International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Customer Segment Prognostic System by Machine Learning using Principal Component and Linear Discriminant Analysis

R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew

Abstract*: Recently, manufacturing industry faces lots of problem in predicting the customer behavior and group for matching their outcome with the profit. The organizations are finding difficult in identifying the customer behavior for the purpose of predicting the product design so as to increase the profit. The prediction of customer group is a challenging task for all the organization due to the current growing entrepreneurs. This results in using the machine learning algorithms to cluster the customer group for predicting the demand of the customers. This helps in decision making process of manufacturing the products. This paper attempts to predict the customer group for the wine data set extracted from UCI Machine Learning repository. The wine data set is subjected to dimensionality reduction with principal component analysis and linear discriminant analysis. A Performance analysis is done with various classification algorithms and comparative study is done with the performance metric such as accuracy, precision, recall, and f-score. Experimental results shows that after applying dimensionality reduction, the 2 component LDA reduced wine data set with the kernel SVM, Random Forest classifier is found to be effective with the accuracy of 100% compared to other classifiers.

Index Terms*: Machine Learning, Churn, Classification, accuracy, precision, recall, log loss and f-score.

I. INTRODUCTION

Customer group and segment analysis is directly connected with the financial profit of the company. So prediction of customer group have direct impact on the total revenue of the company and it is greatly found to be difficult task for each organization. Once the company identifies the customer group, then they can decide the product design and confirm the number of products to be manufactured based on the customer needs and expectation. The successful sales of any product is decided based on the prediction of the customer expectation and the level of customers. With the growth in the online shopping and offline shopping, the organization have lot of sources to predict the customer group. Customers are rapidly growing for the purchase of various wine brands in supermarkets, five star restaurants, wine shops, online shopping portal and mobile shopping. The organization also need to maintain the customer loyalty while designing the product.

The paper is organized in such a way that Section 2 deals with the related works. Section 3 discuss about the proposed work followed by the implementation and Performance Analysis in Section 4. The paper is concluded with Section 5.

II. RELATED WORK

A. Literature Review

In the process of wine production it is necessary to analyses the influence of chemical samples to taste. The analysis indirectly helps in changes in proportion during wine preparation and predicts its demand in market. The association of chemical parameters can revealed by data mining algorithms. Suitable models can be built to determine the combination of chemical parameters for better wine production. Techniques such as Linear Regression, Decision Trees and Artificial Neural Networks have been used to predict the organoleptic parameters [1]. Results state that accuracy level of built model is appreciable.

An elaborate comparison on customer behavior predicted focusing on customer relationship management, approaches and datasets [2]. Studies reveal that compared to statistical methods, data mining techniques perform well in predictions. Among the data mining techniques, Artificial Neural Networks has been proved outperforming.

The consumption rate of wine was assessed using parameters such as product involvement, subjective knowledge, personal traits and socio demographics. It was found that though it was expensive, the quality of wine rewarded the business [3].

Nowadays the sales of wine increase in online compared to outlets. To better understand the buying behavior a stimulus-response model was built to anticipate the sales rate [4]. The brands of wine has tremendous improvement and the market is highly competitive. Structural Equation modelling was developed to assess the models of wine brand and it was stated that wine experience is related its brand [5].

Revised Manuscript Received on 30 July 2019.

***** Correspondence Author

R. Suguna*, Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

M. Shyamala Devi, Associate Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Rincy Merlin Mathew, Lecturer, Department of Computer Science, College of Science and Arts, Khamis Mushayt, King Khalid university, Abha, Asir, Saudi Arabia.

An online survey on relation between wine attributes and behavioral intentions was carried over and the responses are analyzed using hierarchical regression analysis. Results state that the factors influencing trust and taste impacts consumer behavioral intentions [6].

A research on wine buying decisions states that the behavior of consumers varies with involvement level with the wine product [7]. A survey in a restaurant conveys that the dynamics of dining groups, dining group composition directly relates with wine consumption. These information helps to derive wine-related profile descriptions [8].

To categorize the type of wine consumers based on price and Designation of Origin, Latent Class Model was used [9]. Consumers' attitudes and personal characteristics are described using Principal Component Factor Analysis. An investigation on behaviors of wine consumers state that the behaviors are not related to demographics but subjective to personal values. These findings helped the wine marketers to develop strategies based on the expectation of the consumers [10]. The impact of product knowledge in utilization was considered for the study. Product choice cues were identified and multidimensional approach was used to measure the impact [11]. A critical review on wine quality assessment using user centric similarity measure in product clustering has been done and results show that wine quality has high correlation with user preference groups [12]-[18].

