Advanced Statistical Approaches to Quality (INSE 6220)
Student ID: [redacted]
Fall 2021

Classification performance analysis of Birds' ecological category dataset utilizing principal component analysis, logistic regression and decision trees

Mahsasadat Noori Najafi[1]

[1]**Concordia Institute for Information Systems Engineering**

ABSTRACT

This project aims at depicting the importance of the principal component analysis (PCA) for the task of binary classification. PCA, as a multivariate data analysis, extracts the existing correlation between different features and by doing so generates a set of new orthogonal features that are sorted in a descending order with respect to the amount of containing variance. This set of new features are called principal components. In circumstances where the data variance is only distributed along some few principal components, PCA acts as a dimensionality reduction method - this is what PCA is essentially known for. Our data set (source) [1] is a collection of information obtained about birds' physical attributes and their habitats along with their ecological categories (simply referred to as types). As we will in this report, there exists a strong correlation between birds' skeleton and their types. We use two classification algorithms 1) logistic regression and 2) decision trees. Our algorithms successfully predicted the birds' types with 94% and 90% accuracy for logistic regression and decision trees respectively. Also, notably, the first two principal components obtained via PCA covers 98% of the data variance.

Keywords: PCA, Logistic Regression, Decision Trees, Classification, Prediction

1 INTRODUCTION

In this work we aim to predict birds' ecological categories by their appearances and more specifically by their skeleton shape. The original dataset was collected by Dr. D. Liu of Beijing Museum of Natural History (source) [1]. It covers birds' bone information of the following birds' ecological classes: Swimming Birds (SW), Wading Birds (W), Terrestrial Birds (T), Raptors (R), Scansorial Birds (P), Singing Birds (SO). Scientists' studies discovered that birds' skeleton evolution is based on their habitat [2]. On the other hand, by measuring the size of birds' bones and appearances features define the location they belong to [2].

In this report, we focus on separating on two types Singing Birds (SO) and Swimming Birds (SW) which are the dominant classes in the original data set. The independent variables in our data set are the bone measures. Table 1 describes the used features in our data set and their description. In summary, 10 measurements are reported and their units are all millimeters.

feml	Length and Diameter of Femur
tibl/tibw	Length and Diameter of Tibiotarsus
tarl/tarw	Length and Diameter of Tarsometatarsus

Table 1. Feature Description

Fig 1 represent the typical birds' skeleton where you can find the upper leg named Femur. Tarsometatarsus, one the largest bone of birds' lower leg and Tibiotarsus, a large bone between the Femur and Tarsometatarsus.

2 PRINCIPAL COMPONENT ANALYSIS

Although dimensionality reduction may cause some information loss, it extremely speeds up training, filter out some noises along with removing unnecessary details and highly useful for data visualization, which result in performance enhancing [3].

There exist many other techniques for the purpose of

dimensionally reduction such as LLE (Locally Linear Embedding), MDS (Multidimensional Scaling), etc. However, PCA is the most popular one [3].

PCA is a preprocessing step and an unsupervised machine learning algorithm. There are many advantages associated with this method. PCA helps us to compress the original data set in a manner the number of features will be reduced and the first principal component will have the largest variance which is calculated based on largest variances is possible in the data set. the rest of the components has the highest variance amount respectively. practically, it is impossible to calculate principal components manually, this is a main reason, written packages in programming languages like Python, for machine learning algorithms are the current trend in the field of Data science/Data analytics.

In addition to dimension reduction, the training set after applying PCA takes up much less space in compassion to original data set.

