Principal Component Analysis for Abalone Cultivation

INSE6220 : Advanced Statistical Approaches to Quality. Fall, 2021
Concordia Institure for Information System Engineering
Saidul Islam (ID : 40106992)

Abstract—While it is very difficult to analyze and represent a data with a large dimension, then the principal component analysis-PCA is a recognized method to reduce the dimension of any data, which turns the corelated data into uncorrelated data effectively. Here, this technique has applied to analyze data for classification of Abalone. Sorting out the proper Abalone for cultivation from a large quantity is very time consuming and complex task. The selection of Abalone can be considered as a classification task. In this report, different types of classifiers have applied and compared to find out the best classifier based on accuracy, while the Logistic regression classifier performed with the highest percentage of accuracy which is more than 80%. And we have able to sort out the right Abalone for cultivation successfully by reducing complexity of the data.

I. INTRODUCTION

Abalone is a sea snail that lives in coastal saltwater. It is distinct from other sea snail for its shape of shell, foot and mostly for usability. Abalone is a rear species of snail which is only found in the cold waters of New Zealand, Australia, south Africa, Japan and west coast of North America. On the other hand, Abalone is regarded as one of the best sea foods due its deliciousness. At the same time, it is considered as a sign of elegance in the dining table due its beautiful iridescent, polished shell and pearl carapace. As a result, demand of Abalone is very high all over the world over the years. The popularity of abalone leaded to overfishing and nearly brought this shellfish to extinction. Even, Abalone was listed on the endangered species list, and it was illegal to gather wild abalone from the oceans in many parts of the world[1]. Consequently, harvesting of abalone is only way to restrict its extinction. For cultivation, the selection of proper Abalone is very crucial and tough task at the same time. Because the criteria are associated with number of measure and specification like gender, length, height, weight and so on, which means lots of data[2].

While data is very valuable and considered as the effective form of storing and explaining any information, then the Data has been the most assessable material in few recent decades. It happened due to massive growth of internet users in each & every sector of our life, including social media platforms, online business, smart devices and so on. These data can be found in various types and numerous forms, while most of the data is unpredictable, irrelevant and sophisticated for the variety of sources and users. This phenomenon made the data representation and processing task more complex and challenging. At the same time, working with these kinds of data with high dimension and unnecessary feature is time consuming and inefficient as well. However, many features can be considered as outlandish and does not have impact on data. Multiple features may have the same information and similar significance on the result. Using these features all together as an input for machine learning classification algorithms can impact negatively on the output and affect the learning model severely. To resolve this issue, dealing with only the relevant data by removing unnecessary information from any large data will make the classification process more efficient, less time consuming with high accuracy. So, reduction of dimension of the data is the effective step of data processing before applying any kind of machine learning algorithm, as dimensional reduction helps us to deal with only the most significant uncorrelated information from a dataset[3].

To reduce dimensionality of data, two methods are applied effectively. While the one way is creating new features by developing a new combination of the actual features utilizing the dimension reduction methodology and creation of this new feature is done by feature extraction. Another way is called selection of feature and this method does not change the actual features of the data and recuperate a subset from data, while wrapper, filter and embedded methods are the classification of this feature selection[4]. In this practice, it is proven by plenty of research that, making the data into uncorrelated from the correlated data by retaining the relevant features provides us better result with high accuracy. While numerous dimension reduction techniques are available such as linear discriminant analysis (LDA)[5],Factor analysis (FA)[6] and principal component analysis (PCA) to improve the computational performance, then selection of the dimension reduction method is very crucial [7].

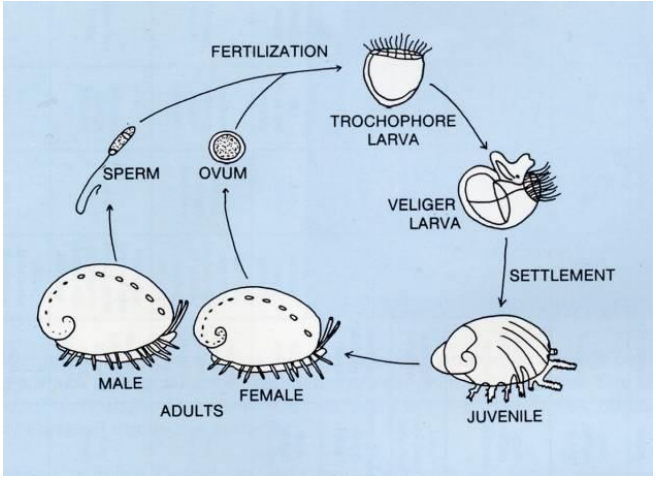

Fig. 1: Abalone Cultivation Process.

In this paper one of the most popular and effective method-principal components (PCA) has applied on Abalone dataset to distinguish between the Abalone which are right for cultivation, and which are not. The Fig. 1 above shows the cultivation stages and life cycle of the Abalone. First, we did some preprocessing of our data (i.e., makes all data in numeric values, remove class & etc.) and we applied PCA on it to process and reduce the dimensionality of our data. Then two types of classification algorithms- Logistic Regression and K-Nearest Neighbor (KNN) has applied to measure the conversion of our data.

Remaining part of this report is organized as follows: Section II – Detailed discussion about PCA, Section III – Introduction of classification algorithms, Section IV –Report and discussion about the results from classification, Section V – Discussion about the results from classification and Section VI – conclusion of this paper.

