Clustering Visualization and Class Prediction using Flask of Benchmark Dataset for Unsupervised Techniques in Machine learning

Ayantika Nath, Shikha Nema

Abstract: ***Cutting edge improved techniques gave greater values to Artificial Intelligence (AI) and Machine Learning (ML) which are becoming a part of interest rapidly for numerous types of researches presently. Clustering and Dimensionality Reduction Techniques are one of the trending methods utilized in Machine Learning these days. Fundamentally clustering techniques such as K-means and Hierarchical is utilized to predict the data and put it into the required group in a cluster format. Clustering can be utilized in recommendation frameworks, examination of clients related to social media platforms, patients related to particular diseases of specific age groups can be categorized, etc. While most aspects of the dimensionality lessening method such as Principal Component Analysis and Linear Discriminant Analysis are a bit like the clustering method but it decreases the data size and plots the cluster. In this paper, a comparative and predictive analysis is done utilizing three different datasets namely IRIS, Wine, and Seed from the UCI benchmark in Machine learning on four distinctive techniques. The class prediction analysis of the dataset is done employing a flask-app. The main aim is to form a good clustering pattern for each dataset for given techniques. The experimental analysis calculates the accuracy of the shaped clusters used different machine learning classifiers namely Logistic Regression, K-nearest neighbors, Support Vector Machine, Gaussian Naïve Bayes, Decision Tree Classifier, and Random Forest Classifier. Cohen Kappa is another accuracy indicator used to compare the obtained classification result. It is observed that Kmeans and Hierarchical clustering analysis provide a good clustering pattern of the input dataset than the dimensionality reduction techniques. Clustering Design is well-formed in all the techniques. The KNN classifier provides an improved accuracy in all the techniques of the dataset.***

Keywords: ***Unsupervised Clustering, Machine Learning Classifiers, Flask-app, UCI datasets.***

I. INTRODUCTION

Machine Learning (ML) and Artificial Intelligence (AI) is a dynamic method for the predictable future. In recent years, it was observed a fast advance in executing numerous modern and improved algorithms. Numerous innovative algorithms were proposed by scholastic, researchers as well as high-tech businesses [1]. Machine learning features a wide assortment of algorithms in numerous real-time applications. Data mining is a sub-stream category of machine learning in this it is utilized to extract valuable and significant data from expansive datasets. One of the broadly utilized real-time methods is the Clustering method which is the foremost important method utilized in data-mining. Clustering is an unsupervised learning mechanism that predicts the output datapoint and organizes it into a cluster format from the given input dataset. The output data points are grouped agreeing to the similitude and dissimilarity measure of the individuals of the dataset. Partitioning and Hierarchical strategies are essential sorts of clustering methods. K-means (Partitioning method) and Hierarchical (Progressive method) clustering are broadly utilized in clustering methods. The growth of data information of the dataset is expanding massively. The dimensionality reduction method is utilized on an expansive scale to downsize the information to fit into the algorithm. The two most prominent and frequently utilized DR methods are Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). K-means is the highest best technique utilized to bunch the data and put it into cluster format. The primary point of this research is to perform a comparative analysis on partitioning strategies like K-means, Hierarchical and Dimensionality reduction strategies such as PCA and LDA by utilizing different datasets so that to plot the good cluster pattern, discover its exactness (accuracy) and Predicting the class of the dataset utilizing flask-app. The Literature survey is depicted in Section II. The Clustering methods and Dimensionality Reduction procedure, Dataset used, and Flask app is depicted in detail in Section III. Datasets used in the program to showcase the cluster output in mentioned Section IV.