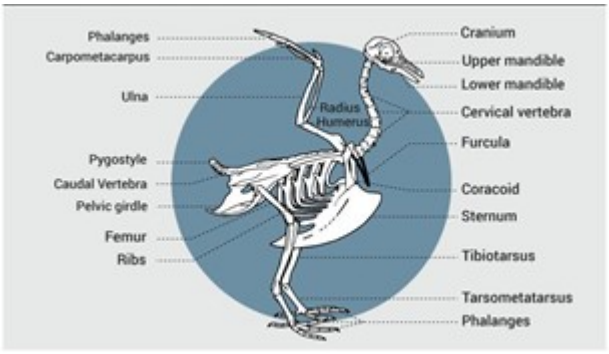

Figure 1. A typical Birds' Skelton

This report includes 2 main parts, at first, we applied PCA, considering unsupervised machine learning technique, were the label excluded. in the second part, label included (binary label- Singing Birds (SO) or Swimming Birds (SW)), 2 classifiers have been used to build the classification model.

3 CLASSIFICATION

In this section, we briefly recall the two methods we will employ in our note. Namely, logistic regression algorithm and decision trees.

3.1 Logistic regression

Numerous regression algorithms like Logistic regression (Logit Regression) could be applied for binary classification. Simply it could predict if an example in the training data set belong to a certain class. Upon fixing a threshold τ, if the estimated probability is greater than τ, the model predicts the instance belong to the class has binary labels equal to 1 or 0. [4] In this note, diagnosing the birds 'ecological classes is predicted using classifiers.

The same as every other regression model, logistic regression model calculates a weighted sum of the input features plus a bias term. Logistic function is as follows:

$$\sigma(t) := \frac{1}{1+\exp(-t)}$$

Logistic regression model estimates the probability of belonging to each class for a linear classifier $\boldsymbol{x}$ according to the following formula

$$\hat{p} = \sigma(\boldsymbol{x}^T \boldsymbol{\theta})$$

Notice that here $\boldsymbol{x}$ is augmented by a dummy dimension to account for the bias term. In this note, for brevity, we omit the details.

3.2 Decision Trees

Decision Trees are multifaceted Machine Learning algorithms, it is even considered as regression task and also multi-output tasks. These algorithms capable of fitting the highest complex data sets.

Decision Trees are the major components of Random Forest, which are one the most powerful and the well-known Machine Learning algorithm these days. Briefly, the training data set considered as a root node. Then, the node will be divided into two or more sub-nodes. Next, if the sub-nodes divided to another sub-nodes, decision node created, else we call them leaf or terminal node. Additionally, a sub-tree is a branch of entire tree subsection. Even, the node is divided to sub-nodes s called parent and respectively, the sub-nodes are the child, Fig.... depicts the decision tree flowchart [5].

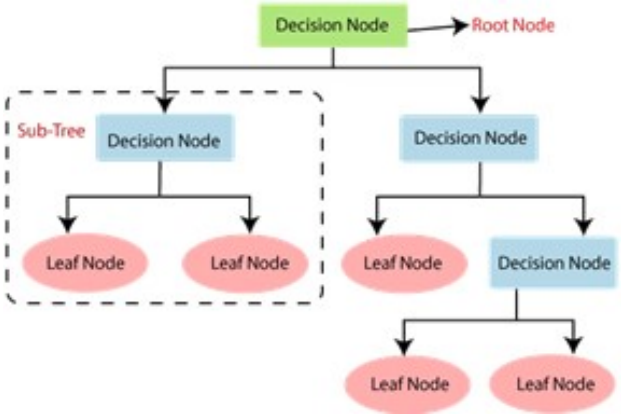

Figure 2. Decision Trees flowchart

4 EXPERIMENT

4.1 Data

Our main focus for the first part of report would be dimension reduction, as it has been mentioned earlier the PCA was applied to reduce the dimension. The dataset is a collection of 240 entries, consist of 5 columns as independent variables include:'huml', 'humw', 'ulnal', 'ulnaw', 'femw'. Which all have been explained earlier in Table 1

For the second part of the report, we considered the dependent variable in binary classification represent 128 Singing Birds (SO) with value equal to 0 and 116 Swimming Birds (SW) cases show by value equal to 1. Fig 3, represent the dataset is equally balance, means the outcome is approximately equal in two binary classes.

