LABWORK 2

PRINCIPAL COMPONENT ANALYSIS

Mai Hải Đăng - BI12-076

01.12.2023

Machine Learning and Data Mining II

Contents

1. Introduction

- This report is dedicated to the understanding of data classification.
- We will tackle many approaches to manage training datasets to reduce variance and avoid overfitting such as: K-fold Cross Validation, Leave-p-out.
- We will understand and implement classification algorithms such as:
    - K-Nearest Neighbors (K-NN)
        - Analyze and compare the results with different values of k.
        - Normalize the dataset to improve convergence rate.
        - Evaluate our clustering quality.
        - Apply PCA (principal components analysis) or SVD (single value decomposition) to see the performance.
    - Perception Classifier
        - Try to optimize the converging rate.
- We will have a brief mention about the concept of data types, data normalization, statistical measurements, PCA, and SVD.
- The original dataset are Iris [1], Breast Cancer Wisconsin (Diagnostic) [2], and Wine [3]

2. Concepts

2.1. Basic concepts

2.1.1. Data

Discrete Data	Continuous Data
Take on distinct, separate values	Take on an infinite number of values within a specified range
Countable (can count the number of different values it can take)	Measurable
Number of students in a classroom Number of cars in a parking lot. Number of books on a shelf.	Height of individuals. Weight of objects. Temperature of a room.

Qualitative	Quantitative
Descriptive in nature and deals with qualities or characteristics.	Measurable quantities and is expressed in numerical terms.
It is non-numerical and is often expressed in words. Relies on the interpretation of meanings, feelings, or attributes.	Can be continuous (measured on a scale) or discrete (countable)

Categories: Qualitative data is often categorical and falls into categories. **Nominal or Ordinal:** It can be nominal (categories with no inherent order) or ordinal (categories with a specific order).	**Continuous or Discrete:** Quantitative data can be continuous (measured on a scale) or discrete (countable). **Ratio or Interval:** It can be ratio (with a true zero point) or interval (without a true zero).

2.1.2. Statistical Measurements

- Mean measures the central tendency and represents the average value of a set of numerical values.

$$\mu(X) = \frac{\sum_{i=1}^{n} X_i}{n}$$

- The Variance measures the spread or dispersion of a set of values from the mean. It gives an indication of how much individual data points differ from the average.

$$Var(X) = \frac{\sum_{i=1}^{n} \left(X \middle| i - \mu(X)\right)^2}{n}$$

- The Covariance measures how two variables change together. A positive covariance indicates a positive relationship, while a negative covariance indicates a negative relationship.

$$Cov(X,Y) = \frac{Var(X) * Var(Y)}{n}$$

- Correlation is a standardized measure of the strength and direction of a linear relationship between two variables. It is a unitless measure that ranges from -1 to 1.

$$Correlation = \frac{Cov(X,Y)}{STD(X) * STD(Y)}$$

Where:

- $STD(variable) = \sqrt{Var(variable)}$

2.1.3. Principal components analysis

- The PCA objective is to project the data onto a lower dimensional linear space such that the variance of the projected data is maximized.
- The PCA objective is to project the data onto a lower dimensional linear space such that the variance of the projected data is maximized.
- **Standardize the Data:**
  - If the features in the dataset are measured on different scales, it's common to standardize the data (subtract mean and divide by standard deviation) to ensure that all features have the same influence on the analysis.
- **Compute the Covariance Matrix:**

    - Calculate the covariance matrix of the standardized data. The covariance matrix provides information about the relationships between different features.
- **Compute Eigenvectors and Eigenvalues:**
    - Compute the eigenvectors and eigenvalues of the covariance matrix. Eigenvectors represent the directions of maximum variance, and eigenvalues indicate the magnitude of variance in those directions.
- **Sort Eigenvectors by Eigenvalues:**
    - Sort the eigenvectors in descending order based on their corresponding eigenvalues. The eigenvector with the highest eigenvalue corresponds to the principal component with the most significant variance.
- **Select Principal Components:**
    - Choose the top k eigenvectors to form the basis of the new subspace, where kk is the desired dimensionality of the reduced space.
- **Project Data onto the New Subspace:**
    - Multiply the original data matrix by the selected eigenvectors to obtain the new lower-dimensional representation of the data.

2.1.4. Single value decomposition

- The SVD objective is to decomposes a matrix into three simpler matrices, facilitating various operation and providing insights into the structure of the original matrix.
- It is represented as: $A = U\Sigma V^T$
    - U is an orthogonal matrix containing the left singular vectors, or the eigenvectors of the matrix.
    - Σ is a diagonal matrix containing the singular values, these values are non-negative and are arranged in descending order, or the eigenvalues of the matrix.
    - The columns of V (or rows of V^T) represent the right singular vector. Like U, they form an orthogonal basis but for the row space of A.

2.2. Classification

2.2.1. Definition

- Data classification is the supervised task of data mining that predicts categorical class labels (discrete or nominal).