II. PRINCIPAL COMPONENT ANALYSIS(PCA)

Principal Component Analysis – PCA is an effective and one of well accepted technique by the scholars for simplifying complex data. It reduces the dimensionality of large data sets by transforming a large set of variables into a smaller number, but it retains most of the information of the original large dataset. PCA is a simple implementation of the true eigen vector based multivariate analyses. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space, PCA can provide us with a lower-dimensional better visualization.

PCA is an orthogonal linear transformation of feature vectors into uncorrelated vectors. The greatest variance by some projection of the data holds by the first coordinate, which called the first principal component. It fixes the direction of most variability in the data. The amount of variability has a directly proportional relationship with the amount of information carries by the component. So, the first principal component-PC has the highest variability, that means it carries the highest amount of information. And the 1st PC supposed to be a straight line in a very close position to the data points. That means, a minimum the sum of square distance exits between data point and the line. Similarly, the second greatest variance on the second coordinate and we get the second principal component. The 2nd PC captures the remaining variance of the dataset those were not carried by the 1st principal component. That means, there is no correlation between the principal components to each other. By this process, we get the more of uncorrelated principal component and PCA turn our dataset uncorrelated [7] [8].

III. CLASSIFICATION ALGORITHOMS

One of the significant tasks of machine learning algorithms is to recognize objects and being able to separate them into categories and this process is called classification. There three types of machine learning are mostly practiced, those are- supervised learning, unsupervised learning and reinforcement learning. For the classification, supervised and unsupervised learning from the machine learning can be used. While the supervised learning algorithms labels the input and the algorithm work to label output data. Then, the labeled data is used to produce more output for future query. On the other hand, Unsupervised learning patterns inferred from the unlabeled input data. Unsupervised learning finds out the structure and patterns from the input data. It does not need any supervision, even it searches and finds out patterns from the data by its own. In this paper, we used labeled dataset for classification. So, we will work with supervised learning classifiers which will distinct the Accepted & Rejected Abalone. After doing plenty of background study and practice of various algorithms for this dataset, we decided to apply Logistic Regression and K-Nearest neighbor [9][10].

A. Logistic Regression

If we talk about the supervised learning for binary classification, then Logistic Regression appears as a famous machine learning algorithm. Because Logistic Regression is an effective and one of simplest algorithms for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems. Instead of fitting a straight line or hyperplane, the logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. Logistic Regression learns a model that predicts the probability that an observation belongs to a certain class, using the sigmoid function[11]. The Logistic Function is as below:

$$Y(p) = \frac{1}{\left(1 + exp^{-\theta^T x}\right)} \tag{1}$$

Where P is the probability (the mean of Y), e is the base of the natural logarithm and x and θ are the parameters of the model in Eq. (1).The value of a yields P when x is zero, and θ adjusts how quickly the probability changes with changing x a single unit. Means, we can standardize and unstandardized θ weights in logistic regression, just as in ordinary linear regression.

B. K-Nearest Neighbor (KNN)

One of the most frequently used classifiers that does a reasonable job is called K-Nearest Neighbors (KNN) Classifier. Like other classifiers, the KNN classifier estimates the conditional distribution of Y given by X. After that, it classifies the observation to the class with the highest estimated probability. There is a positive integer K and a test observation of the classifier that identifies the K points in the data those are closest to x_0. For example, if K equals to 5, then the five closest observations to observation x_0 will be identified. Here, n_0 represents these points and then The KNN classifier calculates the probability under conditions for class j. This is because, the fraction of points in observations in n_0 response equals j, which is showed by the Eq. (2) below.

$$\Pr(Y = j \mid X = x_0) = \frac{1}{K} \sum_{i \in n_0} I(y_i = j) \tag{2}$$

Moreover, altering K produces dramatically different results. For instance, if $K = 1$, the decision boundary is minimally restricted and the KNN models are supposed to deliver a low bias but dramatically high variance. If we increase the value of K, the flexibility of the classifier will be reduced, and the decision boundary will be gradually closer to the linear. Thus, these models provide low variance with high bias[12].

IV. Experimental Results

A. *Dataset Description*

Abalone is regarded as one of the best and expensive sea foods in the world. While this seasnail is in a threat of being endangered, then farming is the only way to restrict extinction and mitigate the high demand. Detect the right Abalone automatically for cultivation and make it classified is a tough task. But it plays a crucial role and helps farmers in Abalone cultivation significantly. We have chosen the Abalone dataset to apply and study effectiveness of the PCA to reduce its dimension. While this dataset has 4177 observations, then it has divided into two groups based on measurements and 2081 labeled as "Accepted" and 2096 as "Rejected" during the analysis that shown by Fig. 2. An Abalone is associated with several measurements of different parts of its body. The main variables are (i) sex [male, female, infant]; (ii) lengths; (iii) diameter; (iv) height; (v) weight of whole and different parts (shucked, viscera & shell) of body.

Fig. 2: Abalone Classes.

The box plot depicts the data after we normalize. The box plot helps to observe the distribution of the data and variability of the data by Fig. 3 below. The data those are out of upper control limit-UCL and Lower control limit-LCL can be found out from box plot. Some outliers can be seen for different features. There are related to type of the data and different measurements, not from any kind of error. That's why, we decided to retain all data points for PCA analysis.