Revised Manuscript Received on May 30, 2020.
* Correspondence Author

Ayantika Nath*, Department of Electronics and Communication, Usha Mittal Institute of Technology, S.N.D.T Women's University, Mumbai, India. Email: ayantika22.an@gmail.com

Shikha Nema, Department of Electronics and Communication, Usha Mittal Institute of Technology, S.N.D.T Women's University, Mumbai, India. Email: shikha.nema@umit.sndt.ac.in

TABLE- I: ACRONYMS AND MEANINGS

Acronym	Meaning
ML	Machine Learning
AI	Artificial Intelligence
DR	Dimensionality Reduction
PCA	Principle Component Analysis
LDA	Linear Discriminant Analysis
KNN	K-Nearest Neighbor
HCA	Hierarchical Clustering Analysis
SVM	Support Vector Machine
LR	Logistic Regression
DTC	Decision Tree Classifier
RFC	Random Forest Classifier
GNB	Gaussian Naïve Bayes

Retrieval Number: G5943059720/2020©BEIESP
DOI: 10.35940/ijitee.G5943.059720
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

II. LITERATURE SURVEY

Broadly utilized clustering and DR procedures are used for multivariate datasets to cluster the data points and predict their classes. The exceptionally well known and broadly used IRIS dataset is compared utilizing two clustering procedures specifically CLARA and K-means. These methods are forms of Partitioning based clustering strategy in machine learning. In comparison, the creator demonstrated that the Clustering Large Applications (CLARA) strategy demonstrated to be the most excellent procedure than the K-Means technique. These two methods are executed using R programming. This paper made a difference in choosing the foremost well-known IRIS dataset and executing the K-means clustering strategy in this project. The utilization of this technique helped in shaping the clustering in the given measurements conjointly helped in knowing the K-means Clustering technique in a distant better way [2].

The reference authors in [3] compared their modern proposed Progressive method with five distinctive datasets to calculate the execution of different classification frameworks like Neural systems, KNN (K-Nearest Neighbors), and numerous more. Accuracy was determined to utilize different classifiers on a distinctive dataset.

Authors [4] gave a profound thought of clustering strategy specifically Simple K-means, Density-Based spatial Clustering of Algorithm with noise (DBSCAN), Hierarchical Clustering Analysis (HCA), Make Density-Based Clustering Algorithm (MDBCA) to calculate it's execution analysis and time complexity by testing it on diverse datasets specifically Abalone, Bankdata, Switch, SMS and Webtk dataset utilizing Weka tool.

In [5], DR methods to be specific LDA and PCA recognition of Breast Cancer, Iris, Glass, Yeast, and Wine dataset is performed to improve the classified wrong information using various classifier. It made a difference in data optimization due to which rate of correct classification increased.

The author referenced in [6], has displayed a flask web-based application on food photography for portable phone clients. The app is based on Thai foods of thirteen distinctive classes and an image recognition model is constructed utilizing Create Training and Test Data (CNN) and VGG19 model. This model demonstration is executed utilizing Tensorflow and Keras.

In 2019, the author of reference [7] designed a Model-View-Controller (MVC) framework using the flask app. MVC framework does not support flask. The authors tried to fit the MVC model to the python flask app thus work quality and speed of mobile users utilizing this app is improved. The loading time of the program of normal python flask and MVC Flask was compared and found MVC flask provides better speed.

III. CLUSTERING AND DIMENSIONALITY REDUCTION TECHNIQUES

Unsupervised machine learning such as Kmeans, HCA, PCA, and LDA is taken since it was observed that the bunching of clusters is taken place easily. These techniques are widely used in many machine learning algorithms, to visualize the large dataset into a small cluster format. Unsupervised learning provides a good predicted output from the given input data. The commonly used Clustering and DR methods namely K-means, Hierarchical clustering, PCA, and LDA provides a better data visualization outcome. The UCI benchmark dataset such as IRIS, Wine, and Seed is taken into consideration since this dataset gives better clustering outcomes for the above techniques. The similarities between these datasets are utilized for clustering and classification reasons. The proposed strategy is specified in Fig. 3.1.

The dataset is fed into clustering and DR program at that point clustering visualization output is obtained. The classifier is used to calculate the accuracy of the given technique. Cohen Kappa precision is calculated to discover that the precision in accuracy and it is utilized to compare classification results.