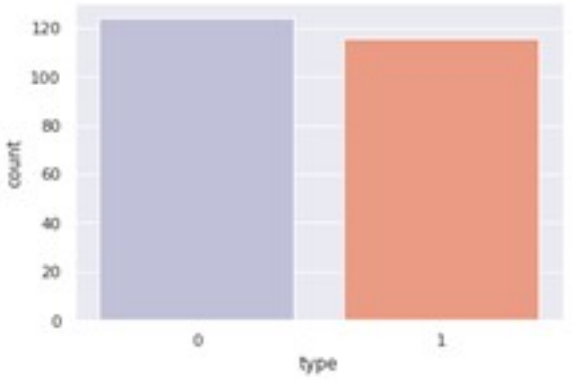

Figure 3. - Birds' ecological binary classes

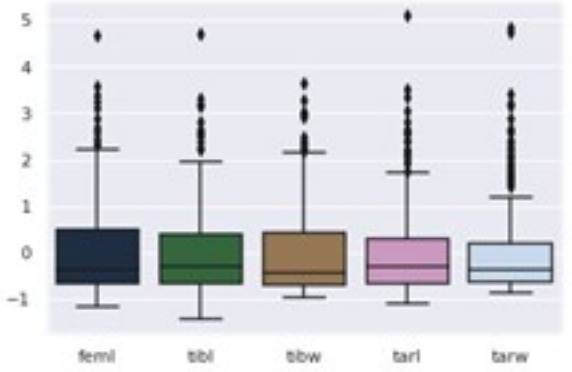

Figure 4. Box-plot of centered feature vectors

Fig 4, represents the bar plot for normalized independent variables, along with the distribution they follow, central values and the data are located outside the of the box-plot named outliers. In summary, the plot depicts the outliers as well as the shape of distribution.

Fig 4 and Fig 5 assessed that all the box-plots follow the "skewed right" distribution with a long right tail, which means the distribution is not symmetrical.

Fig 6, represents the correlation among independent variables (features) by the aid of covariance matrix. The most highly positive correlated features in our dataset are 'tibw' and 'tibl', 'tibw' and 'feml'. In addition, the lowest correlation exist between 'tarw' and 'tarl'. Hight correlations can be clearly seen in Pair-plot, Fig7. All the observations are divided by the hue of classification type in Pair-plot. Furthermore, It shows the data is linearly separable.

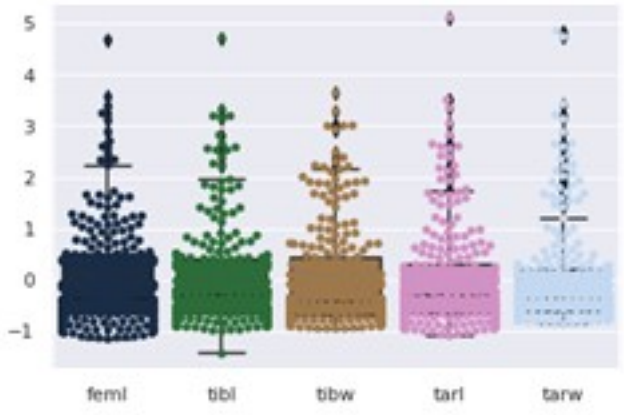

Figure 5. Swarm plot

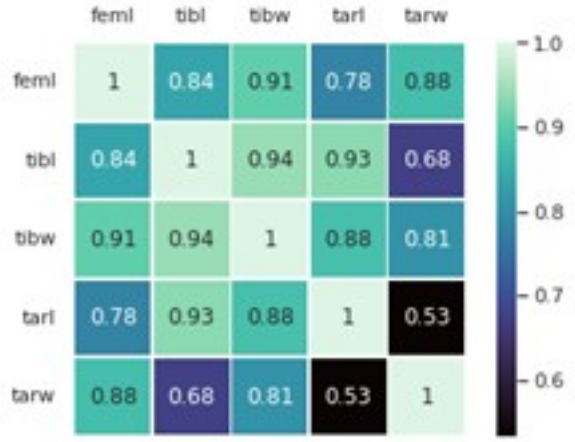

Figure 6. Dimension reduction by PCA

4.2 Dimension reduction by PCA

This section discusses the dimension reduction technique PCA. The original dataset contains 5 features, PCA application reduce the number of independent variables to any number less than 5, for this data set.