2.2.2. Validations

- Confusion Matrix is a table that summarizes the performance of a classification model on a set of test data.
    - The confusion matrix displays the number of:
        - True Positive (TP): Both predicted and actual values are positive.
        - True Negative (TN): Both predicted and actual values are negative.
        - False Positive (FP): The prediction is positive while it's negative.
        - False Negative (FN): The prediction is negative while it's positive.
    - From the confusion matrix we can derive several importance metrics:
        - Accuracy: $ACC = \frac{TP+TN}{TP+TN+FP+FN}$

    - Sensitivity, recall, or true positive rate: $TPR = \frac{TP}{TP+FN}$
    - Specificity, or true negative rate: $TPR = \frac{TN}{TN+FP}$
    - Precision, or positive predictive value: $PPV = \frac{TP}{TP+FP}$
    - Negatively predictive value: $NPV = \frac{TN}{TN+FN}$
- Classification error: Comparing the class labels obtained with the prediction and the original labels of test data.

2.2.3. Prevent overfitting!

- Overfitting: A model is excessively complex, such as having too many parameters relative to the number of training objects. A model has poor predictive performance, as it overreacts due to the training data.
- K-fold cross validation
    - Generalization to k subsets.
    - k-1 subsets for training and 1 subset for testing.
- Leave-P-Out
    - Take p data out from the input dataset of n data, n-p for training and p for testing.

2.3. K-nearest Neighbor Classification

2.3.1. Definition

- A non-parametric algorithm often used for classification which uses proximity to make classifications or predictions about the grouping of an individual data point.
- A class label is assigned based on a majority vote, i.e., the label that is most frequently represented among the k nearest neighbors is used.
- The distance between data points is typically calculated using distance metric like Euclidean distance.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

2.3.2. Algorithm

1. Choose k number of neighbors to consider.
2. Store the training dataset, which consists of labeled data points with their corresponding classes.
3. For a new, unlabeled data point, calculate its distance to all data points in the training set using a distance metric.
4. Sort the distances from closest to furthest.
5. Identify the k-nearest neighbors based on the calculated distances.
6. Repeat step 3, 4 and 5.
7. Stopping criterion: All data points have been labeled.

➔ The time complexity: $O(n)$

- Calculating distance from a new data point to all n data points of dimension d: $O(n \cdot d) = O(n)$
- Sorting algorithm, using quicksort: $O(n \cdot \log(n))$
- Calculating the majority class: $O(k)$

2.4. Perceptron classifier

2.4.1. Definition

- An algorithm for supervised learning of binary classifiers.
- A binary classifier is a function which can decide whether an input, represented by a vector of numbers, belongs to some specific class.
- It's a type of linear classifier, i.e., a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector.

2.4.2. Algorithm

Assume all instances are points $x \in R^n$

Labels $y = \{-1,1\}$

Function $sign(x) = \begin{cases} 1, x > 0; \\ 0, x = 0; \\ -1, x < 0; \end{cases}$

1. Initial zero prediction vector $v = 0$
2. Predicts the new label of a new instance x to be $\hat{y} = sign(v \cdot x)$
3. If the prediction differs from the label y, update the prediction vector $v = v + yx$
4. Repeat step 2 and 3.
5. Stopping criterion: All data points have been labeled.

➔ The time complexity: $O(n)$

- Prediction or update for all data points n in d dimensions: $O(n \cdot d) = O(n)$

3. Data Analysis

3.1. Iris

- The dataset contains 150 instances and 4 features. Each instance is a plant, and the Predicted attributes are classes of iris plant.

Table 1. Characteristics of Iris dataset.

Data	Role	Data Type	Mean	Variance
Sepal length	Feature	Continuous, Quantitative	5.843333	0.685694
Sepal width	Feature	Continuous, Quantitative	3.057333	0.189979

Petal length	Feature	Continuous, Quantitative	3.758000	3.116278
Petal width	Feature	Continuous, Quantitative	1.199333	0.581006
Class	Target	Categorical, Qualitative		

Table 2. Explained variance by the principal components of Iris dataset.

Principal Components	Explained Variance
PC1	0.52875361
PC2	0.52875361

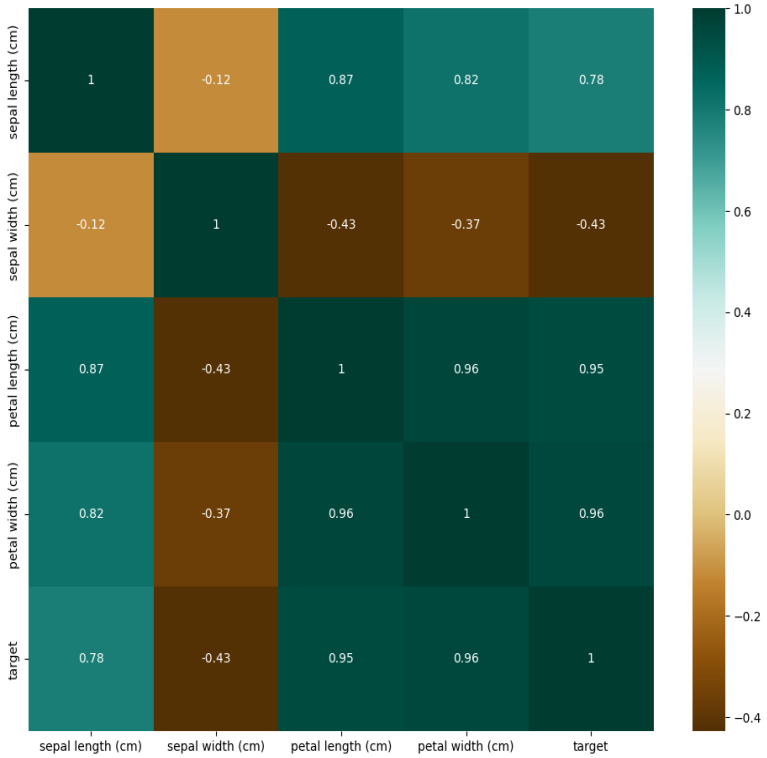

Figure 1. Iris dataset. Correlation matrix.