The covariance matrix helps to explicit the correlation between different parameters. The covariance matrix shows us the covariance matrix of the data in Fig. 4. From the matrix, it is obvious that "sex" is highly corelated with all other variables and it is a strong negative correlation. On the contrary, the highest positive correlation exists between "length" and "diameter". And a further strong positive correlation can be observed between "whole weight" and both "shucked weight" & "viscera weight".

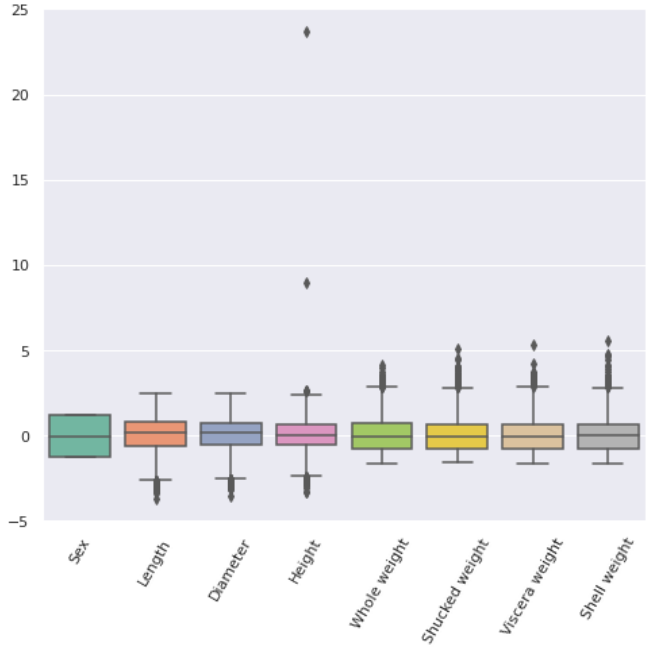

Fig. 3: Boxplot of Centered feature Vectors.

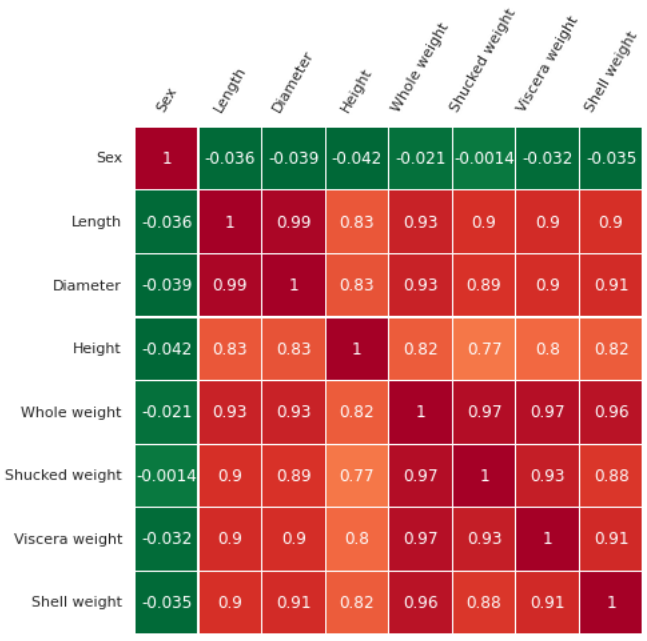

Fig. 4: Covariance Matrix.

In addition, the correlation between the parameters is more transparent by the Pair plot in Fig. 5. The negative correlation between sex and all other variables are clear and obvious from the visualization of pair plot. And the shape of the variables tells more that all kinds of weights have a nearly similar amount of positive correlation with length. Moreover, the length and diameter showed nearly same shape because both are very strongly correlated in positive way.

dataset: https://www.openml.org/d/720

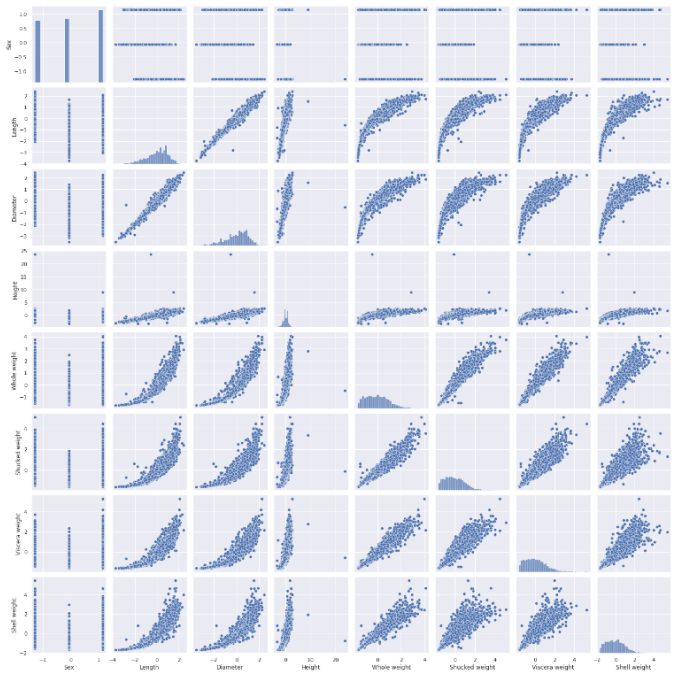

Fig. 5: Pair plot.