III. PROPOSED WORK

In our proposed work, the wine data set is subjected to predict the behaviour of the customer based on the composition of the wine features. Our implementation in this paper is folded in six ways.

(i) Firstly, creating the correlation matrix and identifying the relationship between each features in the wine data set.

(ii) Secondly, projecting the components with high importance in the wine data set.

(iii) Thirdly, displaying the distribution of high feature importance component with the customer segment target variable of the wine data set.

(iv) Fourth, the wine dataset is subjected to various classifiers like Logistic, KNN, SVM, Kernel SVM, Naive Bayes, Random Forest and Decision and the accuracy is analysed for predicting the customer segment.

(v) Fifth, the feature reduction is done using PCA with number of components as 2, 3 and 5. The feature reduced wine data set is applied to various classifiers and the accuracy is analysed for predicting the customer segment.

(vi) Sixth, the feature reduction is done using LDA with number of components as 2. Then the accuracy is analysed for predicting the customer segment with PCA and LDA for number the component as 2.

A. Principal Component Analysis

The linear conversion of large feature data set into small feature data set is principal component analysis and the steps are given below.

(1) Developing the Covariance matrix for the wine data set

(2) Find the Eigen vectors of Covariance matrix

(3) Reiterate the wine data set with Eigen vectors havin maximum Eigen values.

(4) The high variance features are identified as principal components.

B. System Architecture

The propose architecture of our work is shown in Fig. 1

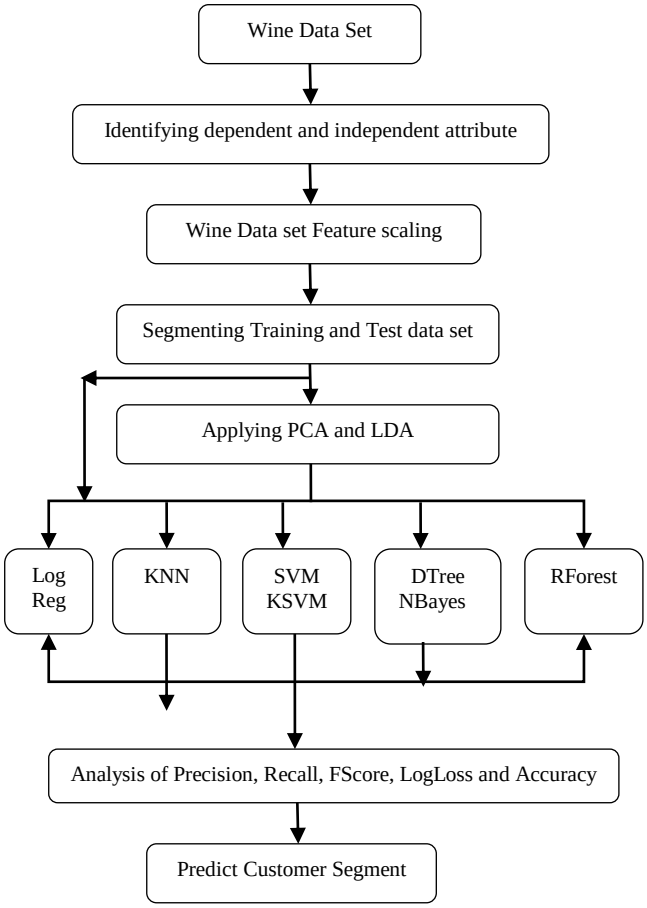

Fig. 1 System Architecture

IV. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Customer Segment Prediction

The Wine dataset from UCL ML Repository is used for implementation with 13 independent attribute and 1 Customer Segment dependent attribute. The attribute are shown below.

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline Churn
14. Customer Segment - Dependent Attribute

Wine data set is executed to create the correlation matrix and identifying the relationship between each features in the wine data set and is shown in Fig. 2.

