Iris Flower Identification Using PCA and Machine Learning

Posted on November 4, 2026

View this essay online:
https://academic-master.com/iris-flower-identification-pca-machine-learning/

Abstract

This study examines iris flower identification using principal component analysis (PCA) and common machine learning classifiers. The dataset contains 50 examples of each of three species: Iris setosa, Iris versicolor and Iris virginica. Each example has four measurements: sepal length, sepal width, petal length and petal width. PCA provides a lower-dimensional representation for exploring the relationships among these measurements. The workflow prepares the data, examines grouping and classification, and compares reported performance with logistic regression, support vector machines (SVM), k-nearest neighbours (KNN) and Naïve Bayes. The reported scores are 0.925 for the PCA-based workflow, 0.904 for logistic regression, 0.980 for SVM, 0.961 for KNN and 0.923 for Naïve Bayes. The study uses Scikit-Learn and presents implementation diagrams and illustrative outputs.

Keywords: machine learning, principal component analysis, iris classification, logistic regression, support vector machines, k-nearest neighbours, Naïve Bayes.

Introduction

Machine learning uses data to identify patterns and support prediction or classification. Its applications include recognising images, text and sound, as well as analysing structured measurements. Different learning algorithms address different tasks, and their usefulness depends on the data representation, learning objective and evaluation procedure.

Supervised learning uses examples with known target values. Classification predicts categories, while regression predicts numerical outcomes. A model learns relationships from training examples and is then evaluated on other observations.

Unsupervised learning examines structure without requiring category labels during fitting. It can support exploration, dimensionality reduction and grouping of observations [1].

PCA is an unsupervised dimensionality-reduction method. It transforms the measured variables into principal components that describe directions of variation. These components can help visualise data and can serve as inputs to a later classifier; PCA itself does not assign species labels.

Iris species can share visible characteristics, making reliable identification from a small set of measurements a useful classification exercise. This project explores how dimensionality reduction represents the four measured attributes and how several classifiers distinguish the three species.

Study Objectives

The study aims to compare classification approaches for identifying iris species from measured flower attributes.

A further objective is to organise the identification process into a reproducible sequence of data preparation, model application and output comparison.

The dimensionality-reduction stage examines whether a more compact representation retains useful information for the subsequent analysis.

The literature review introduces related studies and the learning methods. The proposed system then describes the data and workflow, followed by implementation illustrations, results and the conclusion.

Literature Review

Previous work has examined classification, clustering and visualisation using the iris dataset. The following studies provide context for the proposed workflow.

Pinto and colleagues examined iris species identification using machine learning approaches, separating model training from prediction on other observations [4].

Nath and Nema (2020) presented clustering visualisation and class prediction through a Flask application using benchmark data [2].

Nisha described an iris flower classification example using logistic regression in an online educational resource [3].

Pinto, Kelur and Shetty (2018) compared logistic regression, SVM and KNN for iris species identification. Their comparison considered differences in classifier performance on the dataset [4].

Poojitha and colleagues investigated iris grouping using a neural-network clustering tool in MATLAB. Related work also considered k-means-based grouping [5, 20].

Thirunavukkarasu and colleagues (2018) applied supervised KNN classification to the iris dataset [7, 16].

Dang (2015) evaluated clustering methods across different datasets [8].

Tallapragada and Rajan (2015) examined morphology-based recognition using a decision-tree classifier [9].

Singh and colleagues compared the effect of data types on Random Forest, Naïve Bayes and KNN classification [15].

Gupta and Panda (2019) compared CLARA and k-means clustering using validation measures on the iris dataset. CLARA works with representative medoids, while k-means uses cluster centroids [6].

Principal Component Analysis

PCA transforms correlated measurements into components through an orthogonal projection. It is commonly used in exploratory analysis and in preparing compact inputs for predictive models. The components are ordered by the variation they represent [13].

K-Nearest Neighbours

KNN predicts a category by examining nearby labelled examples in the feature space. The selected distance measure, feature scaling and number of neighbours influence the result. The method can also be adapted to regression tasks.