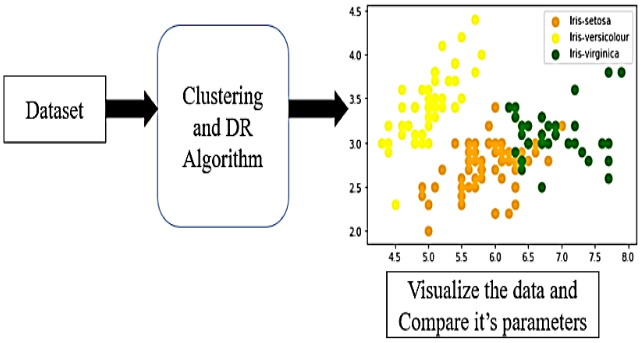


Fig. 3.1 Flow for Proposed Methodology

A. Clustering Technique: K-means

K-means is the simplest and frequently utilized unsupervised learning method to solve the clustering problems. K-means is additionally named as Simple K-means since it is the easiest method to make clustering designs. The basic functionality of K-means is that it bunches cluster according to the dataset. The proposed calculation for K-means is depicted in Algorithm 1. Here the number of clusters taken is three. The maximum iterative in executing k-means is taken as 300 iteration since it gives the best grouping of clusters.

Algorithm 1. PROPOSED CLUSTERING FOR K-MEANS ALGORITHM

Input: 1. Dataset
2. K-means Library
3. Cluster number (n_clusters = 3)
4. Maximum Iterations (max_iter=300)

Output: K-means Cluster

Step 1: Load Dataset.
Step 2: Label Features and Target (class)
Step 3: Initialize K-means
Step 4: Set the clustering number for the Kmeans model
Step 5: Plot Elbow method by allotting Within-Cluster-Sum-of-Squares (WCSS).

```
For i in range (1, 11):
        kmeans = Kmeans (n_clusters = i, max_iter = 300
        End For
Step 6: Fit the model and predict K-means model
Step 7: Plot the K-means model

END
```

B. Clustering Technique: Hierarchical Clustering Analysis

Hierarchical clustering analysis named as HCA is utilized to group the cluster hierarchically. It is one of the data mining methods utilized in a top-down or bottom-up way [1]. This clustering procedure is profoundly sensitive to noise thus the accuracy obtained is less. The proposed algorithm for HCA is clarified in Algorithm 2.

Algorithm 2. PROPOSED CLUSTERING FOR HIERARCHICAL ALGORITHM

```
Input: 1. Dataset
       2. Hierarchical Library
       3. Cluster number (n_clusters = 3)

Output: Hierarchical Cluster

Step 1: Load Dataset.
Step 2: Label Features and Target (class)
Step 3: Visualize Classes of the dataset.
Step 4: Initialize Hierarchical
Step 5: Set cluster number and Fit the Agglomerative Clustering model and predict the data for hierarchy
Step 6: Plot the Hierarchical clustering analysis model
END
```

C. Dimensionality Reduction Technique: Principal Component Analysis (PCA)

The popularly used Dimensionality reduction technique is PCA. Since high dimensional data is difficult to preserve or maintain therefore PCA technique is utilized to gather valuable important data and abbreviate or shorten the size of data. In this, PCA is utilized to gather the data and shape clusters according to the dataset classes. The total clarification in shaping clusters in PCA is stated in Algorithm 3. In this PCA program it goes best with 3 clusters then fits the PCA model and label it as ‘X_r’. At that point this ‘X_r’ variable together with target variable ‘y’ and combined to plot clusters.

Algorithm 3. PROPOSED CLUSTERING FOR PCA ALGORITHM

```
Input:  1. Dataset
        2. PCA Library
        3. Cluster number (n_components = 3)
Output: PCA Cluster
Step 1: Load Dataset and Initialize PCA
Step 2: Label Features and Target (class)
Step 3:Set the number of clusters for the PCA model
        pca = PCA(n_components=3)
Step 4: Fit the PCA model
        X_r = pca.fit(X).transform(X)
Step 5: Create Training and Test Data.
Step 6: Calculate PCA variance and plot Cumulative Proportion of Variance.
Step 7: Train and test PCA until small array dataset is obtained
Step 8: Fit the PCA model
Step 9: Plot the PCA clustering analysis model
END
```

D. Dimensionality reduction technique: Linear Discriminant Analysis

Linear Discriminant Analysis or LDA part is the same as PCA. It is additionally a Dimensionality reduction method that's utilized to diminish High dimensional data to small and important estimation data. For grouping into clusters, it takes the label of the class of a given dataset. Algorithm 4 states the procedure to plot the clusters. The number of clusters is shaped by the labels of the dataset. The ‘x’ and ‘y’ variable fits into the LDA model to plot the clusters.