The main procedure to apply PCA is as below:

1. **Standardization**

   After computing the center of the data set, we should subtract off-column means to obtain the centered dataset $Y = HX$. We would like to emphasize that since the Scikit-Learn PCA API automatically centers the dataset [4], we did not perform the centering phase in the final code. In our initial draft, we performed the centering and compared the result with the non-centered data and we did not observe any significant difference.

2. **Covariance matrix computation**

   This step represents the relationship among variables. as it has mentioned in the definition of PCA, its main goal is to created uncorrelated features. while the

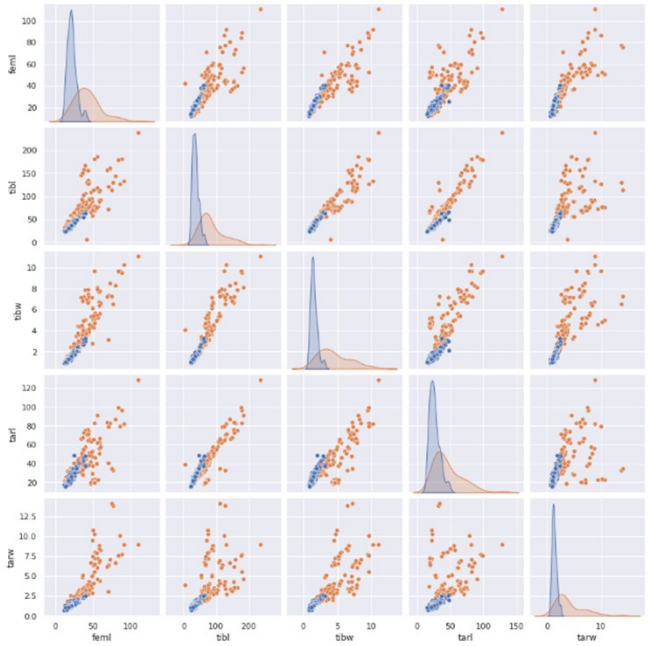

Figure 7. Pair plots

correlation in original data depicts the redundant information or duplicate features in the case of high correlated variables [6, 7].

$$S = \frac{1}{n-1} Y^T Y$$

3. **Eigenspace calculation**

   In order to compute the principal components, we need to calculate the eigenvector, eigenvalue pairs of the covariance matrix:

$$S = A\Lambda A^T = \sum \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T$$

   We then compute the transformed data in the new coordinate as below:

$$Z = YA \in \mathbb{R}^{n \times p}.$$

   The main purpose if principal component is to compressed or squeezed information within the initial components, whereas the correlation among them is zero. principal components cover the maximum amount of variance in the data set, to represent the amount each principal component cover, we use eigenvalues which are linked to Eigen-vectors. After computing the Eigen-vectors, we have to put them in descending order from largest to smallest [6, 7].