Logistic Regression

Logistic regression is a supervised classification method that models class probabilities from input variables. Its name does not make it a substitute for linear regression: the two methods address different target types.

Naïve Bayes

Naïve Bayes applies Bayes' theorem with conditional-independence assumptions about the input

features. It offers a computationally efficient classification approach. Related applications include sentiment classification and spam detection [17, 18].

Support Vector Machines

An SVM learns a decision boundary that separates labelled examples. Its behaviour depends on the kernel and tuning choices. In this study, SVM is one of the classifiers considered alongside the PCA-based workflow, logistic regression, KNN and Naïve Bayes.

Proposed System

The proposed system uses measured iris attributes to identify species. Its purpose is to organise data preparation, dimensionality reduction, grouping and classifier comparison within one workflow. This provides a compact example of applying machine learning to a structured botanical dataset.

The input variables are sepal length, sepal width, petal length and petal width. Although the genus Iris contains approximately 260–300 species, the benchmark used here contains only Iris setosa, Iris versicolor and Iris virginica. PCA reduces the measurement space for exploration. Grouping similar observations and predicting known species labels are distinct steps: the former explores structure, while the latter requires a classification rule.

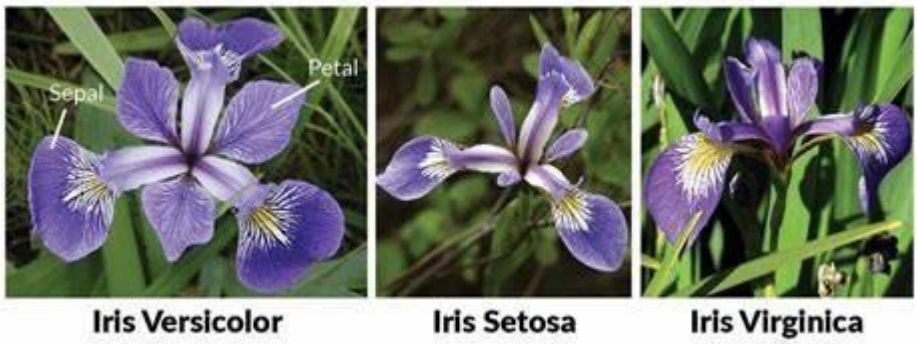


Figure 1 Types of flower

Figure 1 illustrates the three iris species considered in the project. Similar appearance motivates the use of measured attributes rather than relying only on visual impressions.

Figure 2 summarises the project workflow. The dataset supplies four attributes [10]. Measurements are standardised before forming a covariance matrix. Eigenvalues and eigenvectors are calculated and ordered, and the selected components provide a lower-dimensional representation. The

subsequent analysis examines groups and class predictions and compares the reported outputs of the PCA-based workflow with logistic regression, KNN and SVM. Clustering methods and their initialisation are additional background considerations [12].

