Principal Component Analysis and Multi-class Classification on two varieties of Raisins

Anurag Bhaskar, 40196370
Github Link: https://github.com/a-haska/Anurag6220

***Abstract*— In this report, advanced statistical quality control techniques have been applied on a data set using machine learning concepts in Python. Basically, Principal component analysis (PCA) along with multi-class classification is used on raisins data for the raisins grown in Turkey, in order to differentiate between the quality and purity of two types of raisins. Total of 900 observations for each type of raising has been obtained which are present in the dataset. PCA is an advanced statistical method which is used for dimensionality reduction of a data, where a data set with correlated variables or columns are transformed into data with uncorrelated variables at the same time maintaining the variability. After applying four types of classifiers namely Logistic Regression, Decision Tree, K Neighbors and Random Forest, best performing model is evaluated to be Logistic Regression. Also evaluating the best model with PC components turns out to be K Neighbors model.**

***Keywords*— Principal component analysis (PCA), Machine Learning models, multi-class classification, Logistic Regression, Decision Tree, K Neighbors, Random Forest, Raisin types.**

I. INTRODUCTION

Grape fruit is nutritious source of antioxidants, iron, potassium and fiber. Dried grapes which become raisin contain the same properties with addition of being eaten as snacks. Turkey being the one of the pioneer countries in producing grapes can utilise various methods and techniques to know the different variety of raisins which can help in the determining the quality and features of the product. Although old methods based on human judgement can be used, but these methods have proven to be incorrect at various times depending on several factors. [1] So we have utilised advanced statistical techniques to know the variety of raisins using machine learning models. It can help in the field of agriculture and food processing while maintain the standard desired quality of products.

In this report we have used Python programming in order to apply the techniques. First of all, Principal Component Analysis (PCA) is applied to the multi-class data set of raisins to distinguish between two types of raisins called 'Kecimen' and 'Besni'. After the PCA application, four machine learning algorithms have been applied to the transformed data. We have also applied these classifiers to the original data.

Rest of the paper is written in following order: Section II gives the description and analysis of PCA, Section III introduces machine learning algorithms used on the data, Section IV presents the data set, Section V reports and explains PCA results Section VI discusses the classification algorithm results, Section VII describes Explained AI with Shapley values and Section VIII concludes the report.

II. PRINCIPAL COMPONENT ANALYSIS

PCA is an advanced statistical technique which is used for dimensionality reduction of a data, where a data set with correlated variables or columns are transformed into data with uncorrelated variables at the same time maintaining the variability. In simpler terms PCA is a data reduction technique. It is an exploratory multivariate technique which simplifies complex data sets (generally coming from real world scenarios) to a simpler and smaller data set.

PCA essentially finds new set of variables called Principal components (PCs) where it transforms correlated high dimensional data to uncorrelated data with reduced dimensions without hampering the variability of the original data. The motivation behind doing PCA can be due to many reasons. For example, it saves time and space thereby saving more costs, robust nature of model due to simplicity, explanation and understanding becomes easy due to less variables and data can be visualized as well. [2]. The first PC obtained in the PCA reduces the distance between the data and the projection of data onto the PC, which means highest variance in the data is covered by the first PC. The second, third and so on principal components also decrease the distance as first PC, at the same time they become uncorrelated to other PCs obtained before them. [3]

- **PCA Algorithm**

PCA algorithm is designed with four steps with a given data matrix X having n rows and p columns (nXp).

Step 1: Centre the data

To get the centred data matrix, we subtract each column mean from all the respective column elements which is represented by:

$$Y = H\,X$$

Step 2: Compute covariance matrix S

After this covariance matrix represented by S with size of p rows and p columns (pXp) is calculated from the centred data which is given by:

$$S = \frac{1}{N-1} Y'Y$$

Step 3: Eigen vectors and eigen values of covariance matrix S are calculated using eigen-decomposition.

$$S = A \wedge A' = \sum_{j=1}^{p} \lambda_j \, a_j a'_j$$

Step 4: Lastly, transformed data matrix is computed with nXp size.

$$Z= YA$$

$$Z = \left(z'_1, z'_2, \dots.., z'_i, \dots\dots, z'_p \right) = \begin{pmatrix} z_{11} & z_{12} \cdots & z_{1j} & z_{1p} \\ \vdots & & & \vdots \\ z_{n1} & z_{n2} \cdots & z_{nj} & z_{np} \end{pmatrix}$$

III. CLSSIFICATION ALGORITHMS

With the advancements in science and technology, several machine learning methods have been used for analysing data in agriculture advancements. In machine learning there are mainly two technologies known as supervised and un-supervised. In this report we will be using the supervised machine learning algorithm on the raisin dataset. The supervised machine learning is pre-categorized and is of numerical nature where the model learns from the training data. This technique is classified further in the two parts known as classification and regression. Classification means separating data into pre-defined classes having their own labels. We will be using four classification algorithms namely Logistic Regression, Decision Tree, K Neighbors and Random Forest.