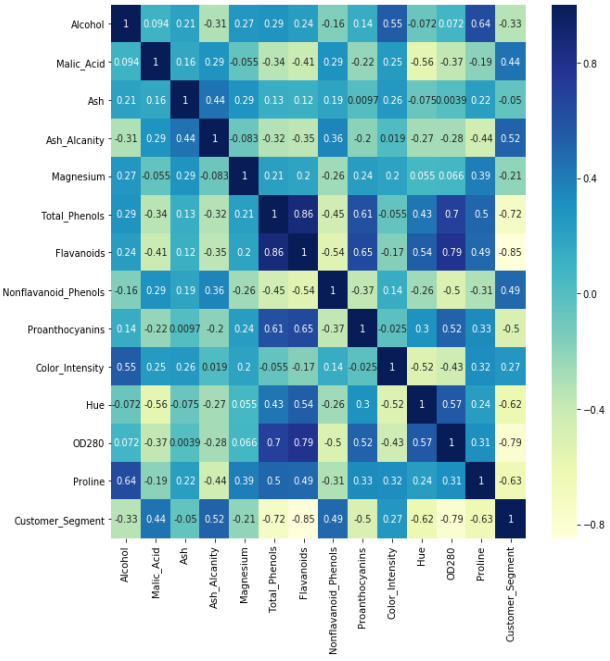

Fig. 2 Correlation Matrix of Wine data set

The feature importance variables of the wine data set is projected as shown in Fig. 3.

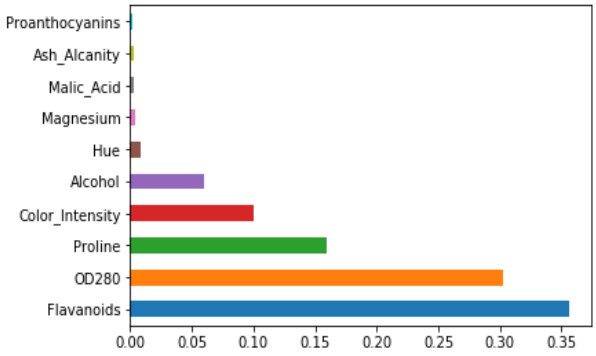

Fig. 3 Feature Importance component Wine data set

The distribution of high feature importance component with the customer segment target variable of the wine data set is done and is shown in Fig. 4- Fig 7.