Algorithm 4. PROPOSED CLUSTERING FOR LDA ALGORITHM

```
Input:  1. Dataset
        2. LDA Library
        3. Cluster number (n_components = 3)

Output: LDA Cluster

Step 1: Load Dataset
Step 2: Initialize LDA
Step 3: Label Features and Target (class)
Step 4: Set the number of clusters for the LDA model
        lda = LDA(n_components=3)
Step 5: Fit the LDA model
        lda_X = lda.fit(X,y).transform(X)
Step 6: Plot the PCA clustering analysis model

END
```

E. Dataset used:

The dataset is taken from the UCI benchmark Repository which supports clustering and classification. The detailed information label of the cluster class in this paper is expressed in underneath Table II.[8][9][10]

TABLE II. DATASET INFORMATION TABLE

Dataset	Instances	Attributes	Application	Label
IRIS	150	5	4 features: Length and width of sepal and petals	Iris setosa Iris virginica Iris versicolor

Wine	178	14	Chemical analysis of wines grown in a particular area	Type 0 Type 1 Type 2
Seed	210	8	Varieties of wheat: Kama, Rosa and Canadian	Target 0 Target 1 Target 2

F. Python Flask web-based Application

Flask is a python web-based application framework. Users can host a web-app on localhost: 5000 with essential knowledge of python language.

For this research, a flask web-based app is utilized to predict the class of the dataset.

The python model is compiled first then the python application program is compiled which directs to browser localhost:5000.

The information is entered and the class of the dataset is predicted.

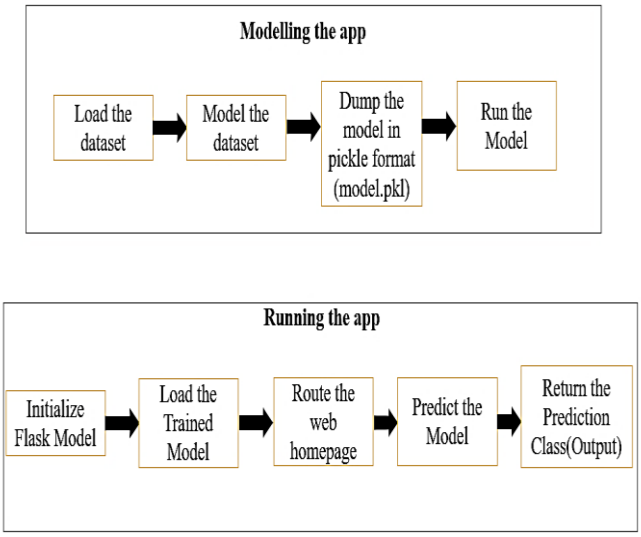


Fig. 3.2 Flask Backend Model

IV. RESULTS AND DISCUSSION

A. Clustering Visualization

The proposed works outcomes are showing us that the Kmeans algorithm provides better clusters than other clustering techniques. PCA and LDA follows the Kmeans clustering accuracy.

The hierarchical clustering accuracy is less than the other due to noise in the data. The accuracy is determined using various ML Classifiers namely KNN, Logistic Regression, Support Vector Machine (SVM), Gaussian Naive Bayes, Decision Tree Classifier, and Random Forest.

The complete accuracy and Cohen Kappa accuracy output values are mentioned in Table IV and Table V.

