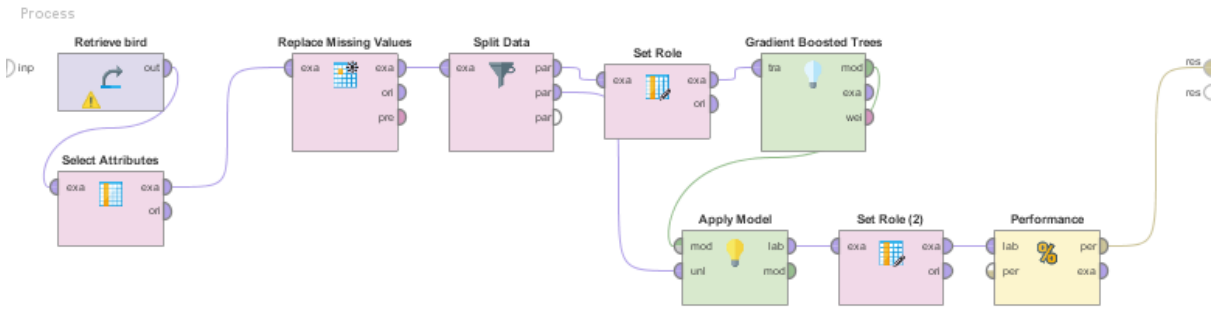
Appendix M

accuracy: 51.19% +/- 7.47% (micro average: 51.19%)

	true SW	true W	true T	true R	true P	true SO	class precision
pred. SW	36	6	1	3	0	0	78.26%
pred. W	45	22	3	5	6	1	26.83%
pred. T	1	1	0	0	0	0	0.00%
pred. R	17	12	3	33	0	0	50.77%
pred. P	8	10	15	9	13	16	18.31%
pred. SO	9	14	1	0	19	111	72.08%
class recall	31.03%	33.85%	0.00%	66.00%	34.21%	86.72%	

Appendix M: Naive Bayes model results

Appendix N



Appendix N: Process for Gradient Boost model using Random sampling

Appendix O

accuracy: 77.78%

	true SW	true W	true T	true R	true P	true SO	class precision
pred. SW	28	5	0	1	0	0	82.35%
pred. W	3	10	0	0	0	0	76.92%
pred. T	0	0	4	0	0	0	100.00%
pred. R	2	0	2	12	0	0	75.00%
pred. P	0	3	3	1	11	0	61.11%
pred. SO	1	5	0	0	2	33	80.49%
class recall	82.35%	43.48%	44.44%	85.71%	84.62%	100.00%	

Appendix O: Results for Gradient Boost