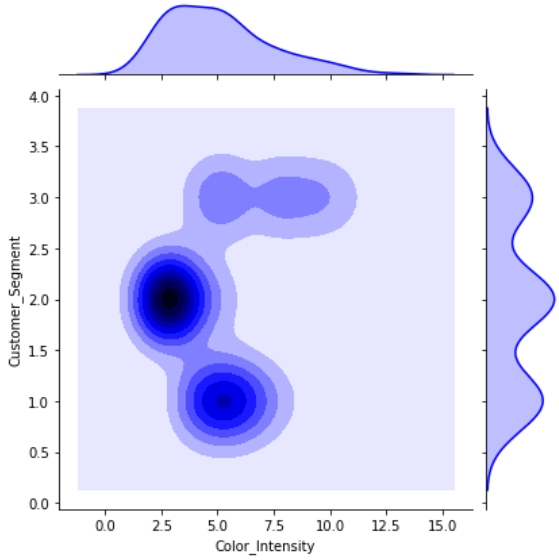

Fig. 4 Color Intensity VS Customer Segment

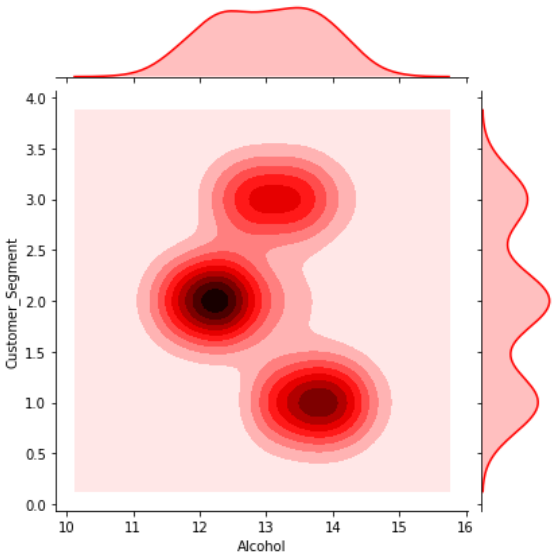

Fig. 5 Alcohol VS Customer Segment

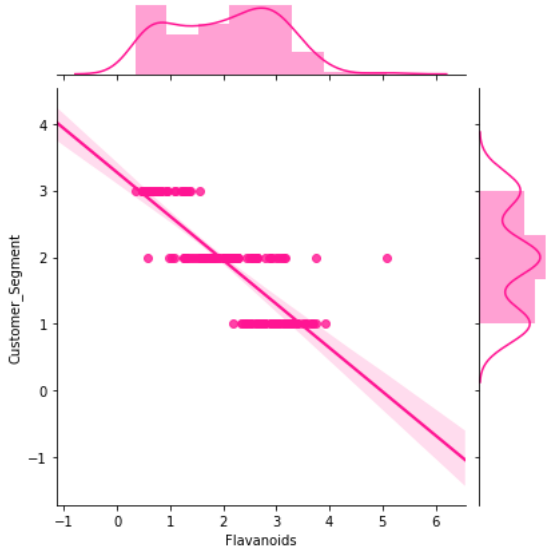

Fig. 6 Flavanoids VS Customer Segment

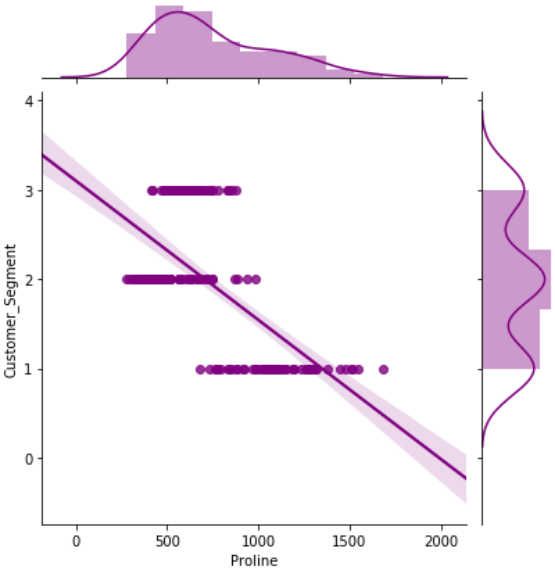

Fig. 7 Proline VS Customer Segment

Wine dataset is applied to various classifiers like Random Forest, SVM, Logistic, Kernel SVM, KNN, Naive Bayes and Decision Tree and the accuracy is compared for predicting the customer segment. The Confusion Matrix is shown in Fig. 8. The performance metric comparison is shown in the Table. 1 and Table. 2.

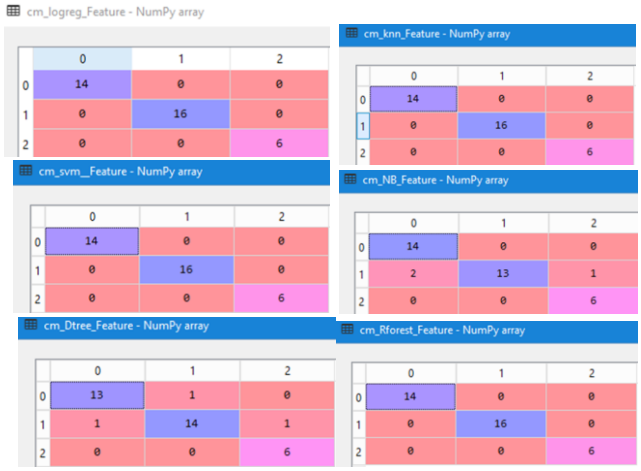

Fig. 8 Confusion Matrix for Random Forest, SVM, Logistic, KNN, Naive Bayes and Decision Tree without PCA

Table. 1 Performance Comparison of Precision, Recall and FScore for all the classifiers without PCA

Classifier	Performance Metrics without PCA		
	Precision	Recall	FScore
Logistic Reg	1.00	1.00	1.00
KNN	0.92	0.97	0.94
SVM	1.00	1.00	1.00
Kernel SVM	1.00	1.00	1.00
Random Forest	1.00	1.00	1.00
Naïve Bayes	0.93	0.94	0.92
Decision Tree	0.92	0.93	0.92

Wine data is applied to PCA with 2,3,5 components and the obtained feature VS Variance of wine data set is shown in the Fig. 9 – Fig. 11