4. **Feature vector**

   To create matrix of vectors, we choose to keep all of the principal components or only few of them with the highest significance. In this regard scree plot and Pareto plot will assist us [6, 7]. Scree plot involves the percentage of variances computed as follows:

$$\ell_j := \frac{\lambda_j}{\sum_{i=1}^{n} \lambda_i}, \ \forall j = 1, \cdots, n.$$

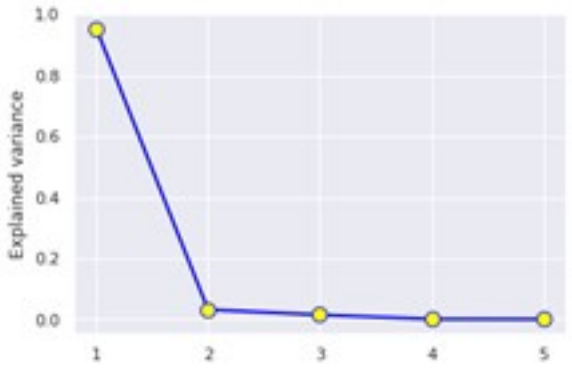

Figure 8. Scree plot

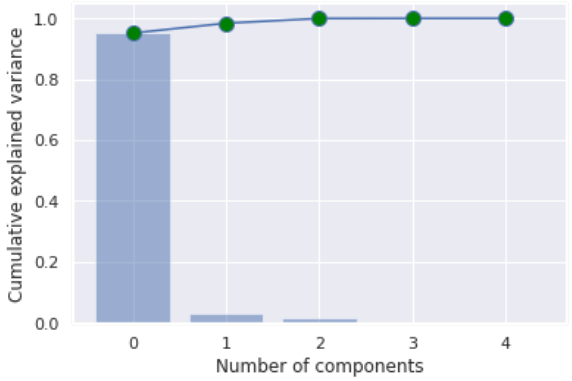

Figure 9. Pareto Plot

Following you can see the first two principal components. Where the 'tibl' and 'feml' have the most contribution in PC1 and PC2 respectively.

$$Z_1 = 0.335X_1 + 0.852X_2 + 0.399X_4$$
$$Z_2 = 0.9256X_1 - 0.2917X_2 - 0.1743X_4 + 0.162X_5$$

The above equations are written based on Eigen vector-matrix :

$$A = \begin{pmatrix} 0.335 & 0.926 & -0.053 & 0.167 & 0.017 \\ 0.852 & -0.297 & 0.433 & 0.048 & -0.012 \\ 0.046 & 0.041 & 0.006 & -0.410 & 0.910 \\ 0.399 & -0.174 & -0.891 & -0.114 & -0.0575 \\ 0.037 & 0.162 & 0.125 & -0.889 & -0.411 \end{pmatrix}$$

Eigenvalue in this report equals the following vector:

$$\Lambda = \begin{pmatrix} 2155.100 \\ 73.434 \\ 35.518 \\ 0.733 \\ 0.209 \end{pmatrix}$$

First two number of eigenvalues are the highest, therefore we keep the first two principal components. It is obviously can be seen from Fig 8 and Fig 9; the first two principal components cover 98% of data set variance.

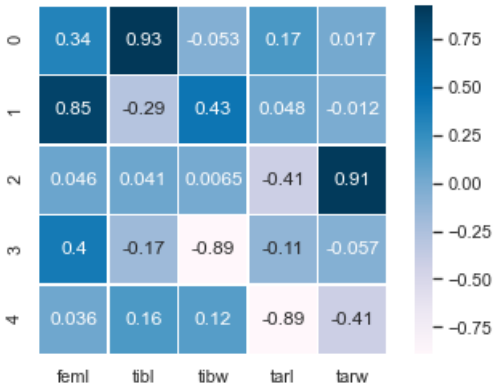

Figure 10. PCs components

Fig 10, and Fig 8, depict the amount of contribution for each variable, 'tibl' and 'femi' as the highest positive involvement (respectively, 0.85 and 0.92). Additionally, the Fig 10 depicts the new uncorrelated variables are created by PCA. Other variables with lowest amount of involvement are located in the middle and bottom of figure. Fig 13, represents the PCA control chart. It shows that the majority of data are located inside the control chart and only a few points are out of the control.

The PCA control limit follow the below equations:

$$UCL = 3\sqrt{\lambda_j}$$
$$CL = 0$$
$$LCL = -3\sqrt{\lambda_j}$$

4.3 Classification

In this section, we used Machine Learning algorithms to predict birds' ecological classes. Two well-known algorithms are chosen for this work.