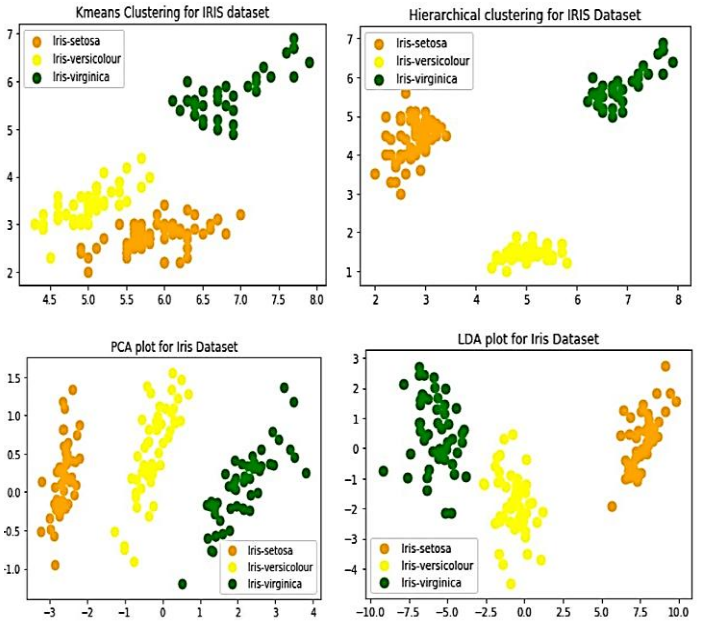


Fig. 4.1 IRIS Dataset Clustering Output

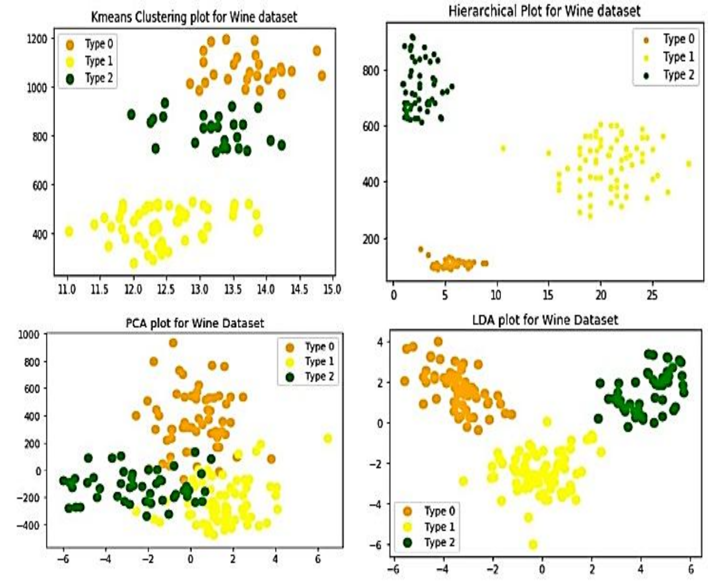


Fig. 4.2 Wine Dataset Clustering Output

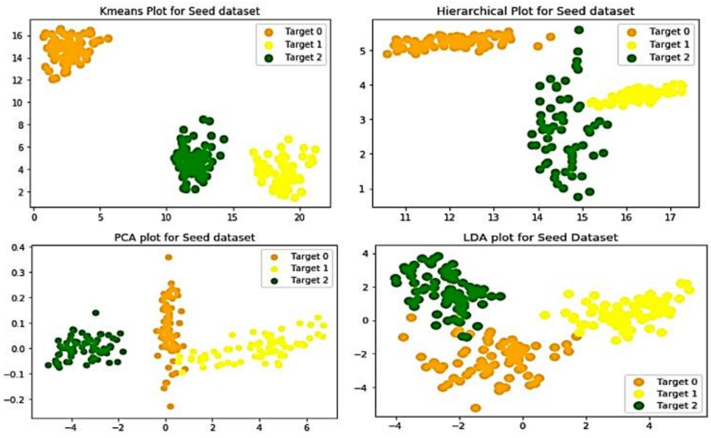


Fig. 4.3 Seed Dataset Clustering Output

TABLE III. ELAPSED TIME FOR VARIOUS DATASETS.