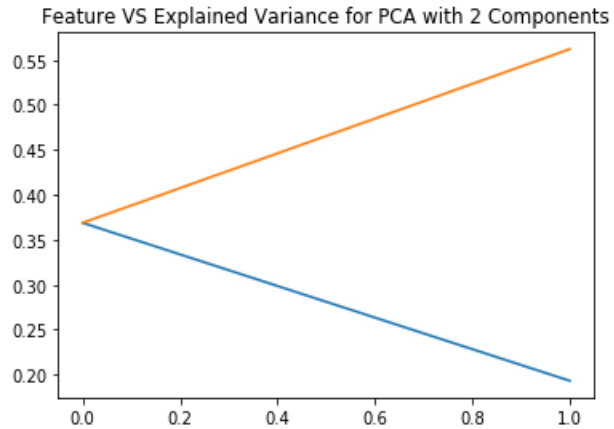

Fig. 9 Feature VS Variance for PCA with 2 components

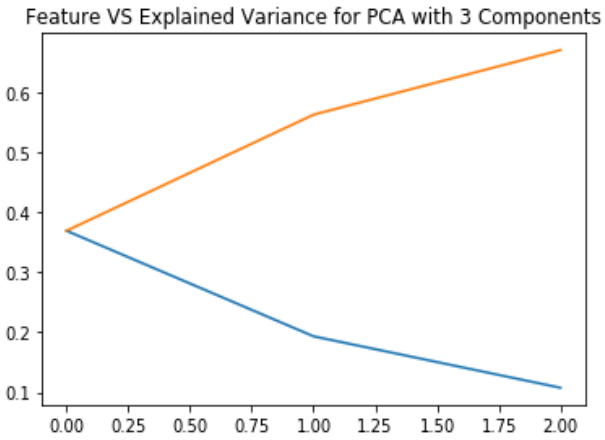

Fig. 10 Feature VS Variance for PCA with 2 components

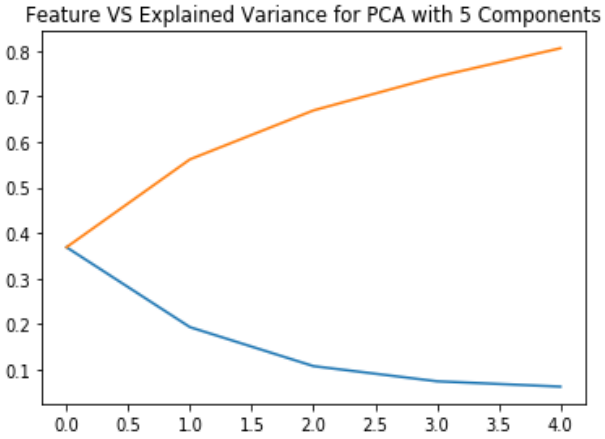

Fig. 11 Feature VS Variance for PCA with 5 components

The PCA reduced Wine dataset is applied various classifiers like Random Forest, SVM, Logistic, Kernel SVM, KNN, Naive Bayes and Decision Tree and the accuracy is compared for predicting the customer segment for PCA with 3 and 5 components. The Confusion Matrix is shown in Fig. 12 – Fig.13. The performance metric comparison is shown in the Table. 2 to Table. 6.

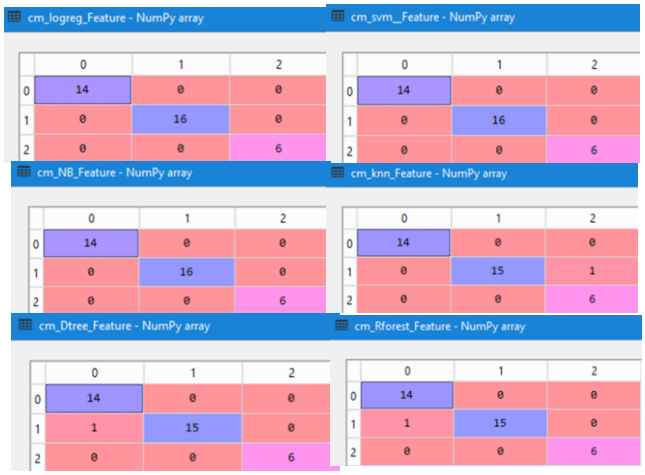

Fig. 12 Confusion Matrix for Random Forest, SVM, Logistic, KNN, Naive Bayes and DTree for PCA with 3 components