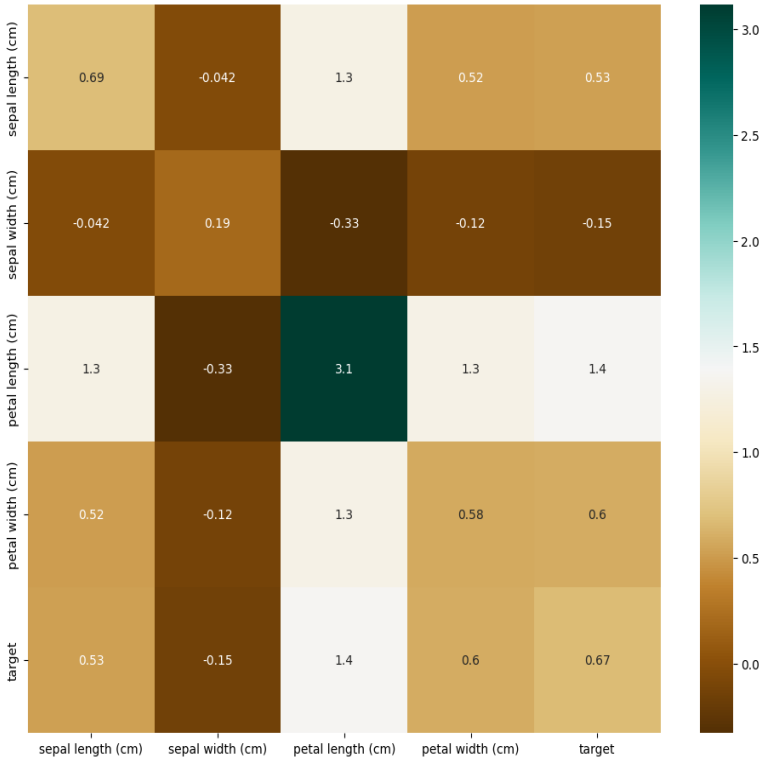

Figure 2. Iris dataset. Covariance matrix.

3.2. Breast Cancer Wisconsin (Diagnostic)

- The dataset contains 569 instances and 30 features. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Table 3. Characteristics of Breast Cancer Wisconsin (Diagnostic) dataset.

Data	Role	Data Type	Mean	Variance
mean radius	Feature	Continuous, Quantitative	14.12729	12.41892
mean texture	Feature	Continuous, Quantitative	19.28965	18.49891
mean perimeter	Feature	Continuous, Quantitative	91.96903	590.4405
mean area	Feature	Continuous, Quantitative	654.8891	123843.6

mean smoothness	Feature	Continuous, Quantitative	0.09636	0.000198
mean compactness	Feature	Continuous, Quantitative	0.104341	0.002789
mean concavity	Feature	Continuous, Quantitative	0.088799	0.006355
mean concave points	Feature	Continuous, Quantitative	0.048919	0.001506
mean symmetry	Feature	Continuous, Quantitative	0.181162	0.000752
mean fractal dimension	Feature	Continuous, Quantitative	0.062798	4.98E-05
radius error	Feature	Continuous, Quantitative	0.405172	0.076902
texture error	Feature	Continuous, Quantitative	1.216853	0.304316
perimeter error	Feature	Continuous, Quantitative	2.866059	4.087896
area error	Feature	Continuous, Quantitative	40.33708	2069.432
smoothness error	Feature	Continuous, Quantitative	0.007041	9.02E-06
compactness error	Feature	Continuous, Quantitative	0.025478	0.000321
concavity error	Feature	Continuous, Quantitative	0.031894	0.000911
concave points error	Feature	Continuous, Quantitative	0.011796	3.81E-05
symmetry error	Feature	Continuous, Quantitative	0.020542	6.83E-05
fractal dimension error	Feature	Continuous, Quantitative	0.003795	7E-06

worst radius	Feature	Continuous, Quantitative	16.26919	23.36022
worst texture	Feature	Continuous, Quantitative	25.67722	37.77648
worst perimeter	Feature	Continuous, Quantitative	107.2612	1129.131
worst area	Feature	Continuous, Quantitative	880.5831	324167.4
worst smoothness	Feature	Continuous, Quantitative	0.132369	0.000521
worst compactness	Feature	Continuous, Quantitative	0.254265	0.024755
worst concavity	Feature	Continuous, Quantitative	0.272188	0.043524
worst concave points	Feature	Continuous, Quantitative	0.114606	0.004321
worst symmetry	Feature	Continuous, Quantitative	0.290076	0.003828
worst fractal dimension	Feature	Continuous, Quantitative	0.083946	0.000326
Class	Target	Categorical, Qualitative		

Table 4. Explained variance by the principal components of Breast Cancer Wisconsin (Diagnostic) dataset.