A. Logistic Regression

Logistic regression is a type of supervised machine learning algorithm which is used to predict the probability of a dependent variable based on an independent input variable. The simplest results of this algorithm can be binary that is it can be in 0 or 1, yes or no, true or false etc. However, for multinomial classification it can have 3 or more values. The output of a dependent variable which is of categorical nature is predicted by this algorithm, so instead of giving values between 0 and 1, it gives probable values lying between 0 and 1. Logistic function is used in this classifier. [4][5]

B. K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a simple and easy to implement algorithm. It is a supervised machine learning algorithm which is non-parametric and also referred to a lazy learning algorithm as it stores all the data related to training data in an n-dimension. This algorithm expects similar things to be near to each other. When this model gets any data value to be predicted, it looks for the 'k' nearest numbers of data saved and returns the average value of the k-nearest neighbors. Also predicts the commonest class for the nearest data. [2] [6]

C. Decision Tree

Decision tree is also a non-parametric supervised machine learning algorithm. It builds classification models in the form of a tree structure for making a decision. This algorithm uses if-then rule to predict a value of a variable in target by learning simple rules set for prediction. It makes sequential decision for predicting the value of a target variable anf thereby removing the tuples whose decision has been made in the sequence. This process ends once the last condition is met in the training data.[2]

D. Random Forest

Random forest algorithm is a set of number of decision trees discussed above. These decision trees together are often referred as ensemble of trees which are trained using bagging method. Each decision tree works in the Random Forest algorithm and predicts a class. Based on all the class predicted, the class having the most votes is selected to be the prediction for the whole model. Although this algorithm can be used for regression, this can give good prediction on the classification.

IV. DATASET DESCRIPTION

This data set is taken for this project is based on two different types of raisin grains made of grapes grown in Turkey. The data from where it is sourced had taken these two types of raisins and took images for image-processing. The results of image processing techniques gave total of 7 characteristic features for the raising types. There are total of 900 grains of each raisin types Besni and Kecimen has been taken.[1] The equal number of data observation is shown in the pie chart below in Fig 1.

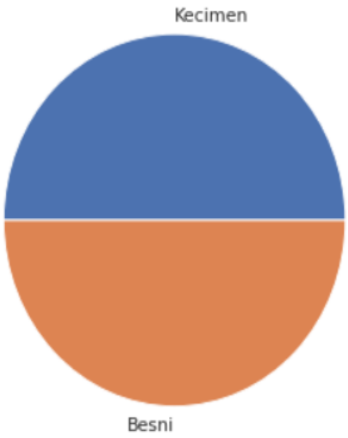


Fig. 1: Pie chart for equal data of raisin variety

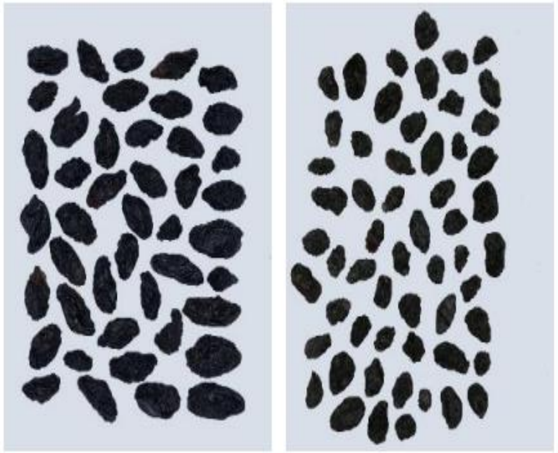


Fig. 2: Sample images of two raisin types used in the study, Left image (Besni), Right Image (Kecimen) [1]

The above figure 2 gives sample images of the raisin types Besni and Kecimen.

Figure 3 below shows the box and whisker plot for the data. Box plots are basically representation of quantitative data with quartiles and whiskers and five number summary. It also helps in showing the outliers present in the data. Here as seen in the Fig. 3 below almost all the variables have outliers present with ecentricity and extent having maximum number of outliers. Also data is either left or right skewed in all the cases except for extent.