- Logistic Regression
- Decision Trees

We applied these algorithms to the original dataset, considering all principal components and only two principal components. Finally, we compared the results with the aid of performance evaluation metrics such as OC – curve, confusion matrix, accuracy, precision, recall, and F-score.

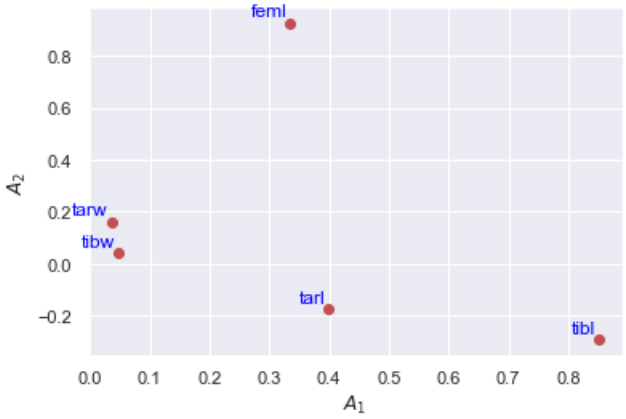

Figure 11. Scree plot

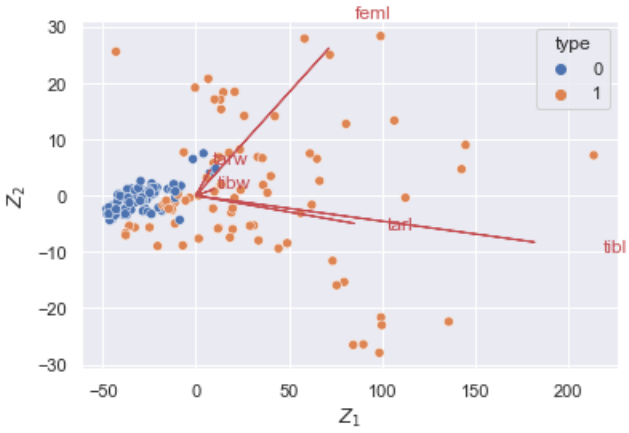

Figure 12. Bi-plot

4.4 Classification Report

Firstly, we applied cross-validation for both Logistic Regression and Decision Trees algorithms. The number of folds is equal to 5 in this work. In the other words, folds mean the entire dataset is divided randomly by 5 folds without replacement. Moreover, applying cross-validation gives us the expected level of fitness of a model to our data set.

Fig 14 and Fig 15, are the cross-validation report for each fold fit time and test accuracy. The first two-component fit time is extremely smaller than the original data. It clearly shows that PCA reduces the time complexity. Test accuracies are achieved by Logistic Regression fitted on original data is approximately the same as considering all principal components. While the result for test accuracies obtained by Decision Trees depicts the model fitted on all principal components worked better than original data. In the rest of the sections, we apply other classification metrics on original data to evaluate the performance of used algorithms by details.

Secondly, we used a confusion matrix to evaluate the created classification models.Fig 16, represents the confu-