Principal Components	Explained Variance
PC1	0.44272026
PC2	0.18971182

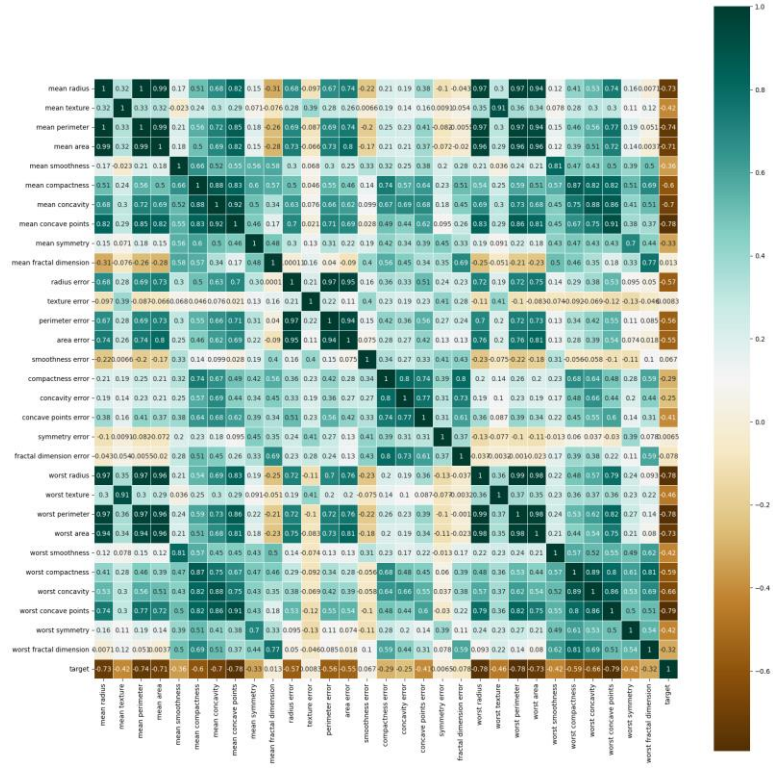

Figure 3. Breast Cancer Wisconsin (Diagnostic) dataset. Correlation matrix.

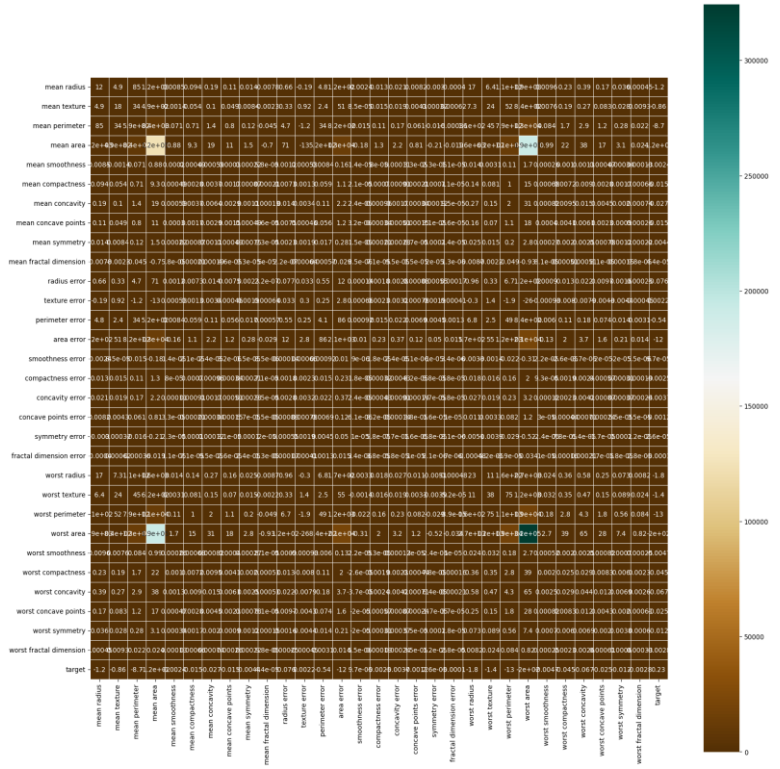

Figure 4. Breast Cancer Wisconsin (Diagnostic) dataset. Covariance matrix.

3.3. Wine

- The dataset contains 178 instances and 13 features. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.

Table 5. Characteristics of Wine dataset.

Data	Role	Data Type	Mean	Variance
alcohol	Feature	Continuous, Quantitative	13.00062	0.659062
malic_acid	Feature	Continuous, Quantitative	2.336348	1.248015
ash	Feature	Continuous, Quantitative	2.366517	0.075265
alcalinity_of_ash	Feature	Continuous, Quantitative	19.49494	11.15269
magnesium	Feature	Continuous, Quantitative	99.74157	203.9893
total_phenols	Feature	Continuous, Quantitative	2.295112	0.39169
flavanoids	Feature	Continuous, Quantitative	2.02927	0.997719
nonflavanoid_phenols	Feature	Continuous, Quantitative	0.361854	0.015489
proanthocyanins	Feature	Continuous, Quantitative	1.590899	0.327595
color_intensity	Feature	Continuous, Quantitative	5.05809	5.374449
hue	Feature	Continuous, Quantitative	0.957449	0.052245
od280/od315_of_diluted_wines	Feature	Continuous, Quantitative	2.611685	0.504086

proline	Feature	Continuous, Quantitative	746.8933	99166.72
Class	Target	Categorical, Qualitative		

Table 5. Explained variance by the principal components of Wine dataset.