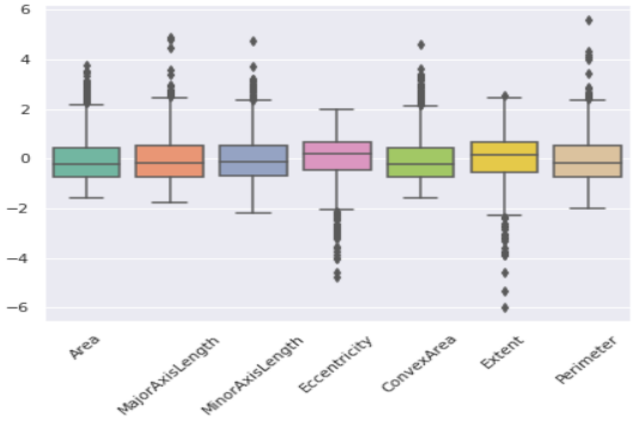


Fig. 3: Box and Whisker plot of centred feature vectors

Below Fig. 4 shows swarm plot or strip plot. Swarm plots are an extension of the box plot where they show all the data points on top of the boxes along with their distribution. The data is normally distributed however there are some outliers existing.

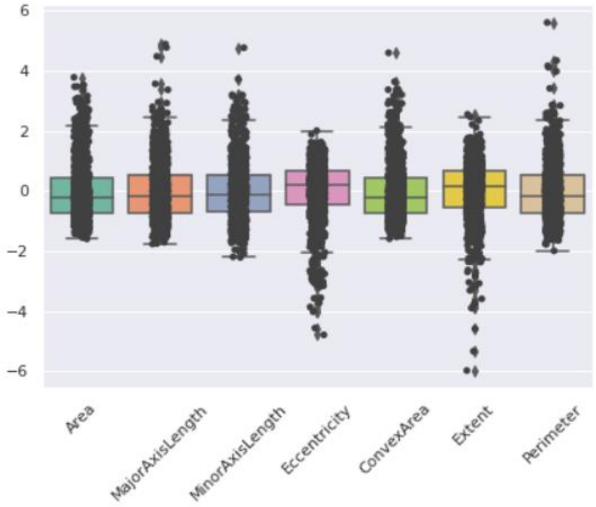
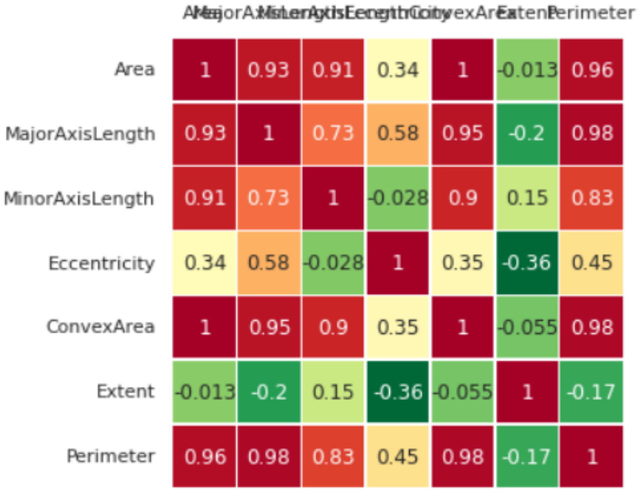


Fig. 4 Swarm Plot to show datapoints on top of the boxes



Area	1	0.93	0.91	0.34	1	-0.013	0.96
MajorAxisLength	0.93	1	0.73	0.58	0.95	-0.2	0.98
MinorAxisLength	0.91	0.73	1	-0.028	0.9	0.15	0.83
Eccentricity	0.34	0.58	-0.028	1	0.35	-0.36	0.45
ConvexArea	1	0.95	0.9	0.35	1	-0.055	0.98
Extent	-0.013	-0.2	0.15	-0.36	-0.055	1	-0.17
Perimeter	0.96	0.98	0.83	0.45	0.98	-0.17	1

Fig. 5 Covariance Matrix

In the figure above covariance matrix has been shown. A covariance matrix shows the relation between variables of the dataset. If one variable with high value corresponds with the high value of other variable or low value of one variable corresponds to the low value of another variabl then the covariance is said to be positive. Whereas if the low value of a variables is matching with the high value of another variable then the covariance is negative. The diagonal values happen to be one becauase of the self corelation of a variable. [7] In Fig. 5, area is positively related with Major Axis length, minor axis length, convex area and perimeter. Similarly major axis length is positively related with perimeter. On the other hand area and extent are negatively correlated.

V. PCA RESULTS

We have applied the PCA on the raisin data set and got the results as discussed follows. The dimensions or the columns were reduced from the original data set post the application of PCA as discussed above in the PCA analysis step. Original data set with 7 variables is reduced to 3 variables or the principal components (PCs) which is the reduced dimension. Previous data which was of the size n x p has reduced to a smaller size of the Eigen vector matrix denoted by A.