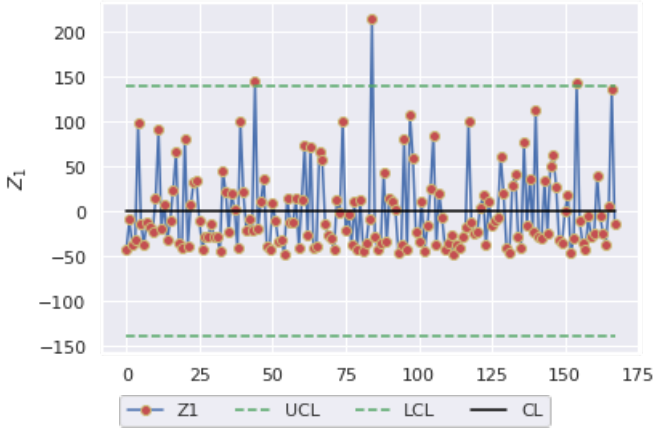

Figure 13. PCA control chart

sion matrix terminology. It gives us the number of true positives, false negatives, false positives, and true negatives. The confusion matrix gives us a more detailed analysis than only considering test accuracy. Fig 17 shows the model created by Decision Trees predict correctly 35 among 39 actually Positive outcomes and 30 among 33 actually negative outcomes.Fig 18 shows the model created by Logistic Regression predict correctly 39 among 39 actually positive outcomes and 29 among 33 actually negative outcomes.

Logistic Regression						
Fold	Fit time			Test Accuracy		
	Original Data	All Principal Components	First Two Component	Original Data	All Principal Components	First Two Component
0	0.041	0.017	0.008	0.853	0.853	0.765
1	0.026	0.021	0.009	0.971	0.971	0.941
2	0.022	0.014	0.008	1.000	1.000	0.912
3	0.021	0.021	0.008	0.939	0.939	0.758
4	0.026	0.016	0.009	0.909	0.909	0.909

Figure 14. Cross validation report-Logistic Regression

Decision Trees						
Fold	Fit time			Test Accuracy		
	Original Data	All Principal Components	First Two Component	Original Data	All Principal Components	First Two Component
0	0.003	0.001	0.001	0.882	0.941	0.882
1	0.003	0.001	0.000	0.89	0.97	0.91
2	0.003	0.001	0.002	0.97	0.97	0.941
3	0.003	0.001	0.001	0.82	0.939	0.848
4	0.003	0.002	0.001	0.970	0.96	0.909

Figure 15. Cross validation report-Decision Trees

Moreover, we assess the main classification metrics including: Precision, Recall, F_1-Score and Support.

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}},$$
$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}},$$
$$F_1\ \text{Score} = 2 \cdot \frac{\text{Recall} \times \text{Precision}}{\text{Recall}+\text{Precision}}.$$

Moreover, the Support column is defined as follows

Support =The number of True responses of each class

Let us define the terminology used in the preceding displayed equations.TN (True Negative) means when a case was negative and predicted negative. TP (True Positive) means when a case was positive and predicted positive. FN (False Negative) means when a case was positive but predicted negative. FP (False Positive) means when a case was negative but predicted positive. This can be summarized in Fig 16.

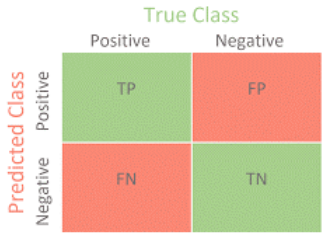

Figure 16. Confusion matrix terminology

	Actually Positive	Actually Negative
Predicted Positive	35	4
Predicted Negative	3	30

Figure 17. Confusion matrix for Decision Trees algorithm

	Actually Positive	Actually Negative
Predicted Positive	39	0
Predicted Negative	4	29

Figure 18. Confusion matrix for Logistic Regression

Fig 19 and Fig 20 represent the classification reports applied to the original data set. The figures help us to compare both algorithms. By comparing the weighted average

value for all scores, we conclude that Logistic Regression is slightly better than Decision Trees.