B. PCA for dimensionality reduction

After observation of our data, we applied PCA on the dataset and the purpose of this implementation is to reduce the dimensionality of the data from $p = 8$ vectors to r feature vectors $r < 8$ (where $r << p$ always). The number of components can be determined by the scree plot Fig. 6. And it can be observed once more by the Pareto Diagram Fig. 7. We can examine from both scree plot and pareto diagram that 91.96% of the variance can be explained by the two first components, those are 79.45 and 12.51. Therefore, the optimal component number for this dataset is chosen as r = 2 and reduced the complexity of the data by retaining the first two principal components (PC).
We applied the following four steps for PCA on our Abalone dataset[13][14]:

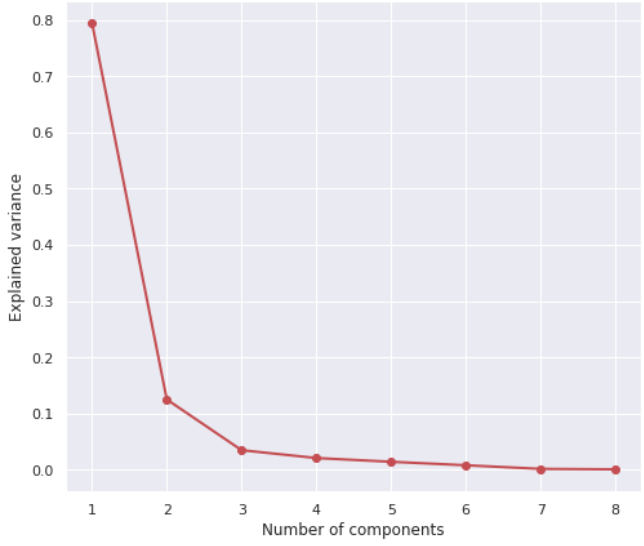

Fig. 6: Scree plot.

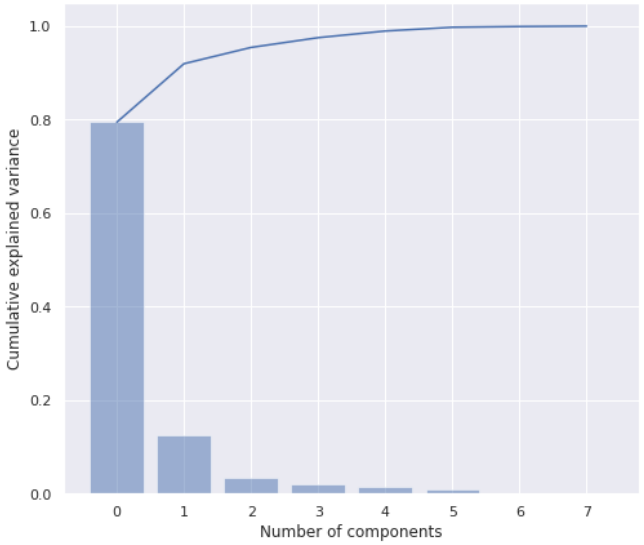

Fig. 7: Pareto Diagram.

i) Standardize the data by centralization

First, we need to standardize the dataset. For that, we need to calculate the mean of each column and subtracting that from every value. That means, compute the centered data matrix by subtracting off-column means. The centered data matrix represented by Eq. (3) below.

$$Y = HX \tag{3}$$

ii) Calculate the Covariance Matrix for whole dataset

In the second step of PCA, we find out how the variables of the input data set are varying from the mean with respect to each other. That means, we identify the correlation between variables. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, in order to identify these correlations, we compute the covariance matrix (S). The below Eq. (4) represents the covariance matrix of whole data.

$$S = \left(\frac{1}{n-1}\right) Y'Y \tag{4}$$

iii) Calculate eigenvalues and eigen vectors

In this third step of PCA, we calculate the eigenvectors and eigenvalues from the covariance matrix in order to determine the principal components of the data. The eigenvectors and eigenvalues are the linear algebra concepts. An eigenvector is a nonzero vector which changes at most by a scalar factor when that linear transformation is applied to it. The corresponding eigenvalue is the factor by which the eigenvector is scaled. The eigenvectors are in directions of the data that explain a maximal amount of variance, to the lines or axes that capture most information of the data. The relationship between variance and information is directly proportional. It means the larger the variance carried by an axis, the more the information it carries. The Eq. (5) is as follow.

$$S = A\Lambda A' = \sum_{j=1}^{p} \lambda_j a_j a'_j \tag{5}$$

iv) Principal Components (PC)

In the fourth step of PCA, we calculated the transformed data matrix (Z) of size $n \; x \; p$, shown in Eq. (6). The columns represent the principal component (PC) then the rows are observations. The number of principal components is similar number of original variable and r will be always less than P.

$$Z = YA \tag{6}$$

We found the eigenvector matrix as below:

$$A = \begin{bmatrix} -0.015 & -0.999 & 0.034 & -0.014 & -0.016 & -0.019 & 0.001 & -0.001 \\ 0.383 & 0.002 & 0.036 & -0.594 & 0.087 & -0.042 & -0.7 & -0.024 \\ 0.384 & 0.005 & 0.063 & -0.586 & 0.006 & -0.01 & 0.711 & 0.016 \\ 0.348 & 0.017 & 0.868 & 0.312 & 0.164 & 0.027 & -0.01 & 0.001 \\ 0.391 & -0.017 & -0.232 & 0.231 & -0.052 & 0.111 & 0.021 & -0.851 \\ 0.378 & -0.039 & -0.344 & 0.229 & 0.495 & 0.549 & 0.011 & 0.371 \\ 0.381 & -0.005 & -0.253 & 0.272 & 0.147 & -0.807 & 0.024 & 0.205 \\ 0.379 & 0.0 & -0.06 & 0.165 & -0.834 & 0.177 & -0.06 & 0.307 \end{bmatrix}$$