Dataset	Algorithms	Elapsed Time (sec)
IRIS	K-Means HCA PCA LDA	0.029 0.148 0.0214 0.066
Wine	K-Means HCA PCA LDA	0.030 0.181 0.0920 0.0174
Seed	K-Means HCA PCA LDA	0.0176 0.0147 0.0162 0.0212

TABLE IV. ACCURACY TABLE USING VARIOUS CLASSIFIERS

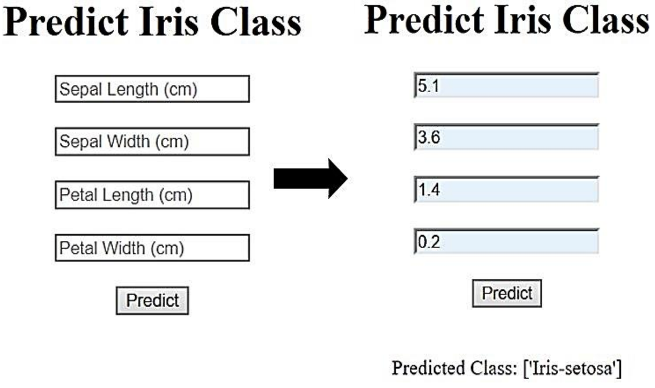
Datasets	Algorithms	Logistic Regression (LR)	KNN	Support Vector Machine (SVM)	Gaussian Naive Bayes (GNB)	Decision Tree Classifier (DTC)	Ran Fo Clas (R
IRIS	Kmeans HCA PCA LDA	91.66 84 96.66 92.10	97.36 96.66 91.66 94.73	97.36 97.77 97.33 97.36	96.66 93.3 94.66 89.47	96.66 93.3 94.66 89.47	96 93 94 89
Wine	Kmeans HCA PCA LDA	94.44 92.52 95.55 93.05	95.5 82.22 93.33 97.77	52.77 50 48.61 42.05	91.66 91.11 97.77 94.38	91.66 94.44 97.77 94.38	91 94 97 94
Seed	Kmeans HCA PCA LDA	92.85 88.09 96.22 92.06	98.1 89.28 93 94.33	88.88 82.10 87.61 92.45	88.09 96.22 92.45 88.67	88.09 96.22 92.45 88.67	88 96 92 88

TABLE V. COHEN KAPPA ACCURACY TABLE USING VARIOUS CLASSIFIERS

Datasets	Algorithms	Cohen Kappa for LR	Cohen Kappa for KNN	Cohen Kappa for SVM	Cohen Kappa for GNB	Cohen Kappa for DTC	Cohen Kappa for RFC
IRIS	Kmeans HCA PCA LDA	87.41 76.19 94.83 88.18	96.03 94.83 87.41 92	96.02 96.61 95.95 96. 02	94.70 88.63 91.85 84.16	93.92 96.61 93.92 96.61	92.39 89.74 83.66 89.74
Wine	Kmeans HCA PCA LDA	91.17 88.63 93.02 89.38	93.05 71.83 88.87 96.65	17.52 19.80 11.20 12.92	86.86 86.29 96.53 91.43	91.52 91.52 89.27 86.79	95.55 96.66 93.32 95.55
Seed	Kmeans HCA PCA LDA	89.25 82.01 94.31 87.84	97.1 83.87 88.88 91.48	83.15 88.09 81.44 88.31	82.08 94.27 88.51 82.93	87.43 90.30 80.04 82.67	92.85 90.29 83.15 91.37

B. Python Flask Prediction Application

In this, the UCI Benchmark dataset namely IRIS, Wine, and Seed dataset is considered. Prediction in keeping with their Class, Type, Target, or Species the value of the dataset is entered and output prediction is received. Firstly, the modeling of a specific dataset is performed and saved in pickle format. Later an application program is built to predict the output of the dataset. In Front end index.html file is for the reason that determines the tabs and button on the localhost. This complete HTML file with Flask Backend app displays the output as shown in Fig. 4.5 for the IRIS dataset, Fig. 4.6 for Wine Dataset, and Fig. 4.7 for Seed dataset. This flask app model is predicting the results very well without any mistakes.