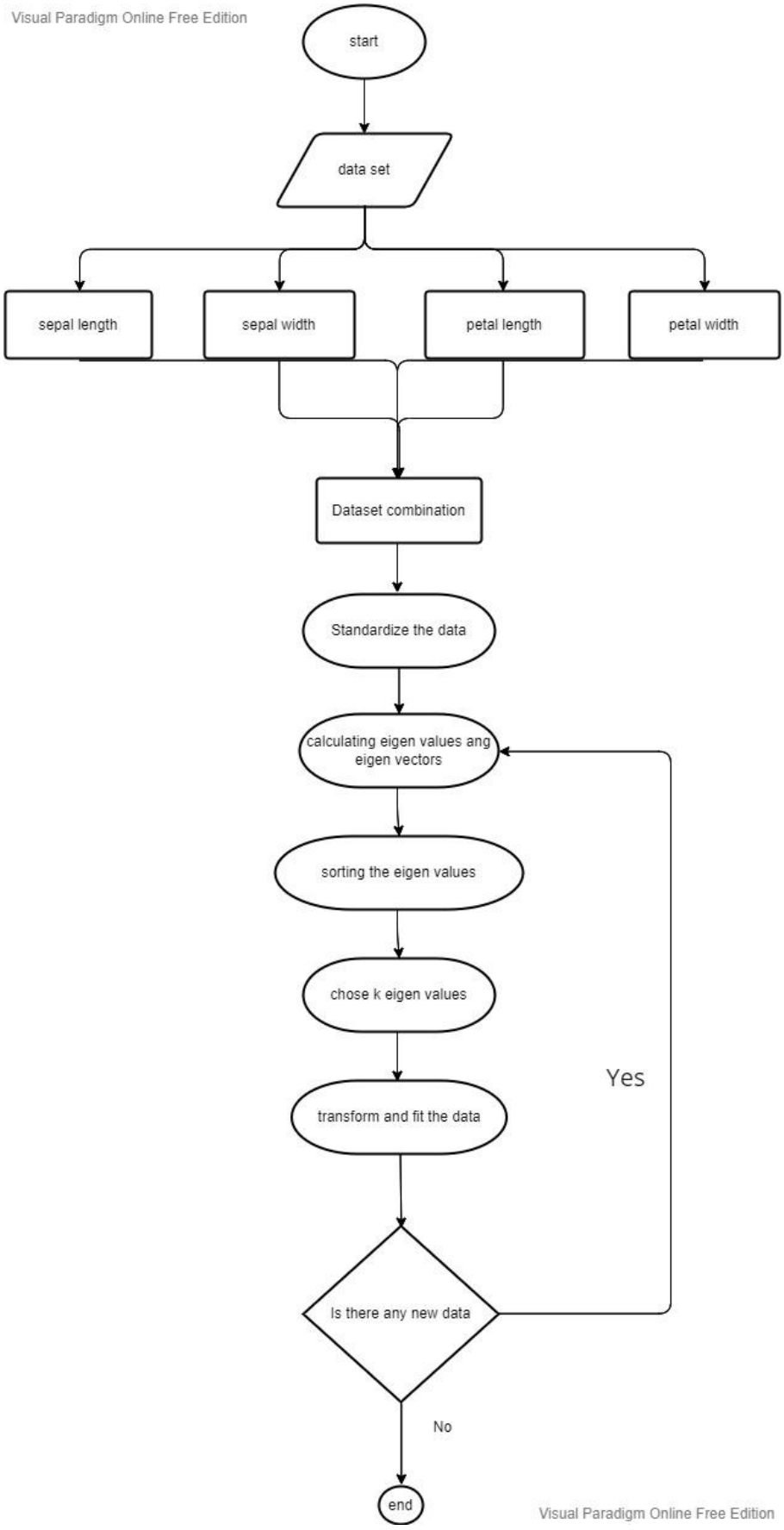

Visual Paradigm Online Free Edition
start
data set
sepal length
sepal width
petal length
petal width
Dataset combination
Standardize the data
calculating eigen values ang eigen vectors
sorting the eigen values
chose k eigen values
transform and fit the data
Is there any new data
Yes
No
end
Visual Paradigm Online Free Edition