Principal Components	Explained Variance
PC1	0.36198848
PC2	0.1920749

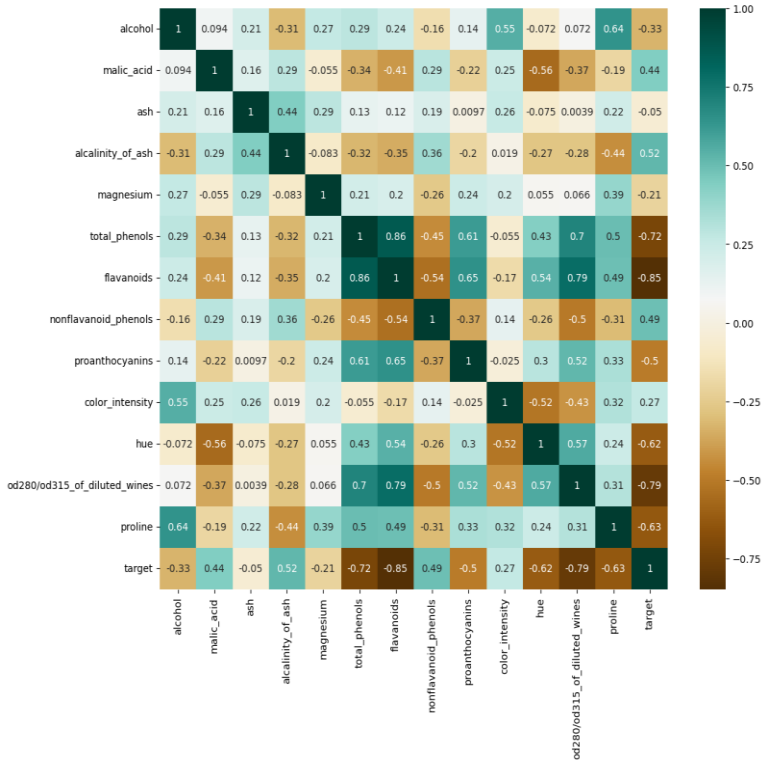

Figure 5. Wine dataset. Correlation matrix.

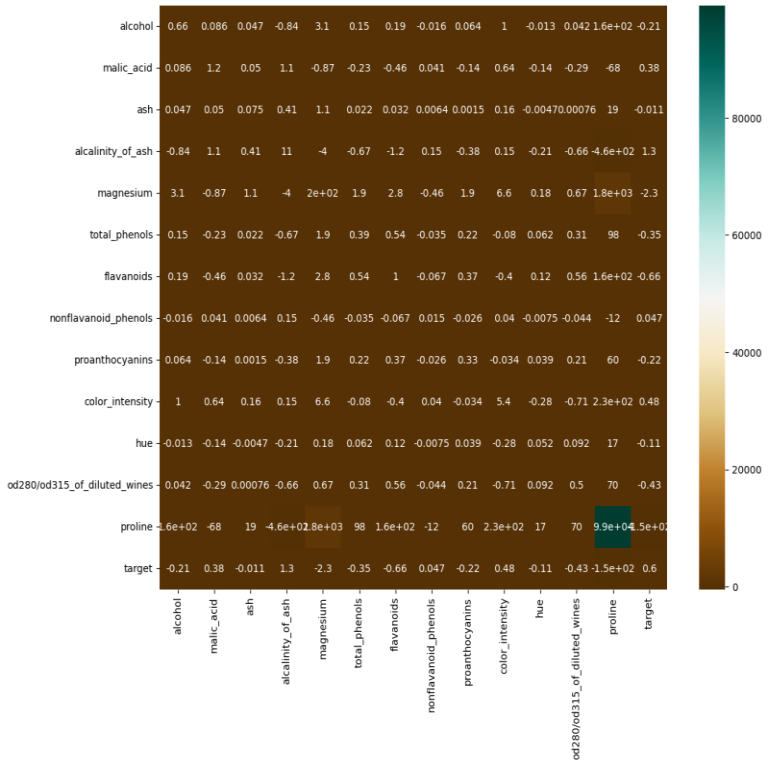

Figure 6. Wine dataset. Covariance matrix.

4. Experiment

4.1. Protocol

4.1.1. KNN

- Retrieve the datasets.
- Select 2 features. Plot for the original data visualization. Try Normalize the dataset. Split the data for training and testing. Plot the result after applying K-nn algorithm. Trying different value k.
- For improve performance apply PCA or SVD. Select 2 principal component for data visualization. Then split the data for training and testing. Plot the result after applying K-nn algorithm, trying different value of k

4.1.2. Perception

- Retrieve the datasets.
- Apply PCA to the dataset.
- Plot for the original data visualization\
- Split the data for training and testing. Plot the result after applying perceptron algorithm,