The first component can be found as Eq. (7). While it can be observed that X_5 (whole weight), X_3 (diameter) and X_2 (length) have the most contribution of the first component.

$$Z_1 = 0.383X_2 + 0.384X_3 + 0.348X_4 + 0.391X_5 + 0.378X_6 + 0.381X_7 + 0.379X_8 \tag{7}$$

Likewise, the second component is as below Eq (8):

$$Z_2 = -0.999X_1 \tag{8}$$

It is obvious that only the X_1(sex) contributes most to the second component.

The vector below expresses the eigenvalues in Eq (9).

$$\lambda = \begin{bmatrix} 79.45 \\ 12.51 \\ 3.48 \\ 2.09 \\ 1.42 \\ 0.80 \\ 0.16 \\ 0.08 \end{bmatrix} \tag{9}$$

Two largest values of this eigenvalue of correlation matrix and their associated eigenvectors helps us to find out the principal components.

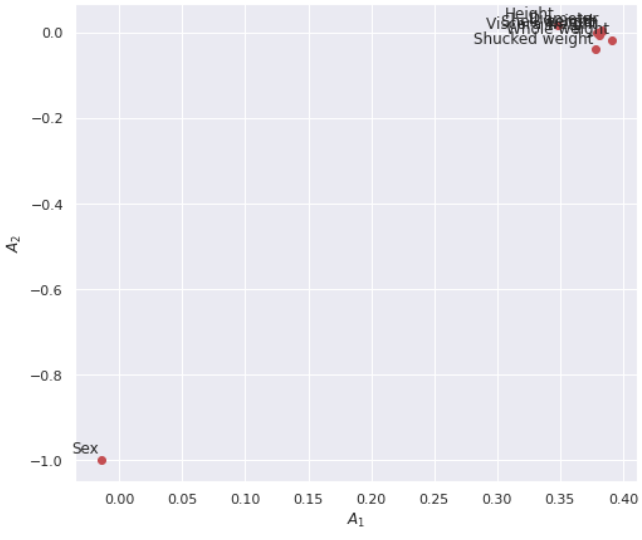

Fig. 8: Scatter Plot.

The scatter plot of PC2 coefficients vs PC1 coefficients is shown in Fig. 8. This plot helps understand which variables have a similar involvement within PCs. As can be seen from Fig. 8 that the variable sex is located at the bottom left of the plot (-1.0) and all other variables are location in the top right of the plot, which is highly consistent with the values of the coefficients of the Z_2 and Z_1 , that we have already shown in Eq. (7) & (8). This is consistent with the values of the coefficients of PC1 and PC2. The height, diameter, whole weight and other variable related to weight have almost the same involvement within the principal components. In Fig. 9, we have already shown that these variables are highly correlated.

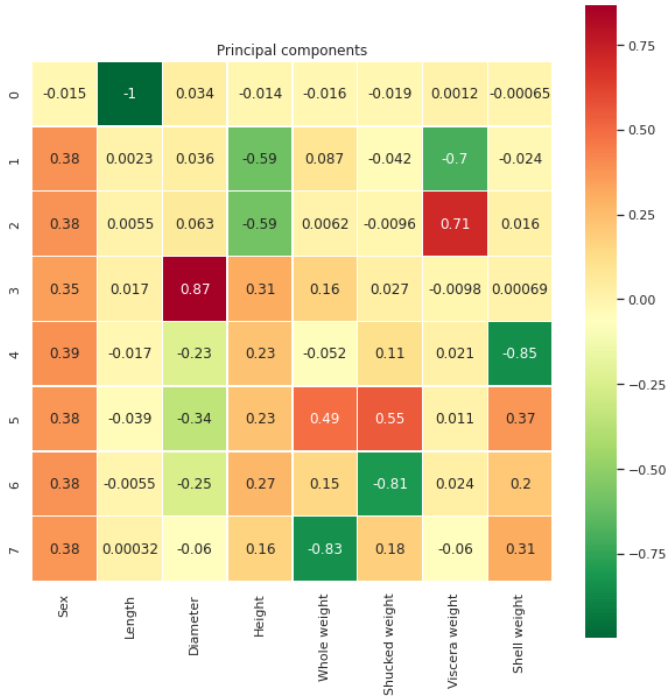

Fig. 9: PCs Components

The biplot in Fig. 10 helps visualize both the principal component coefficients for each variable and the principal component scores for each observation in a single plot. The axes in the biplot represent the principal components (columns of eigenvector A), and the observed variables (rows of eigen vector A) are represented as vectors. Each observation (row of Z) is represented as a point in the biplot. The color of point gives indication of classes associated to it.