Figure 2 Iris identification workflow

Module Design

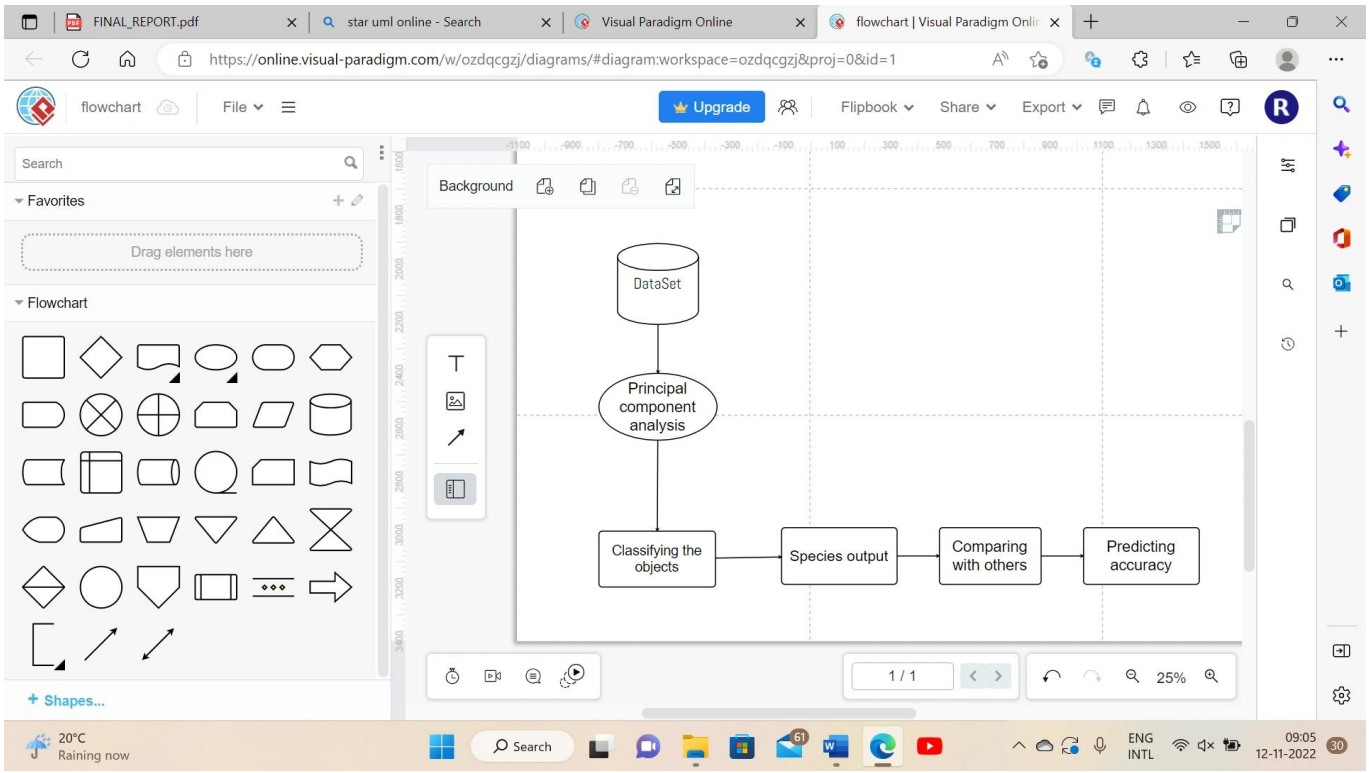


Figure 3 Modules of the proposed system

Figure 3 organises the system into three modules: standardising the data, examining class assignments and comparing performance.

The first module accepts sepal length, sepal width, petal length and petal width. It calculates feature means and standard deviations and prepares the covariance matrix for the component analysis. Related iris classification work also uses these measurements [19].

The second module examines observations with similar measurements and relates the resulting groups or predictions to the three species. Cluster membership and species identity should be distinguished, since similarity alone does not supply a botanical label [20].

The third module compares the reported outputs with those from logistic regression, SVM, KNN and Naïve Bayes. Interpretation depends on using the same evaluation basis when comparing methods.

Data Standardisation

Figure 4 presents the source matrix for the four input measurements. Standardisation and covariance analysis prepare the measurements for dimensionality reduction and the subsequent classification stage.

	f1	f2	f3	f4
f1	0.8	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8	0.51121	0.4945
f3	0.03849	0.51121	0.8	0.75236
f4	-0.14479	0.4945	0.75236	0.8

Figure 4 Four-feature matrix in the data preparation example

Classification

Figure 5 illustrates the classification output for the three species. The workflow uses similarities in measured attributes to examine the observations and their class assignments.