4.2. Results

4.2.1. Iris - KNN

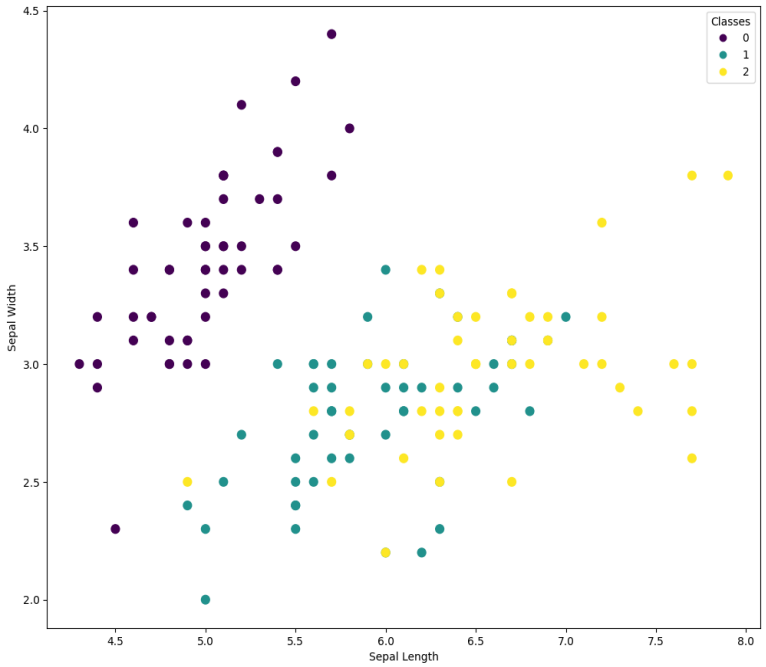

Figure 7. Iris dataset. Sepal Width and Sepal Length scatter plot.

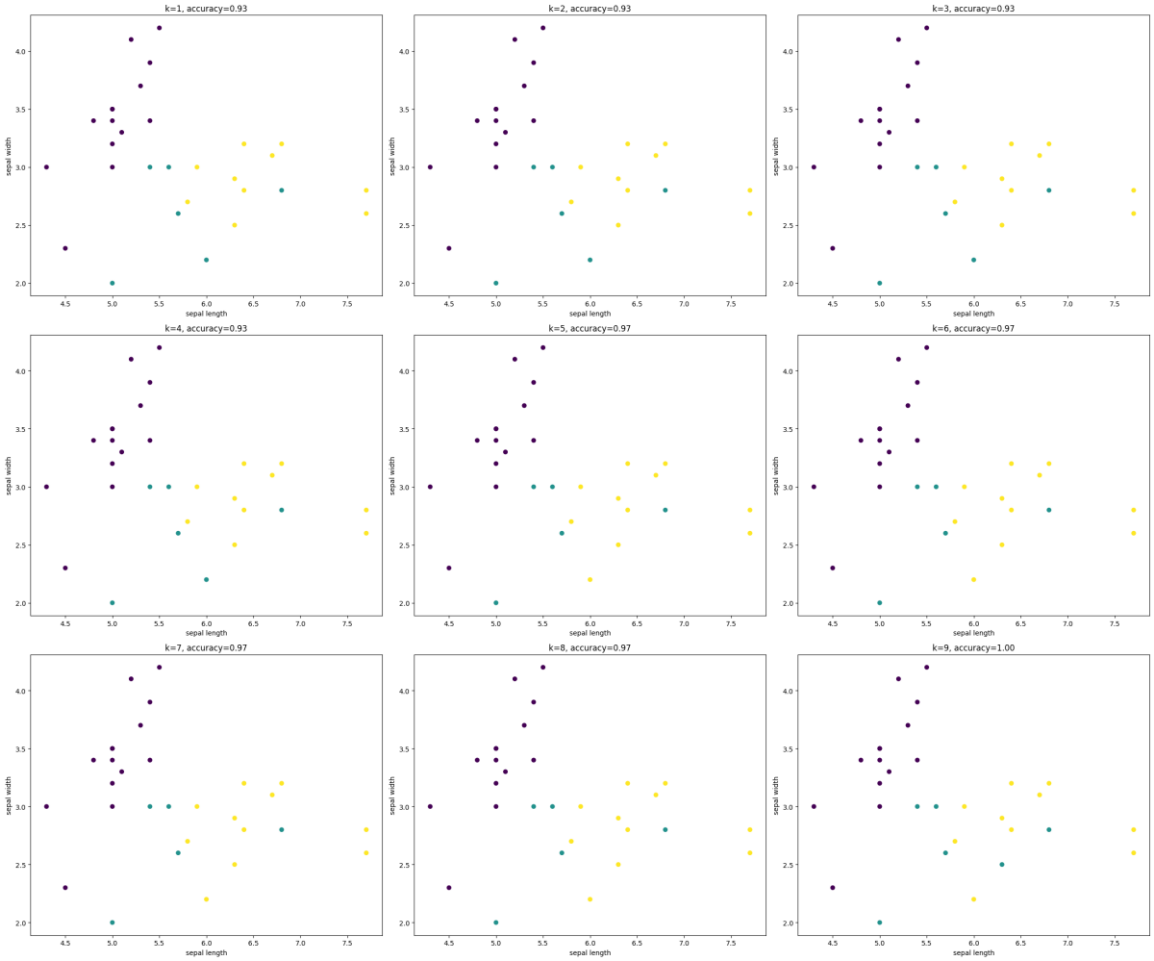

Figure 8. Iris dataset. Sepal Width and Sepal Length scatter plot. Applied K-nn with different value of k and their corresponding accuracy.