It is obvious from the biplot that the first component has a positive coefficient for length, diameter, height and various weight variables with almost the same value, which we can see from in Eq. (7) as well. These seven variables are nearly equally contributed variables for the first component. And the angle between these variables and the first component is very narrow, which is explicit by Fig. 10. The second component has negative coefficient for sex variable and that can be seen also in Eq. (8) and it is the only and largest contributor variable in the second component, as this variable has the longest arrow and that is clear by the biplot as well. It indicates that these components distinguish between observations that have high values for the first set of variables and low for the second, and observations that have the opposite.

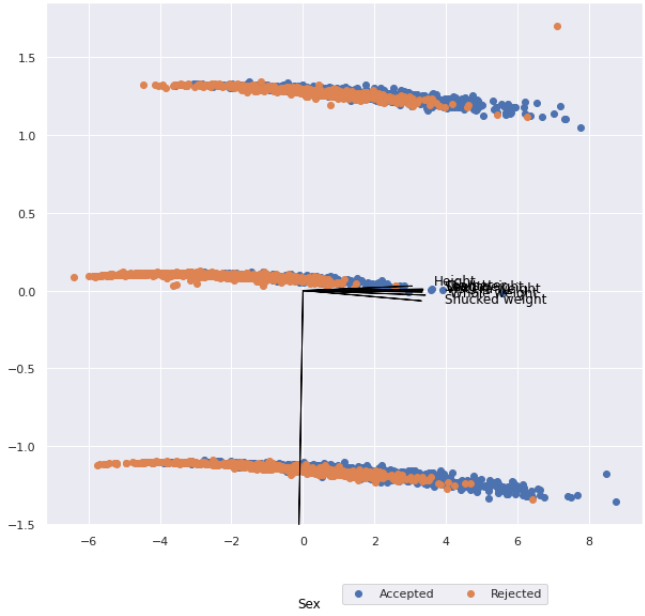

Fig. 10: Biplot.

In this biplot, the positions of points mean the score of each observation for the two principal components. It can be easily observed that the first component in the horizontal axis has a higher variation (most of the variables) than the second component. The color of the data point indicates whether it is accepted or rejected. Each of the 4177 examples is represented in this plot by a point, and their locations indicate the score of each observation for the two principal components in the plot. Though both accepted(blue) and rejected(orange) score points difference is not so huge, while the rejected points near the left of this plot have the lowest scores for the first principal component, they correspond to accepted abalone. These are plotted in accordance with their scores on the first two PCs. It is obvious that accepted abalone spreads far to the right in the direction of ray for the height, length, diameter and all kinds of weight variable. It means that accepted abalones have a better height, length, diameter and weight compared to rejected abalones.

C. PCA Control Chart

The principal Component control chart studied to find out the outliers and for the first component. This control chart is measured by the equations below Eq (10).

$$\begin{aligned} UCL &= 3\sqrt{\lambda_j} \\ CL &= 0 \\ LCL &= -3\sqrt{\lambda_j} \end{aligned} \tag{10}$$

The control chart below is showed by Fig. 11 and the control chart depicted that all the data was within upper control limit-UCL & lower control limit-LCL, and It means that no point is out of control.

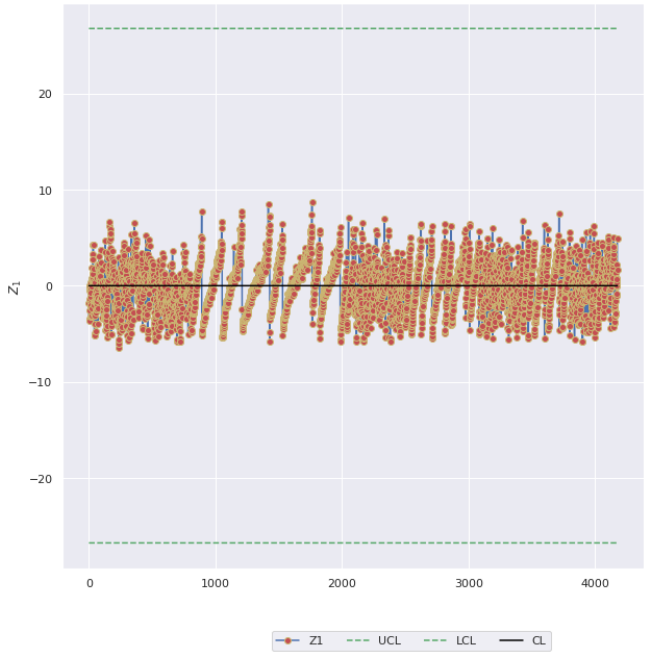

Fig. 11: Control Chart.

V. CLASSIFICATION

In this part of the report, two different classification algorithms will be applied on three sets of data- (i) original Abalone full data, (ii) transformed features (Z) and (iii) first two components (Z_{12}). At the end, we will compare the results and analyze them based on accuracy and fit time. Here, the goal is to assess the impact of PCA on two different classifiers and compare the performance of the regarding models. There are six experiments have been conducted as below:

- Logistic Regression:
  - Full dataset
  - All principal components (Z)
  - First two components (Z_{12})
- K-Nearest Neighbor (KNN)
  - Full dataset
  - All principal components (Z)
  - First two components (Z_{12})

The classifier evaluation is very dependable on the type and pattern of the data that we are working on. Here, we will use the accuracy and the fit time as a parameter of evaluation. We compared different models' complexity and evaluated with 5-fold cross validation scheme. We choose 5-fold cross validation because of the size of our dataset, which is not so huge consisting of 4177 observations.