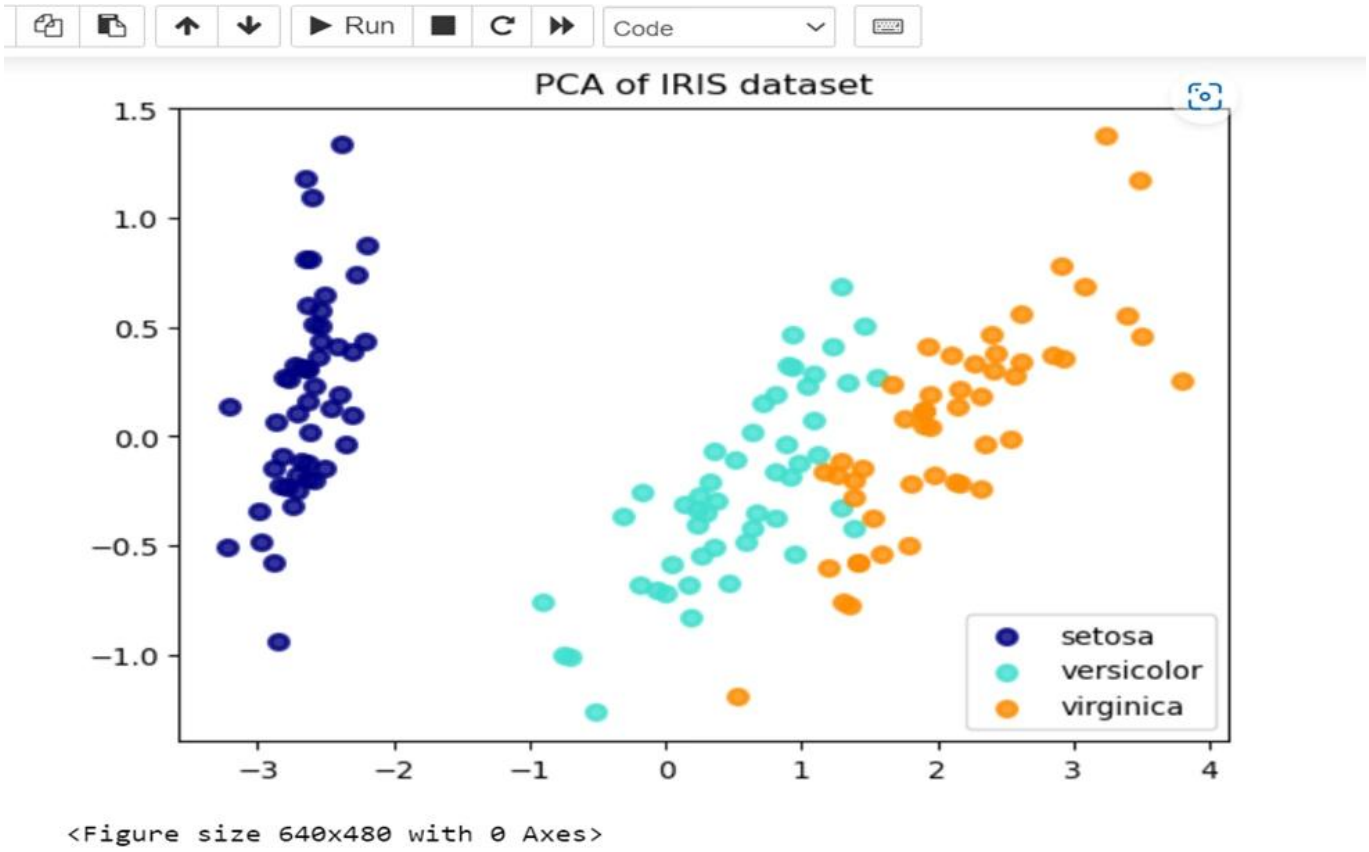


Figure 5 Classifying the data

Performance Comparison

Figure 6 presents the reported comparison of the proposed workflow with other classification methods. Such a comparison should distinguish the dimensionality-reduction step from the classifier applied to the resulting representation.