Eigen vector matrix A is given as below:

$$A=\begin{bmatrix} 0.4482 & -0.1160 & 0.0054 & -0.1111 & -0.6110 & -0.0998 & -0.6243 \\ 0.3893 & -0.3749 & 0.2360 & -0.6558 & 0.3845 & -0.2390 & 0.1299 \\ 0.4432 & 0.1365 & -0.1005 & 0.4952 & 0.0875 & -0.6855 & 0.2277 \\ 0.2029 & 0.6108 & -0.6285 & -0.4262 & 0.0751 & 0.0535 & 0.0204 \\ 0.4509 & -0.0876 & 0.0366 & 0.05581 & -0.3924 & 0.4712 & 0.6391 \\ -0.0563 & -0.6673 & -0.7319 & 0.1090 & 0.0568 & 0.0234 & -0.0016 \\ 0.4508 & 0.0341 & 0.0443 & 0.3398 & 0.5551 & 0.4872 & -0.3639 \end{bmatrix}$$

The below matrix represents the Eigen value matrix which is denoted by lambda (λ).

$$\lambda = \begin{bmatrix} 4.8376 \\ 1.4548 \\ 0.6291 \\ 0.0568 \\ 0.0218 \\ 0.0064 \\ 0.0010 \end{bmatrix}$$

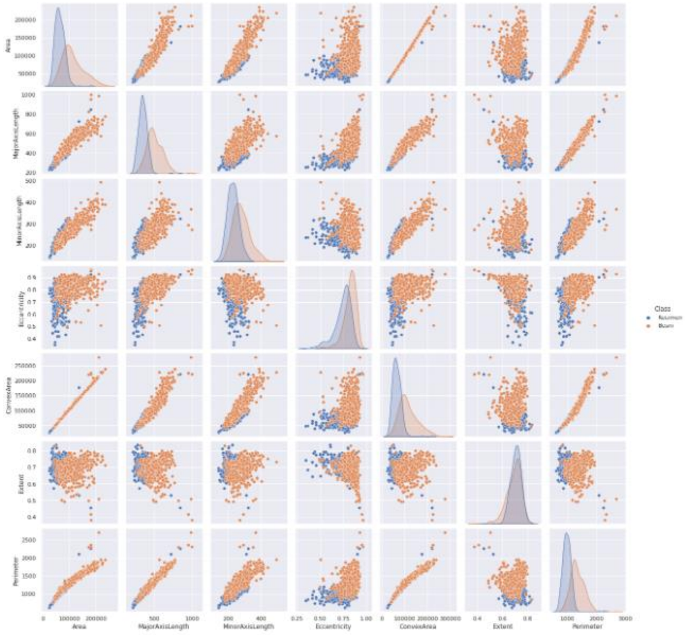


Fig. 6 Pair Plot

Pair plot gives relationship between the variables to see if they are of continuous or categorical nature. This plot is in form of a matrix where x axis gives rows of the data and y-axis is the columns of the data.[8] In fact, it can be visualized as the matrix of scatterplots to understand the relation between variables.[9]

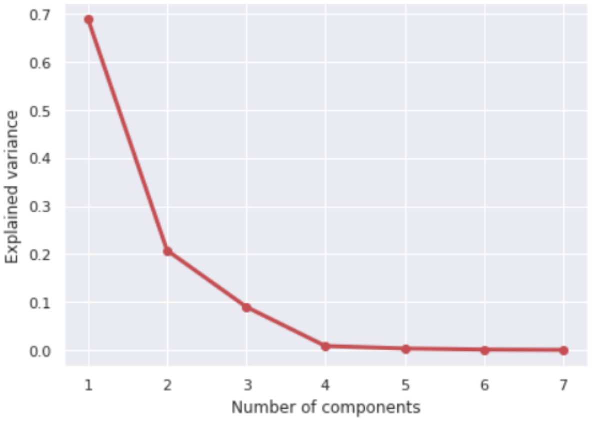


Fig. 7 Scree Plot / Elbow Plot

Variability in the data recorded by each principal component is given by Eigen values. To understand the variance in the data recorded by PCs we have analyzed Scree plot and Pareto plot shown in Fig. 7 and Fig. 8.

Explained variance is the percent value of variance given by the jth principal component. Formula to calculate the same has been given below: [2]

$$\ell_j = \frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j} \times 100\%, \quad j = 1, \ldots, p.$$

where, λ_j is the eigen value and variance of jth PC.

By analyzing the scree plot, we can see that first two PC account for around 89% of the total variance in the original data. Also, for the first three PCs account for 99.58% of the variance as shown in the Pareto plot in Fig. 8.