It is very clear from the provided TABLE I and TABLE II that, we got similar performance for logistic regression from original data and first two component of PCA, while logistic regression performed better for original full dataset and all PCs component than first two component of PCA, but the fit time is highest for original dataset and lowest for the first two PC. On the other hand, the K-Nearest Neighbor classification shows similar characteristics as logistic regression. It means,

KNN performed same and provided best accuracy for all dataset and all components of PCA and lowest accuracy for first two components of PC, whereas the least amount of fit time takes for first two PCs.

By comparing the both classifiers, we can observe that the Logistic regression better than the K-Nearest Neighbor in overall cases and the highest more than 80% accuracy has found from the logistic regression in all components of PCA. To perceive the difference of performance between two classifiers, we visualized the performance of All components of PCA by a bar diagram in Fig. 12.

TABLE I: 5-Fold Cross validation accuracy.

Fold	Accuracy					
	Logistic Regression			KNeighbor		
	Original data	All PCs	First 2-PCs	Original data	All PCs	First 2-PCs
1	0.81	0.81	0.76	0.76	0.76	0.72
2	0.79	0.79	0.71	0.75	0.75	0.73
3	0.78	0.78	0.74	0.77	0.77	0.71
4	0.79	0.79	0.71	0.77	0.77	0.72
5	0.78	0.77	0.70	0.75	0.75	0.72
Avg	0.79	0.79	0.73	0.76	0.76	0.72

TABLE II: 5-Fold Cross validation Fit time.

Fold	Fit time					
	Logistic Regression			KNeighbor		
	Original data	All PCs	First 2-PCs	Original data	All PCs	First 2-PCs
1	0.074	0.089	0.025	0.009	0.005	0.003
2	0.101	0.082	0.049	0.011	0.006	0.003
3	0.089	0.071	0.020	0.010	0.005	0.003
4	0.097	0.092	0.017	0.010	0.006	0.003
5	0.081	0.091	0.011	0.009	0.005	0.004
Avg	0.088	0.084	0.024	0.009	0.005	0.003

We have generated the Receiver operating characteristics- ROC curve to analysis more our perceived results for both classifiers. As the diagnostic capability of binary classified in varied, so the ROC curve is used to depicts the capability by a graphical plot, where the false positive rate is fixed, and true positive rates appears by corresponding average[15]. We can see from the ROC curve of both classifier that logistic regression classifier covers area under the curve- AUC of 0.81 in Fig. 13 and the K-Neighbor covers the area under the curve-AUC of 0.78 in Fig. 14. So, we can conclude from the ROC curve that performance of Logistic Regression outweighed the performance of K-Nearest neighbor for our Abalone data.

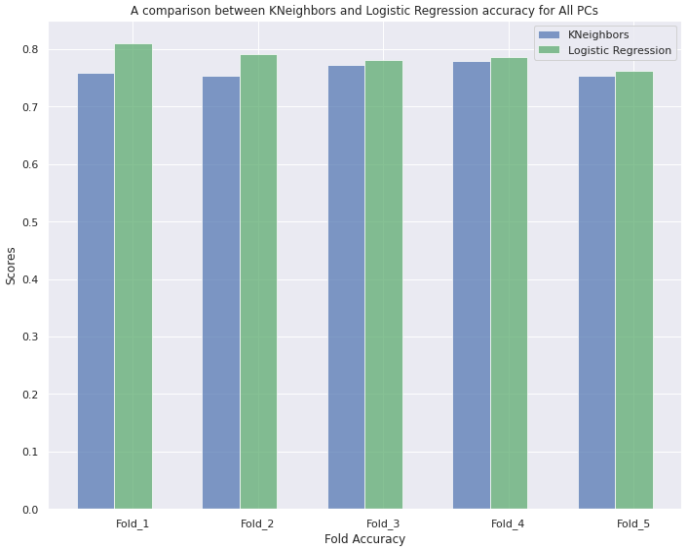

Fig. 12: A comparison between KNeighbor (KNN) and Logistic Regression performance for all PCs.

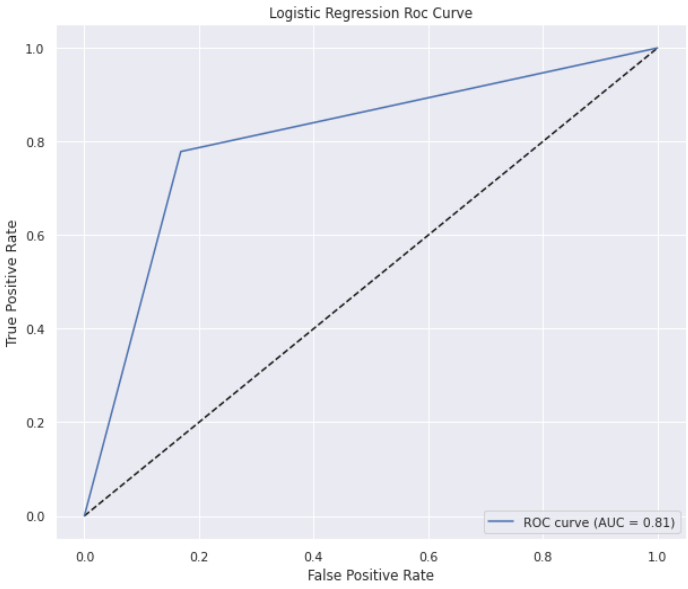

Fig. 13: Logistic Regression ROC curve.