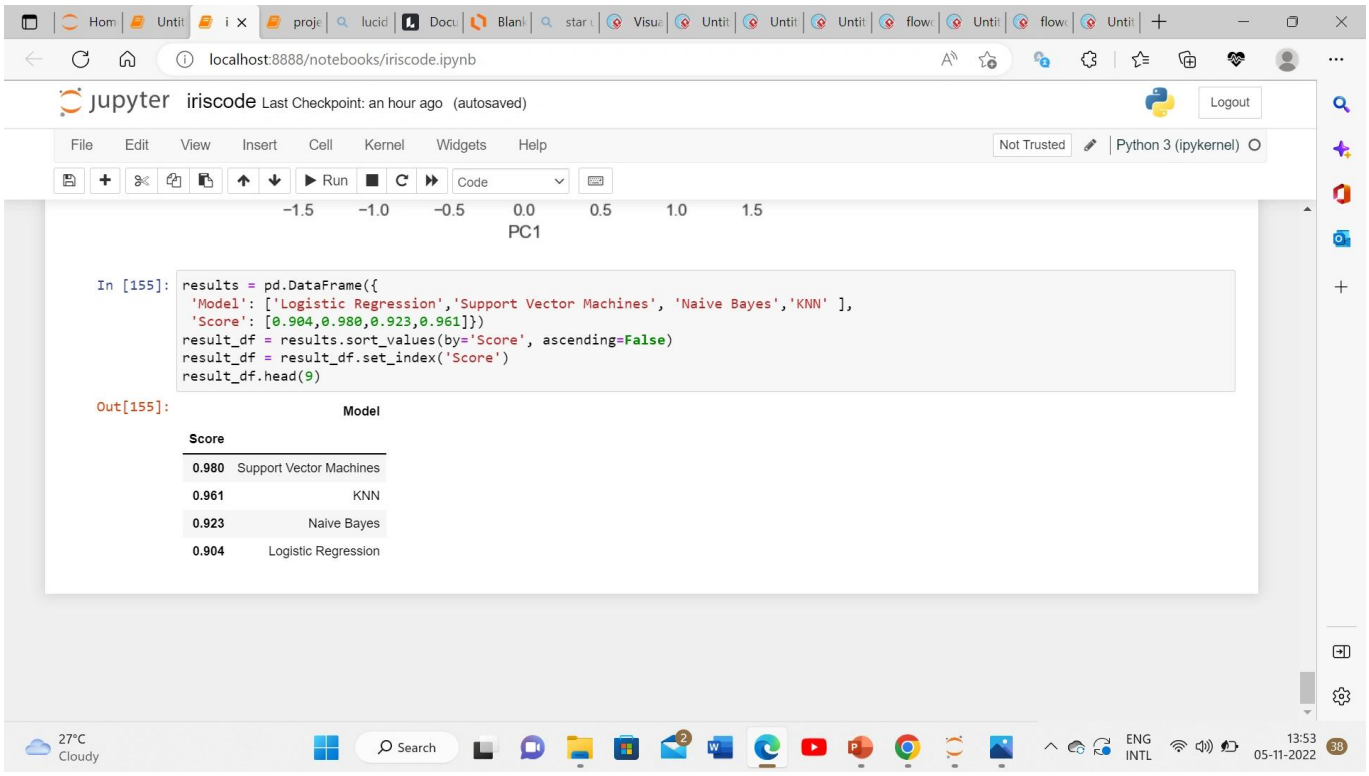


Figure 6 Comparing the accuracy

Implementation and Results

The implementation uses a Kaggle-hosted version of the iris dataset [10]. It contains the three species and four measurement attributes described above. Figure 7 shows KNN classification regions and test observations after PCA. The workflow is implemented using Scikit-Learn, with the earlier comparison output showing the methods evaluated.