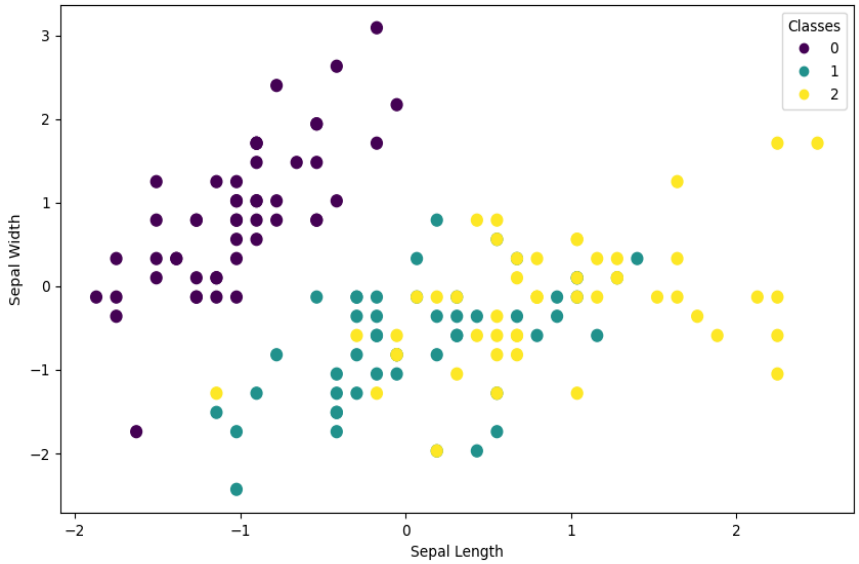

Figure 9. Iris dataset. Sepal Width and Sepal Length scatter plot after normalization.

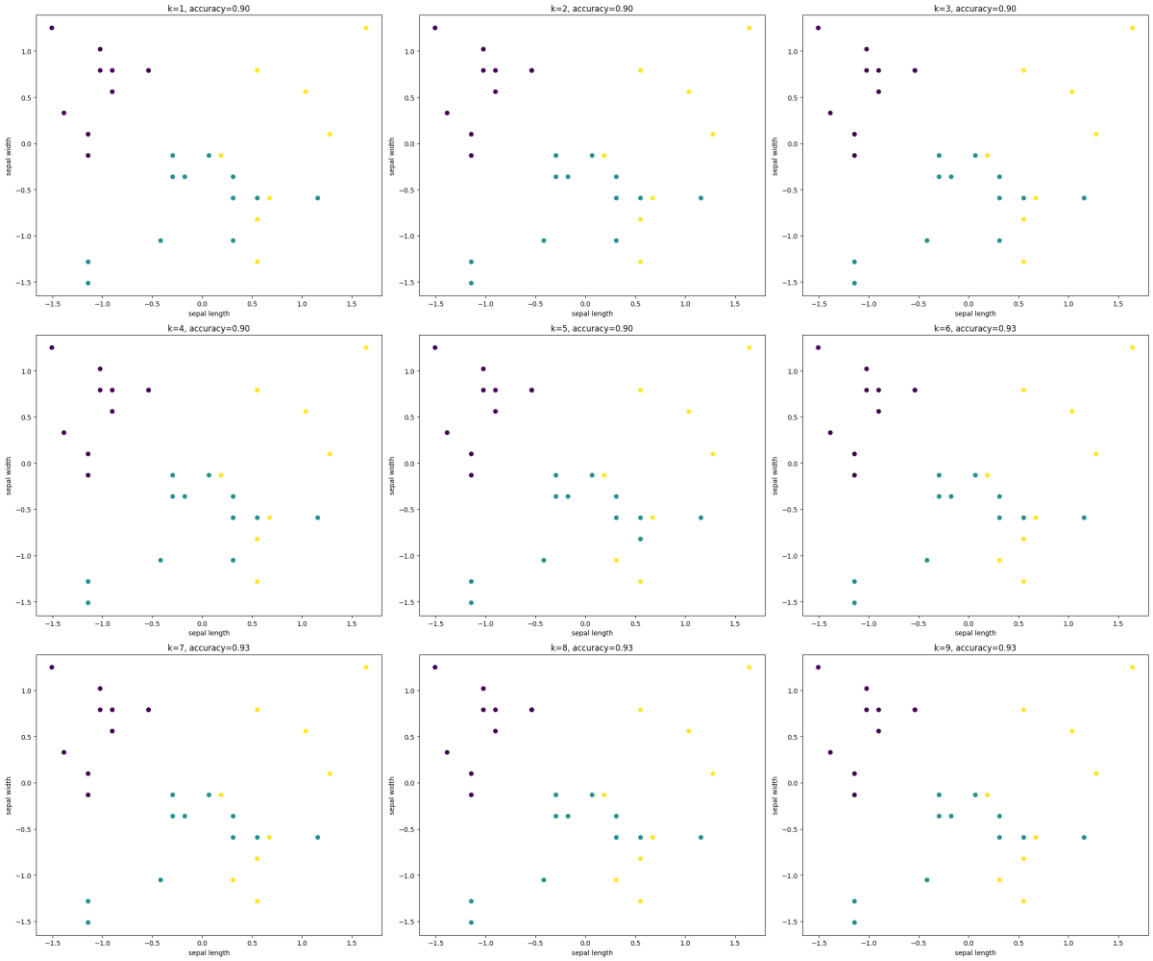

Figure 10. Iris dataset. Sepal Width and Sepal Length scatter plot with normalization. Applied K-nn with different value of k and their corresponding accuracy.