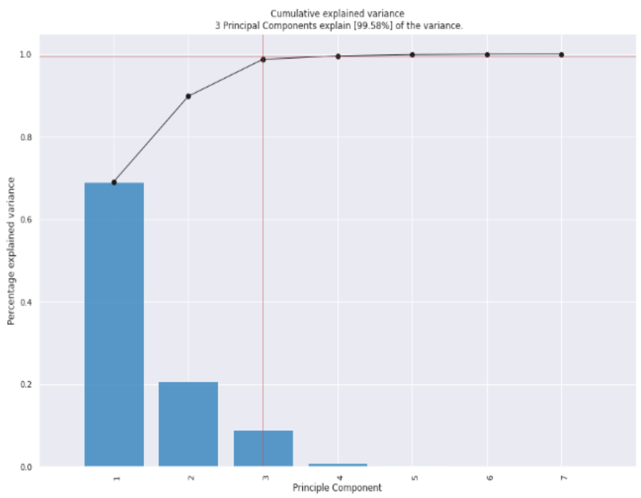


Fig. 8 Pareto Plot

Hence, for further analysis we have taken PC1, PC2 and PC3 and the transformed data has three columns with r=3 and 900 rows. Below we will be giving the equation for the first three PCs denoted by Z1, Z2 and Z3 respectively.

Equation for the first Principal component (PC1) is given as:

$$Z1 = 0.4482\ X1 + 0.4432\ X2 + 0.3893\ X3 + 0.2029\ X4 + 0.4509\ X5 - 0.0563\ X6 + 0.4508\ X7$$

Here in Z1 highest contributions are from X5, X7 and X1 along with others have some contributions to the first principal component. Similary, we have calculated Z2 and Z3 to know the contributions accordingly.

Equation for the second Principal component (PC2) is given as:

$$Z2 = -0.1160\ X1 + 0.1365\ X2 - 0.3749\ X3 + 0.6108\ X4 - 0.0876\ X5 - 0.6673\ X6 + 0.0341\ X7$$

Equation for the third Principal component (PC3) is given as:

$$Z3 = 0.0054\ X1 - 0.1005\ X2 + 0.2360\ X3 - 0.6285\ X4 + 0.0366\ X5 - 0.7319\ X6 + 0.0443\ X7$$

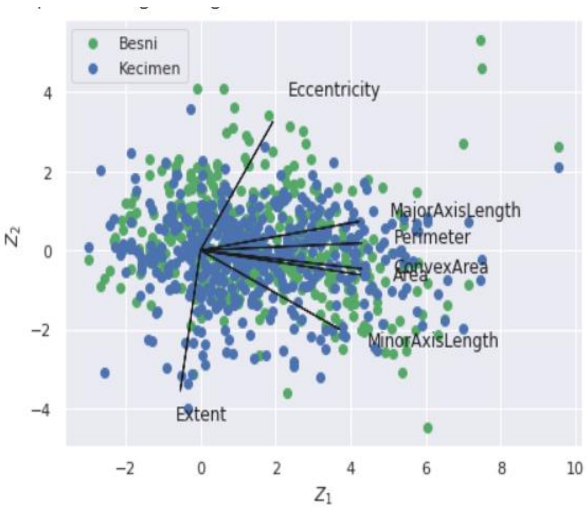


Fig. 9 Biplot

The above Fig. 9 gives the biplot which provides the visual combination of the Principal coefficients of vectors and PC scores. All the dots represent rows observations and vectors represent the rows of the Eigen vector matrix. The angle and length of the vectors represents the contribution of each componet, with small angle means high contribution and vice versa. Green dots represent the Besni type of the raisin and the blue ones are for Kecimen.