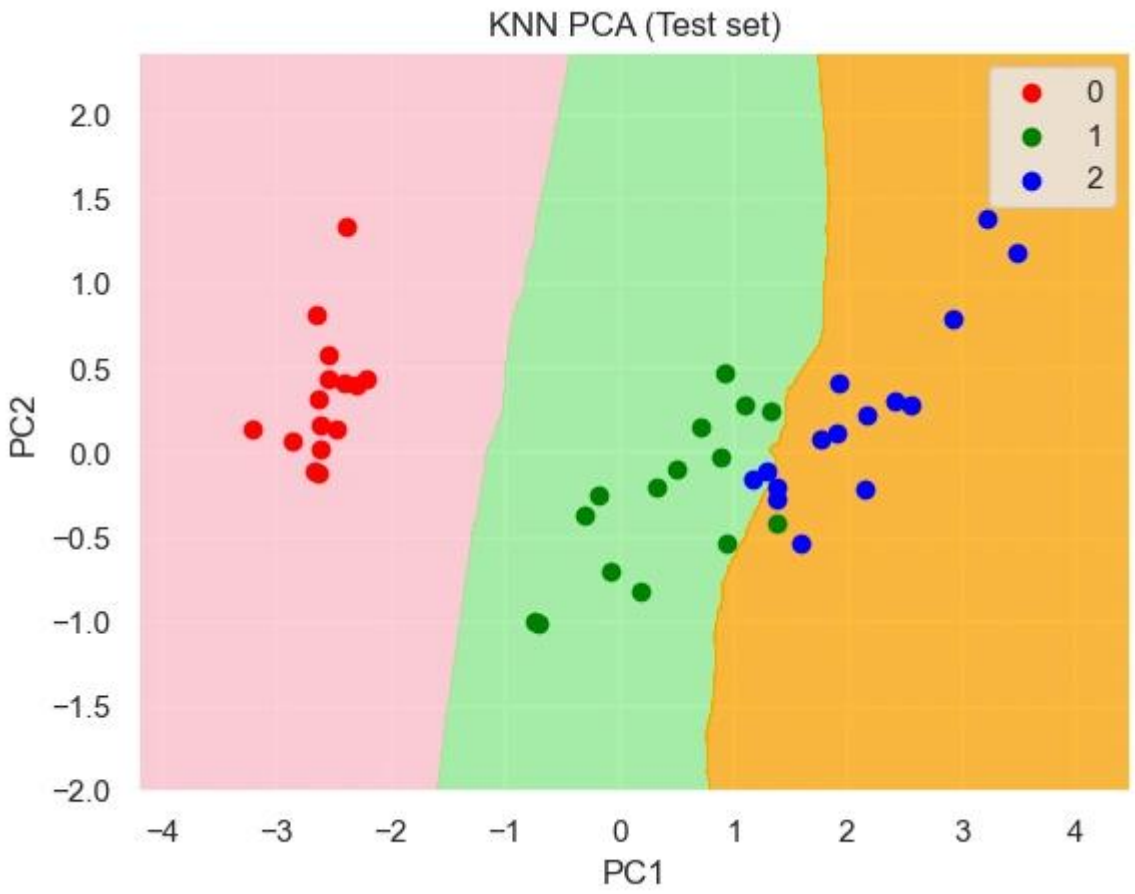


Figure 7 KNN classification after PCA on the test set

Conclusion and Future Work

This study presents an iris identification workflow combining data preparation, PCA-based representation and comparisons among common machine learning classifiers. The four flower measurements provide the inputs, while PCA supports examination of a compact feature space. The reported comparison includes SVM, KNN, logistic regression and Naïve Bayes, with SVM receiving the highest score in the abstract's comparison. The implementation diagrams describe how the stages connect. Further development could make the evaluation protocol more explicit and examine how component selection and classifier settings affect performance. The work concerns species identification; its illustrated outputs do not establish a method for diagnosing flower diseases.