Predict Iris Class

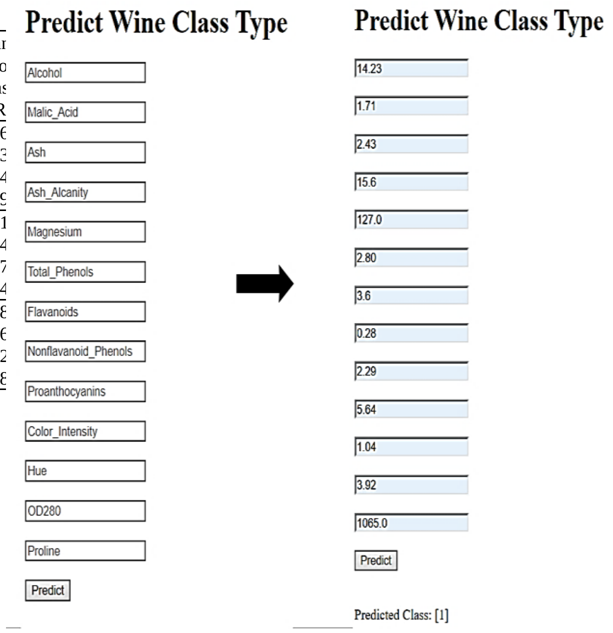
Sepal Length (cm)
Sepal Width (cm)
Petal Length (cm)
Petal Width (cm)

Predict Iris Class

5.1
3.6
1.4
0.2

Predicted Class: ['Iris-setosa']

Fig. 4.5 Flask-app Prediction for IRIS dataset


Predict Wine Class Type

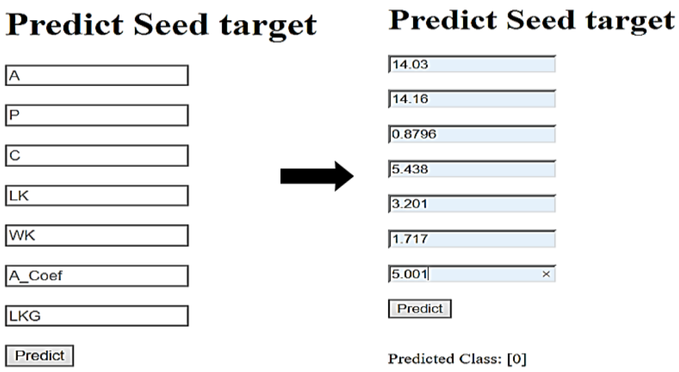
Alcohol
Malic_Acid
Ash
Ash_Alcanity
Magnesium
Total_Phenols
Flavanoids
Nonflavanoid_Phenols
Proanthocyanins
Color_Intensity
Hue
OD280
Proline

Predict Wine Class Type

14.23
1.71
2.43
15.6
127.0
2.80
3.6
0.28
2.29
5.64
1.04
3.92
1065.0

Predicted Class: [1]

Fig. 4.6 Flask-app Prediction for Wine dataset


Predict Seed target

A
P
C
LK
WK
A_Coef
LKG

Predict Seed target

14.03
14.16
0.8796
5.438
3.201
1.717
5.001

Predicted Class: [0]

Fig. 4.7 Flask-app Prediction for Seed dataset

V. CONCLUSION

An unsupervised clustering method such as K-means, HCA, PCA, and LDA clustering visualization is obtained utilizing IRIS, Wine, and Seed Benchmark datasets. From the graph it is observed that K-means and Hierarchical clustering strategies give superior clustering patterns with a given dataset. For better cluster arrangement, the number of clusters taken is three. It is observed Kmeans calculation gives a fast and efficient cluster arrangement. HCA, PCA, and LDA gives a similar clustering visualization as Kmeans. The elapsed time of PCA and LDA algorithm is lesser than other clustering techniques which indicates way better performance evaluation and high speed. The accuracy table provides the exactness in different ML techniques Kmeans,

HCA, PCA, and LDA according to the given dataset. KNN classifier provides a good accuracy overall in different techniques for various datasets. The Cohen Kappa accuracy shows the exactness or trueness of the obtained classification accuracy values. Flask app is predicting all the dataset class of the output accurately without any error. Since the dataset are large in size cannot be handled manually so the Flask app can be used to predict the class of any large dataset.