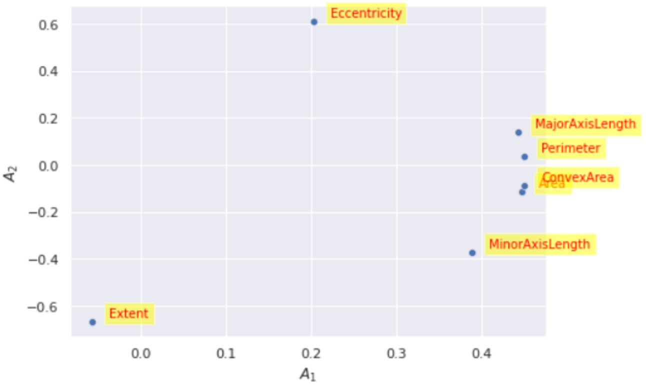


Fig. 10 PC Coefficient Plot

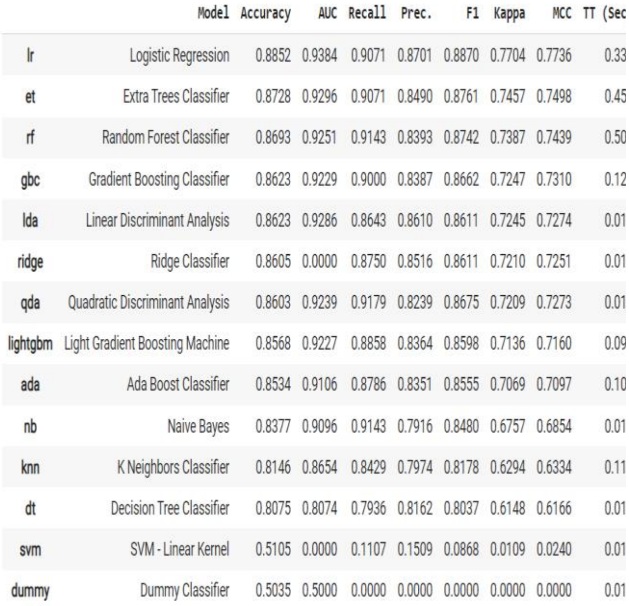
The plot above in Fig. 10 depicts the principal coefficient plots which basically shows variables having similar involvement as first three PCs.

VI. CLASSIFICATION RESULTS

In this section we have discussed the classification results and their visualization using Pycaret and Shap libraries in Python on the raisin data set. The data has been tested on four classification algorithms namely

Logistic Regression, K-Nearest Neighbors, Decision tree and Random Forest. Moreover, the data has been tested on all the models and we have selected best models based on the original data set and the PCA transformed data set.

The raisin data set has been divided in the 70-30 ratio with 7- being training data and 30 being testing data. We have analysed the effect of PCA on the the above-mentioned algorithm and compare their model performance on various criteria.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec
lr	Logistic Regression	0.8852	0.9384	0.9071	0.8701	0.8870	0.7704	0.7736	0.33
et	Extra Trees Classifier	0.8728	0.9296	0.9071	0.8490	0.8761	0.7457	0.7498	0.45
rf	Random Forest Classifier	0.8693	0.9251	0.9143	0.8393	0.8742	0.7387	0.7439	0.50
gbc	Gradient Boosting Classifier	0.8623	0.9229	0.9000	0.8387	0.8662	0.7247	0.7310	0.12
lda	Linear Discriminant Analysis	0.8623	0.9286	0.8643	0.8610	0.8611	0.7245	0.7274	0.01
ridge	Ridge Classifier	0.8605	0.0000	0.8750	0.8516	0.8611	0.7210	0.7251	0.01
qda	Quadratic Discriminant Analysis	0.8603	0.9239	0.9179	0.8239	0.8675	0.7209	0.7273	0.01
lightgbm	Light Gradient Boosting Machine	0.8568	0.9227	0.8858	0.8364	0.8598	0.7136	0.7160	0.09
ada	Ada Boost Classifier	0.8534	0.9106	0.8786	0.8351	0.8555	0.7069	0.7097	0.10
nb	Naive Bayes	0.8377	0.9096	0.9143	0.7916	0.8480	0.6757	0.6854	0.01
knn	K Neighbors Classifier	0.8146	0.8654	0.8429	0.7974	0.8178	0.6294	0.6334	0.11
dt	Decision Tree Classifier	0.8075	0.8074	0.7936	0.8162	0.8037	0.6148	0.6166	0.01
svm	SVM - Linear Kernel	0.5105	0.0000	0.1107	0.1509	0.0868	0.0109	0.0240	0.01
dummy	Dummy Classifier	0.5035	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.01

Fig. 11 Comparing all models on original data

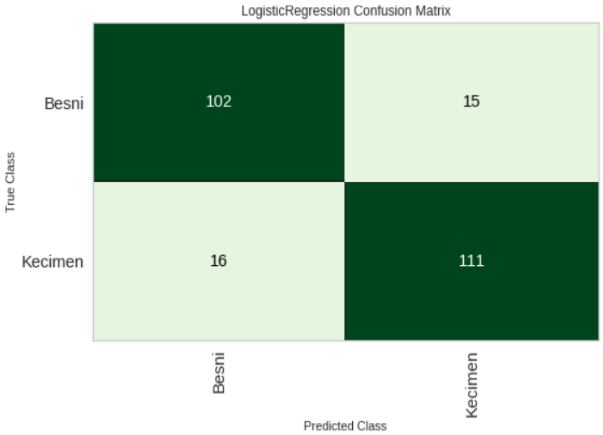
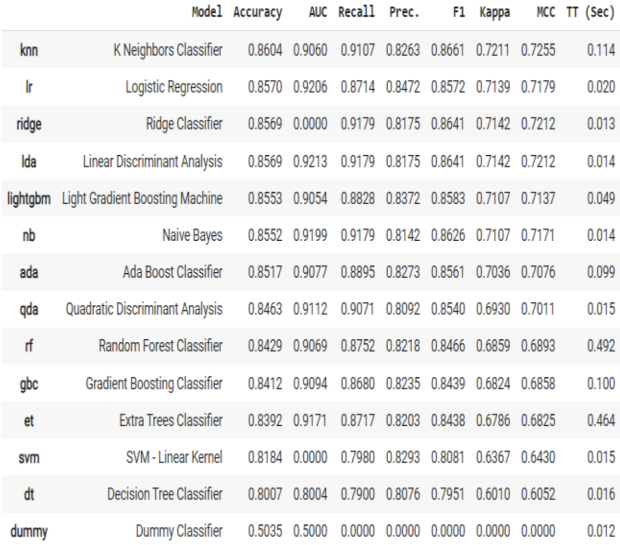