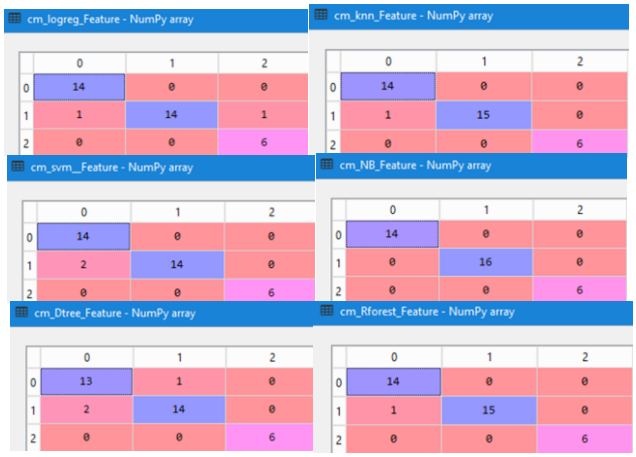

Fig. 13 Confusion Matrix for Random Forest, SVM, Logistic, KNN, Naive Bayes and DTree for PCA with 5 components

Table. 2 Performance Comparison of Precision, Recall and FScore for all the classifiers after applying PCA

Classifier	Performance Metrics with PCA for 3 components		
	Precision	Recall	FScore
Logistic Reg	1.00	1.00	1.00
KNN	0.98	0.97	0.97
SVM	1.00	1.00	1.00
Kernel SVM	1.00	1.00	1.00
Random Forest	0.97	0.97	0.97
Naïve Bayes	1.00	1.00	1.00
Decision Tree	0.97	0.97	0.97

Table. 3 Performance Comparison of Precision, Recall and FScore for all the classifiers after applying PCA

Classifier	Performance Metrics with PCA for 5 components		
	Precision	Recall	FScore
Logistic Reg	0.95	0.94	0.94
KNN	0.97	0.97	0.97
SVM	0.95	0.94	0.94
Kernel SVM	1.00	1.00	1.00
Random Forest	0.92	0.92	0.92
Naïve Bayes	1.00	1.00	1.00
Decision Tree	0.97	0.97	0.97

Table. 4 Performance Comparison of Logarithmic Loss and Accuracy for all the classifiers before applying PCA

Classifier	Accuracy % with PCA for 3 components	Accuracy % with PCA for 5 components
Logistic Reg	100	94.4
KNN	97.2	97.2
SVM	100	94.4
Kernel SVM	100	100
Random Forest	97.2	97.2
Naïve Bayes	100	100
Decision Tree	97.2	91.6

Table. 5 Performance Comparison of Precision, Recall and FScore for all the classifiers after applying PCA

Classifier	Metrics with PCA for 2 components		
	Precision	Recall	FScore
Logistic Reg	0.97	0.97	0.97
KNN	0.97	0.97	0.97
SVM	0.97	0.97	0.97
Kernel SVM	0.97	0.97	0.97
RForest	0.97	0.97	0.97
Naïve Bayes	0.97	0.97	0.97
Decision Tree	0.97	0.97	0.97

Table. 6 Performance Comparison of Precision, Recall and FScore for all the classifiers after applying LDA

Classifier	Metrics with LDA - 2 components		
	Precision	Recall	FScore
Logistic Reg	1.00	1.00	1.00
KNN	0.97	0.97	0.97
SVM	0.97	0.97	0.97
Kernel SVM	1.00	1.00	1.00
Random Forest	1.00	1.00	1.00
Naïve Bayes	0.97	0.97	0.97
Decision Tree	0.97	0.97	0.97

The accuracy of the prediction of customer segment for PCA and LDA with 2 components is given in Table. 7

Table. 7 Performance Comparison of Logarithmic Loss and Accuracy for all the classifiers before applying PCA

Classifier	Accuracy % without PCA	Accuracy % with PCA for 2 components	Accuracy % with LDA for 2 components
Logistic	100	97.2	100
KNN	100	97.2	97.2
SVM	100	97.2	97.2
KSVM	100	97.2	100
RForest	100	97.2	100
NBayes	91.6	97.2	97.2
D Tree	91.6	97.2	97.2

V. CONCLUSION

This paper attempts to predict and cluster the customer behaviour that results in the increase of profit and decision making process of manufacturing design. The correlation matrix of the wine data set is identified between each features in the wine data set. The high featured components of the wine data set is projected. The distribution of high feature importance component with the customer segment target variable of the wine data set is done with the graphical analysis. The performance of all the classifiers of the wine data set is compared before and after applying PCA and LDA. Experimental results shows that after applying dimensionality reduction, the 2 component LDA reduced wine data set with the kernel SVM, Random Forest classifier is found to be effective with the accuracy of 100% compared to other classifiers.