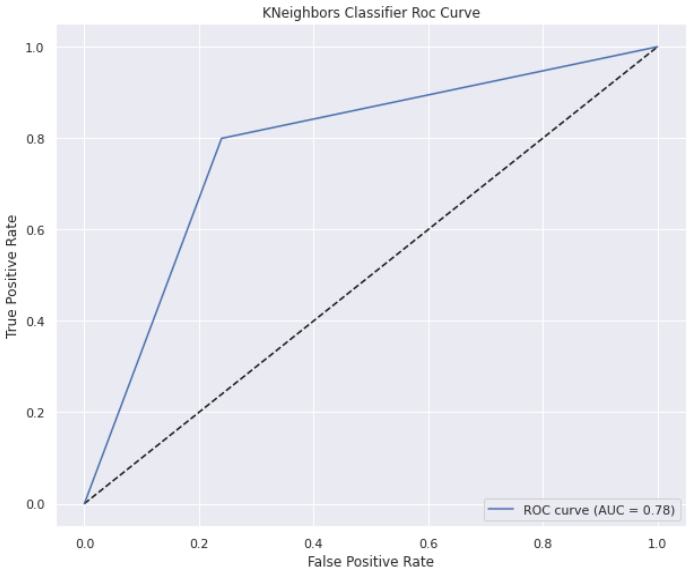

Fig. 14: K-Nearest Neighbor ROC curve.

VI. CONCLUTION

To recapitulate, in this report we applied Principal Component Analysis on Abalone for cultivation, and we utilized the results to analyze the impact of PCA on different classifiers. The paper is divided into two parts, while PCA has applied on a dataset of 4177 abalone. We have found 91.95 % of the explained variance in the first two components of PCA. As a result, we decided to deal with the first two components. Studying the first component's control chart, we found that no data points was out of control. On the second part, we used the resulted data to build six different classifiers. The classifier should determine whether an Abalone is accepted or rejected for cultivation. The performance of each classifier is measured by the fit time and accuracy. Considering both parameters, we found out that Logistic Regression performed better than the K-Nearest Neighbor Classifier, while both classifiers performed better with all principal components of PCA than first two components, and a same accuracy as all PC components was observed with full data from the classifiers, though the fit time was considerably less for all PCs. Finally, we successfully discriminated accepted abalone from rejected ones for cultivation with a high accuracy (more than 80%) for Logistic Regression.

REFERENCES

[1] P. A. Cook, "*The Worldwide Abalone Industry*," Modern Economy, vol. Vol.05, no. 13, pp. 1-5, 2014.

[2] J. R. A. S. M. M. A.-Z. B. S. A.-N. &. S. S. A.-N. Ghaida Riyad Mohammed, "*Predicting the Age of Abalone from Physical Measurements Using Artificial Neural Network*," International Journal of Academic and Applied Research (IJAAR), pp. 7-12,2020.

[3] E. P. a. J. v. d. H. Laurens van der Maaten, "*Dimensionality Reduction: A Comparative*," Tilburg centre for Creative Computing, 5000 LE Tilburg, The Netherlands, 2009.

[4] K. B. a. N. A. Jovic, "*A review of feature selection methods with applications*," 38th international convention on information and communication technology,electronics and microelectronics (MIPRO)., vol. , no. IEEE, pp. 1200-1205, 2015.

[5] A. G. T. I. A. A. E. Tharwat, "*Linear discriminant analysis: A detailed tutorial*," AI Communications, vol. 30, no. 3, pp. 169-190, 2017.

[6] S. P. e. a. A.G Yong, "A beinner guide to factor analysis: Fosucing on exploratory factor analysis," Tutorial in quantitative methods for psychology,, vol. 9, no. 2, pp. 79-94,2013.

[7] S. M. A. A. M. M. Z. A. H. Sasan Karamizadeh, "*An Overview of Principal Component Analysis*," Journal of Signal and Information Processing, vol. 4, no. 1, pp. 173-175, 2013.

[8] I. Jollffe, "*Principal Component Analysis*," Technometrics, vol. 45, no. 3, p. 276, 2003.

[9] R. Feldman, "*Technique and applications for sentiment analysis*," Commnications of the ACM, vol. 56, no. 4, pp. 82-89, 2013.

[10] R. Xu and D. Yadav, "*Survey of Clustering algorithms*," IEEE Transaction on Neural networks, vol. 16, no. 3, pp. 654-678, 2005.

[11] K. L. L. G. M. I. Chao-Ying Joanne Peng, "*An Introduction to Logistic Regression Analysis and Reporting*," The Journal of Educational Research, vol. 2, no. 1, pp. 3-14, 2010.

[12] D. K. T. N. Suguna, "*An Improved k-Nearest Neighbor Classification Using*," IJCSI International Journal of Computer Science Issues, vol. 7, no. 4, pp. 18-21, 2010.

[13] A. Ben Hamza, "*Advanced statistical approaches to quality*" Unpublished.

[14] Ringnér, M. "*What is principal component analysis?*". Nat Biotechnol. Vol.26, pp. 303–304, 2008.

[15] S. U. A. W. Jerome Fan, "*Understanding receiver operating characteristic (ROC) curves*," Cambridge University Press, Ontario, 2015.