	Logistic Regression				
		Precision	Recall	F1-score	Support
	0	0.91	1	0.95	39
	1	1	0.88	0.94	33
Accuracy				0.94	72
Macro Avg	0.95	0.94		0.94	72
Weighted Avg	0.95	0.94		0.94	72

Figure 19. Classification Report for Logistic Regression

	Decision Trees				
		Precision	Recall	F1-score	Support
	0	0.9	0.92	0.91	38
	1	0.91	0.88	0.9	34
Accuracy				0.9	72
Macro Avg	0.9	0.9		0.9	72
Weighted Avg	0.9	0.9		0.9	72

Figure 20. Classification Report for Decision Trees algorithm

Finally, we assessed the ROC (receiver operating characteristics) curve for both algorithms. We used the AUC (area under the curve) ROC curve, to visualize the performance of both classifications. In summary, ROC represents the probability curve and the area under the curve depicts the measure of separability. Fig 21 and Fig 22, depict the ROC curve for Logistic Regression and Decision Trees fitted on original data. It clearly shows the AUC for Logistic Regression is slightly higher than Decision Trees. Therefore, the same as other measurement performance techniques we applied earlier, we can conclude that the Logistic Regression perfume is better for this data set.

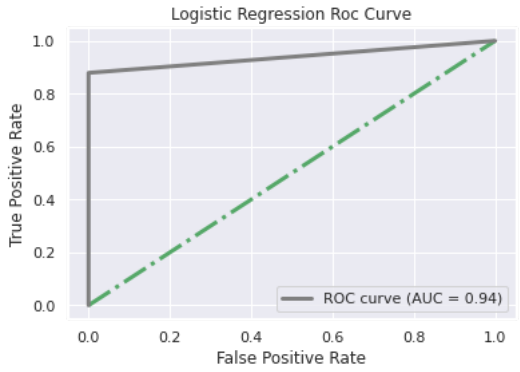

Figure 21. Logistic Regression ROC curve

Figure 22. Decision Trees ROC curve.png

5 CONCLUSION

In this work, we assessed some statistical aspects of the Birds' ecological category dataset. The first section discussed the data set description. We conclude that the data set is approximately balanced (128 Singing Birds (SO) and 116 Swimming Birds (SW)), features are highly correlated to each other. Moreover, based on the hue in Pair-plot we could clearly see the data set is linearly separable. In the second section, we applied Principal Component Analysis (PCA) as a dimension reduction technique. 98% of explained variance is covered by the first two principal components. The newly generated uncorrelated features helped us to have a better visualizing of data. We also applied the PCA components' control chart. It shows only 4 points are out of control in our data set. In the last part of the report, we worked on classification algorithms (Logistic Regression and Decision Trees) to see how Machine Learning algorithms predict the birds' ecological classes in our dataset. We applied the algorithms to original data, all transformed data by PCA and the first two components. The test accuracy for models fitted on original data was better than the models fitted on transformed data, specifically for Logistic Regression. But for the Decision Trees, the test accuracies between original data and transformed data are too close. Lastly, we decided to compare both algorithms by other classification metrics so we applied a confusion matrix and classification report (include: precision, recall, f1-score, and support). In all classification metrics, Logistic Regression performance was slightly better than Decision Trees. In closing, we could successfully forecast Birds' ecological classes by Logistic Regression and Decision Trees algorithms with high accuracy 94% and 90% respectively.

ACKNOWLEDGMENTS

I would like to thank Professor Ben Hamza for offering a great course. I truly learnt a lot from it.

REFERENCES

[1] Bird Bones and Living Habits kernel description. https://jy2014.github.io/BirdBones/birdbones.html. Accessed: 2021-12-17.

[2] Dumont, E. R. Bone density and the lightweight skeletons of birds. *Proc. Royal Soc. B: Biol. Sci.* **277**, 2193–2198 (2010).

[3] Abdi, H. & Williams, L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2**, 433–459 (2010).

[4] Geron, A. Hands-on machine learning with scikit-learn, keras, and tensorflow, isbn: 978-1492032649 (2019).

[5] Quinlan, J. R. Induction of decision trees. *Mach. learning* **1**, 81–106 (1986).

[6] Vidal, R., Ma, Y. & Sastry, S. S. Principal component analysis. In *Generalized principal component analysis*, 25–62 (Springer, 2016).

[7] BenHamza, A. Statistical process and quality control. In *INSE 6220 course content*, Chapter 5 (Concordia University, Fall,2021).