REFERENCES

1. Jorge Ribeiro , Jose Neves , Juan Sanchez , Manuel Delgado , Jose Machado, and Paulo Novais, "1. Wine Vinification prediction using Data Mining tools", Computing and Computational Intelligence, 2016
2. T. Afolabi Ibukun, Olufunke Oladipupo, E. Rowland Worlu, and O. Akinyemi," An Open Access Journal Available Online A Systematic Review of Consumer Behaviour Prediction Studies", 2. Covenant Journal of Business & Social Sciences., vol. 7, no. 1, June 2016.
3. David Cox, "Predicting Consumption, Wine Involvement and Perceived Quality of Australian Red Wine", Journal of Wine Research., vol. 20, no. 3, 2009, pp. 209-229.
4. Constanza Bianchi, "Consumer Brand Loyalty in the Chilean Wine Industry", Journal of Food Products Marketing., vol. 21, no. 4, 2015, pp. 442-460.
5. D. Veena Parboteeah, D. Christopher Taylor, and A. Nelson Barber, "Exploring impulse purchasing of wine in the online environment", Journal of Wine Research., vol. 27, no. 4, 2016, pp. 322-339.
6. Hyojin Kim, and A. Mark Bonn, "The Moderating Effects of Overall and Organic Wine Knowledge on Consumer Behavioral Intention", Scandinavian Journal of Hospitality and Tourism., vol. 15, no. 3, 2015 pp. 295-310.
7. Johan Bruwer, Nicole Burrows, Sylvia Chaumont, Elton Li, and Anthony Saliba, "Consumer involvement and associated behaviour in the UK high-end retail off-trade wine market", The International Review of Retail, Distribution and Consumer Research., vol. 24, no. 2, 2014, pp. 145-165.
8. Johan Bruwer, Justin Cohen, and Kathleen Kelley, "Wine involvement interaction with dining group dynamics, group composition and consumption behavioural aspects in USA restaurants", International Journal of Wine Business Research., vol. 3, no.1, 2019, pp.12-28.
9. Gabriele Scozzafava, Francesca Gerini, Andrea Dominici, Caterina Contini, and Leonardo Casini. "Reach for the stars: The impact on consumer preferences of introducing a new top-tier typology into a PDO wine", Wine Economics and Policy., vol. 7, no. 2, 2018, pp. 140-152.
10. Renata Schaefer, Janeen Olsen, and Liz Thach, "Exploratory wine consumer behavior in a transitional market: The case of Poland", Wine Economics and Policy., vol. 7, no. 1, 2018, pp. 54-64.
11. Johan Bruwer, Polymeros Chrysochou, and Isabelle Lesschaeve., "Consumer involvement and knowledge influence on wine choice cue utilization". British Food Journal., vol. 119, no. 4, 2017,pp. 830-844.
12. Y. Subba Reddy and Prof. P. Govindarajulu, "An Efficient User Centric Clustering Approach for Product Recommendation Based on Majority Voting: A Case Study on Wine Data Set", International Journal of Computer Science and Network Security., vol..17, no. 10, October 2017.
13. N. Modani, K. Dey, R. Gupta, and S. Godbole, "CDR Analysis Based Telco Churn Prediction and Customer Behavior Insights: A Case Study", Web Information Systems Engineering – WISE 2013, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg., vol. 8181, 2013
14. Shyamala Devi Munisamy, and Suguna Ramadass Aparna Joshi, "Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification (Accepted for publication)", Springer's book series entitled "Learning and Analytics in Intelligent Systems, Springer, 2019 to be published.
15. Suguna Ramadass, and Shyamala Devi Munisamy, Praveen Kumar P, Naresh P, "Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers(Accepted for publication)", Springer's book series entitled "Learning and Analytics in Intelligent Systems, Springer, 2019 to be published.
16. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, and Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification , Journal of Statistics and Management Systems, Taylor Francis, vol. 22, no. 4, 25 June 2019, pp. 729-739. DOI: 10.1080/09720510.2019.1609729
17. M. Shyamala Devi, Rincy Merlin Mathew, and R. Suguna,"Attribute Heaving Extraction and Performance Analysis for the Prophesy of Roof Fall Rate using Principal Component Analysis", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2319-2323.
18. R. Suguna, M. Shyamala Devi, and Rincy Merlin Mathew, " Customer Churn Predictive Analysis by Component Minimization using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2329-2333.