References

1.Lewis DD, Catlett J. Heterogeneous uncertainty sampling for supervised learning. InMachine learning proceedings 1994 1994 Jan 1 (pp. 148-156). Morgan Kaufmann.

2.Ayantika Nath, Shikha Nema- Clustering Visualization and Class Prediction using Flask of Benchmark Dataset for Unsupervised Techniques in Machine learning. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-7, May 2020.

3.Nisha. Logistic regression and iris flower classification problem Available at:https://www.goeduhub.com/758/logistic-regression-iris-flowers-classificationproblem.

4. J. P. Pinto, S. Kelur and J. Shetty, “Iris Flower Species Identification Using Machine Learning Approach”, 2018 4th International Conference for Convergence in Technology (I2CT), pp. 1-4, 2018.

5. V Poojitha and Shilpi Jain, “A Collocation of IRIS Flower Using Neural Network CLusterimg tool in MATLAB”, International Journal on Computer Science and Engineering (IJCSE).

6.Tanvi Gupta,Supriya P Panda Clustering Validation of CLARA and K-Means Using Silhouette & DUNN Measures on Iris Dataset .2019 International Conference on Machine Learning,Bigdata,Cloud and Parallel Computing(COMIT Con).

7. Thirunavukkarasa.K, Ajay S. Singh, Prakhar Rai, Sachin Guptha. Classification flower dataset using classification base KNN Algorithm Supervised learning. 2018 4th international conference on computing communication and automation.(ICCCA).

8. Dang, Shilpa. (2015). Performance Evaluation of Clustering Algorithm Using Different Datasets. IJARCSMS. 3. 167-173.

9. Satyanarayanan Tallapragada, Dr.E.G.Rajan. Morphology based Non ideal recognition using Decision tree classifier. 2015 International Conference on pervasive Computing.

10. Fisher, R.A. 1936. UCI Machine Learning Repository: Iris Data Set. Available at: https://www.kaggle.com/datasets/arshid/iris-flower-dataset.

11. Hui Zhang, Xiangfeng Guan. Iris Recognition based on grouping KNN and Rectangle convertion. 2012 IEEE International Conference on computer science and automation engineering.

12.Qin, X. and Zheng, S. 2009. A new method for initializing the K-means clustering algorithm. Knowledge Acquisition and Modeling, 2009. KAM ’09. Second International Symposium on. vol. 2, pp.

41–44.

13. Ledisi G.Kabari and Believe B.Nwamae . Principle Component Analysis(PCA)-An effective Tool in Machine Learning.2019 International Journals of Advanced Reasearch in Computer Science and Software Engineering.

14. Robin. 2010. Machine-Learning – Artificial Intelligence. Available at: http://intelligence.worldofcomputing.net/category/machine-learning. Consulted 22 AUG 2013.

15. Singh, A; N., M; Lakshmiganthan, R, Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms,International Journal of Advanced Computer Science and Applications,2017.

16. K Thirunavukkarasu, Ajay S Singh, Prakhar Rai, Sachin Gupta Classification of IRIS dataset using classification based KNN algorithm in supervised learning2018 4th International Conference on Computing Communication and Automation (ICCCA), 1-4, 2018.

17.Prabhat A, Khullar V. Sentiment classification on big data using Naïve Bayes and logistic regression. In2017 International Conference on Computer Communication and Informatics (ICCCI) 2017 Jan 5 (pp. 1-5). IEEE.

18. Preethi S, Neelofar S, Chandralekha T, Rishikanth S, Mohiddin SK. An Enhanced Efficient Approach For Spam Detection In IOT Devices Using Machine Learning.

19. Pachipala Y, Maddipati HC, Udimudi PS, Thota V, Sree VN, Burra LR. Iris Flower Classification by using Random Forest in AWS. In2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS) 2022 May 25 (pp. 1772-1777). IEEE.

20. Poojitha V, Bhadauria M, Jain S, Garg A. A collocation of IRIS flower using neural network clustering tool in MATLAB. In2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) 2016 Jan 14 (pp. 53-58). IEEE.