REFERENCES

1. Bhattacharya, Sambit & Czejdo, Bogdan & Agrawal, Rajeev & Erdemir, Erdem & Gokaraju, Balakrishna. (2018). Open Source Platforms and Frameworks for Artificial Intelligence and Machine Learning. 1-4. 10.1109/SECON.2018.8479098. Sambit Bhattacharya, “Open Source Platforms and Frameworks for Artificial Intelligence and Machine Learning”
2. T. Gupta and S. P. Panda, "Clustering Validation of CLARA and K-Means Using Silhouette & DUNN Measures on Iris Dataset," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 10-13.
3. L. B. Goncalves, M. M. B. R. Vellasco, M. A. C. Pacheco and Flavio Joaquim de Souza, "Inverted hierarchical neuro-fuzzy BSP system: a novel neuro-fuzzy model for pattern classification and rule extraction in databases," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 36, no. 2, pp. 236-248, March 2006.
4. Dang, Shilpa. (2015). Performance Evaluation of Clustering Algorithm Using Different Datasets. IJARCSMS. 3. 167-173. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
5. N. Panahi, M. G. Shayesteh, S. Mihandoost and B. Zali Varghahan, "Recognition of different datasets using PCA, LDA, and various classifiers," 2011 5th International Conference on Application of Information and Communication Technologies (AICT), Baku, 2011, pp. 1-5.
6. U. Tiankaew, P. Chunpongthong and V. Mettanant, "A Food Photography App with Image Recognition for Thai Food," 2018 Seventh ICT International Student Project Conference (ICT-ISPC), Nakhonpathom, 2018, pp. 1-6.
7. M. R. Mufid, A. Basofi, M. U. H. Al Rasyid, I. F. Rochimansyah and A. rokhim, "Design an MVC Model using Python for Flask Framework Development," 2019 International Electronics Symposium (IES), Surabaya, Indonesia, 2019, pp. 214-219.
8. Fisher,R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
9. S. Aeberhard, D. Coomans and O. de Vel, Comparison of Classifiers in High Dimensional Settings, Tech. Rep. no. 92-02, (1992), Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland. (Also submitted to Technometrics).
10. M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak, 'A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010, pp. 15-24.

AUTHORS PROFILE

Ms. Ayantika Nath
Research Scholar (M. TECH Engineering)
Currently Pursuing a Degree as a Master in Technology in the branch of Electronics and Communication Engineering from Usha Mittal Institute of Technology, Shreemati Nathibai Damodar Thackersey (S.N.D.T) Women’s University, Mumbai, Maharashtra, India.Completed Bachelor of Technology (B.Tech) Degree in Electronics Engineering from Usha Mittal Institute of Technology, Shreemati Nathibai Damodar Thackersey (S.N.D.T) Women’s UniversityMumbai, Maharashtra, India in the year 2017.Published one journal research paper on this topic “Clustering Using Dimensional ReductionTechniques for Energy Efficiency in WSNs: A Review” in Journal of Information and Computational Science, Volume 13 Issue 3 – 2020. This paper is also available in Google Scholars.

Dr. Shikha Nema
Professor and Head of the Department of Electronics and Communication Branch at Usha Mittal Institute of Technology, Shreemati Nathibai Damodar Thackersey (S.N.D.T) Women’s University, Mumbai, Maharashtra, India from June 2012 till present.
Professor of Electronics and Telecommunication Department at Vivekanand Education Society’s Institute of Technology Chembur, Mumbai. Maharashtra, India in the year July 2001- June 2012. Received a Ph.D. degree in 2010 and an M.tech degree in Digital Communication from Maulana Azad National Founded of Technology, Bhopal, Madhya Pradesh, India in 2002. Having 14 years of Educating and 3.5 years of Industrial experience. Ten papers published in International and National Journals. Over Fifty Papers published in different International Conferences and National Conferences.