Fig. 12 Confusion Matrix for Best Model Logistic Regression on original data

Fig. 11 shows the performance of all the models based on different categories like accuracy, are under the curve etc. Looking at the table it is clear that the Logistic Regression turns out to be the best performing model for the original data set.

In Fig. 12 confusion matrix for the logistic regression has been given. It is evident from the matrix that 102 times Besni has been correctly classified as Besni and 15 times incorrectly classified as Kecimen. Similarly Keciment correct and incorrect classification has been done 111 and 16 times respectively.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
knn	K Neighbors Classifier	0.8604	0.9060	0.9107	0.8263	0.8661	0.7211	0.7255	0.114
lr	Logistic Regression	0.8570	0.9206	0.8714	0.8472	0.8572	0.7139	0.7179	0.020
ridge	Ridge Classifier	0.8569	0.0000	0.9179	0.8175	0.8641	0.7142	0.7212	0.013
lda	Linear Discriminant Analysis	0.8569	0.9213	0.9179	0.8175	0.8641	0.7142	0.7212	0.014
lightgbm	Light Gradient Boosting Machine	0.8553	0.9054	0.8828	0.8372	0.8583	0.7107	0.7137	0.049
nb	Naive Bayes	0.8552	0.9199	0.9179	0.8142	0.8626	0.7107	0.7171	0.014
ada	Ada Boost Classifier	0.8517	0.9077	0.8895	0.8273	0.8561	0.7036	0.7076	0.099
qda	Quadratic Discriminant Analysis	0.8463	0.9112	0.9071	0.8092	0.8540	0.6930	0.7011	0.015
rf	Random Forest Classifier	0.8429	0.9069	0.8752	0.8218	0.8466	0.6859	0.6893	0.492
gbc	Gradient Boosting Classifier	0.8412	0.9094	0.8680	0.8235	0.8439	0.6824	0.6858	0.100
et	Extra Trees Classifier	0.8392	0.9171	0.8717	0.8203	0.8438	0.6786	0.6825	0.464
svm	SVM - Linear Kernel	0.8184	0.0000	0.7980	0.8293	0.8081	0.6367	0.6430	0.015
dt	Decision Tree Classifier	0.8007	0.8004	0.7900	0.8076	0.7951	0.6010	0.6052	0.016
dummy	Dummy Classifier	0.5035	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.012

Fig. 13 Comparing all models on PCA transformed data

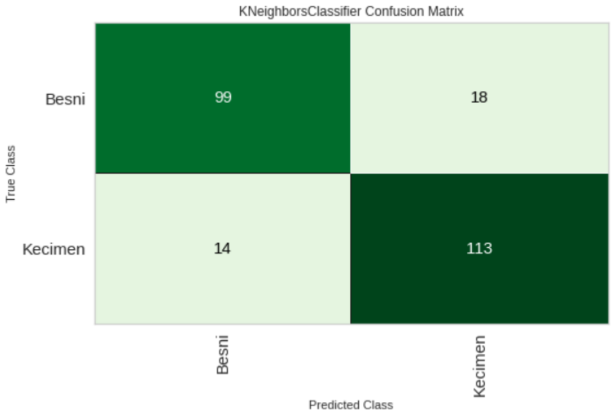


Fig. 14 Confusion Matrix for Best Model K-Nearest Neighbor on PCA

After the application of PCA on the original data and evaluating the data against all the classifiers, KNN comes to be the best performing model as shown in Fig. 13. In the Fig 14, the confision matrix of the best performing model KNN has been shown. This shows the diagonal values in dark green for raisin types Besni and Kecimen has been correctly tagged with 99 and 113. Whereas in the light blue colours are the numbers when these types were incorrectly classified 14 and 18 times.

VII. EXPLAINED AI WITH SHAPLEY VALUES

Explained AI is a type of artificial intelligence in which the machine makes decision which can be understood by humans. This concept is different with the 'black box' concept of machine learning where the decision of AI becomes difficult for the coder or designers to explain. [12] Shap means Shapley Additive Explanations which is essentially a technique to describe individual prediction on the basis of game theory. Shapely values are really popular approach from the cooperative game theory.[11] In this report, Random Forest model has been used along with principal components to distinguish two different types of raisins called Besni and Kecimen. Below in Fig. 15 shows the Shap summary plot which shows feature value arranged as per the sum of the Shap value. Also, in Fig. 16 and Fig. 17 we have shown single and multi-prediction visualization which describe the effect of three principal component on the model of Random forest.

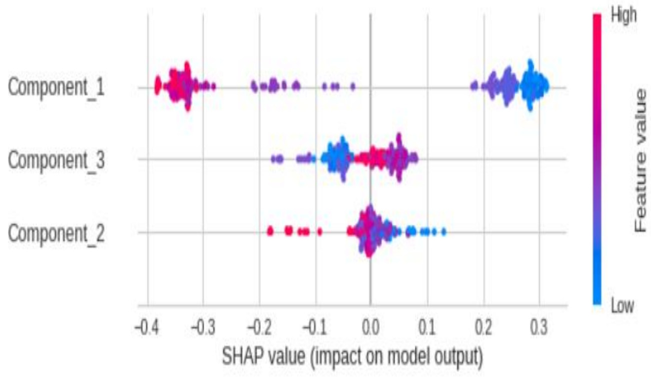


Fig. 15 SHAP Summary Plot

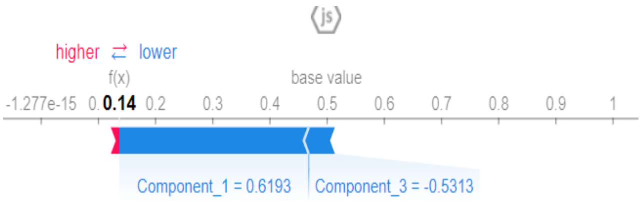


Fig. 16 Single Prediction Visualization

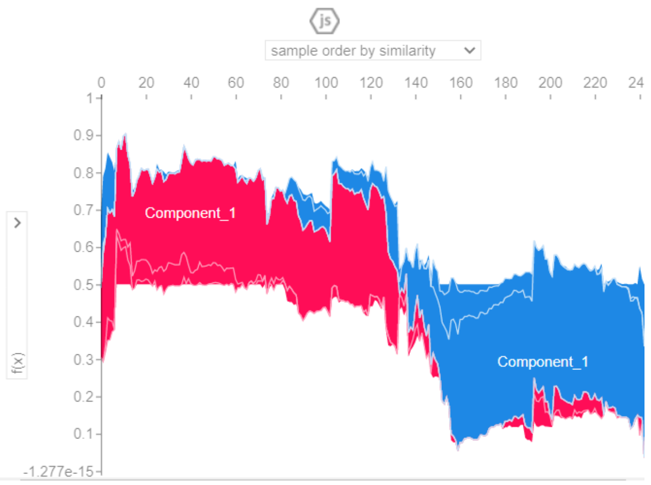


Fig. 17 Multi Prediction Visualization

VIII. CONCLUSION

Using the PCA we got more than 95% of the variations recorded with three principal components. On the application of various machine learning algorithms, we analyzed that Logistic regression turns out to be the best model for the data. However, working the three PCs and testing them other classifiers we found KNN to be best performing on the transformed data. To conclude, we have successfully analysed the difference between the raisin types with the application of principal component analysis and several machine learning algorithms.

REFERENCES

[1] CINAR I., KOKLU M. and TASDEMIR S., (2020). Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Methods, Gazi Journal of Engineering Sciences, vol. 6, no. 3, pp. 200-209, December, 2020, DOI: 10.30855/gmbd.2020.03.03

[2] A. Ben Hamza, Advanced Statistical Approaches to Quality, unpublished.

[3] Lever, J., Krzywinski, M. & Altman, N. Principal component analysis. Nat Methods 14, 641–642 (2017). https://doi.org/10.1038/nmeth.4346

[4] https://www.tutorialspoint.com/mach

[5] ine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic_regression.htm

[6] https://www.javatpoint.com/logistic-regression-in-machine-learning

[7] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[8] https://en.wikipedia.org/wiki/Covariance

[9] https://medium.com/analytics-vidhya/pairplot-visualization-16325cd725e6

[10] https://www.statology.org/pairs-plot-in-python/#:~:text=A%20pairs%20plot%20is%20a,different%20variables%20in%20a%20dataset.

[11] https://towardsdatascience.com/explainable-ai-xai-with-shap-regression-problem-b2d63fdca670#:~:text=SHAP%20(Shapley%20Additive%20Explanations)%20by,that%20come%20with%20desirable%20properties.

[12] https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

[13] https://www.kaggle.com/datasets/muratkokludataset/raisin-dataset