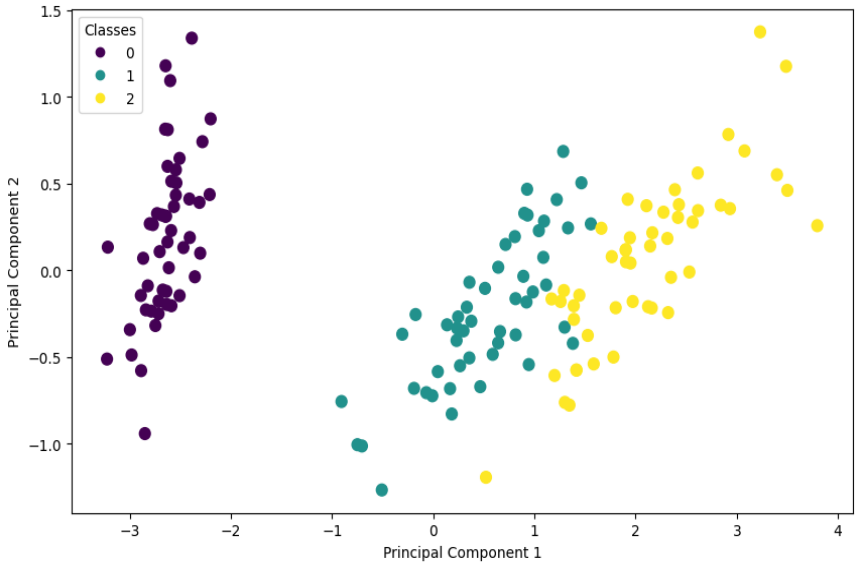

Figure 11. Iris dataset. Principal component 1 and 2 scatter plot after PCA.

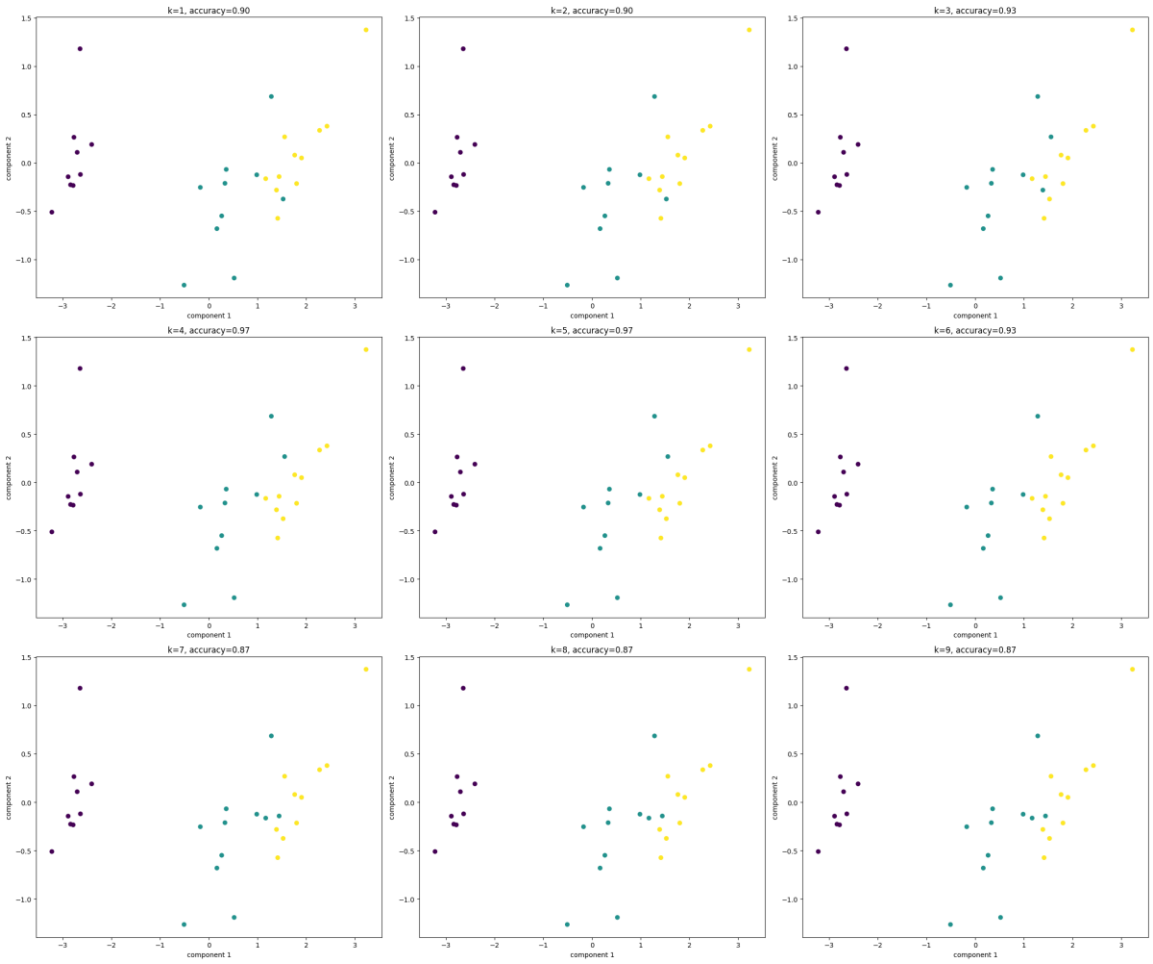

Figure 12. Iris dataset. Principal component 1 and 2 scatter plot after PCA. Applied K-nn with different value of k and their corresponding accuracy.

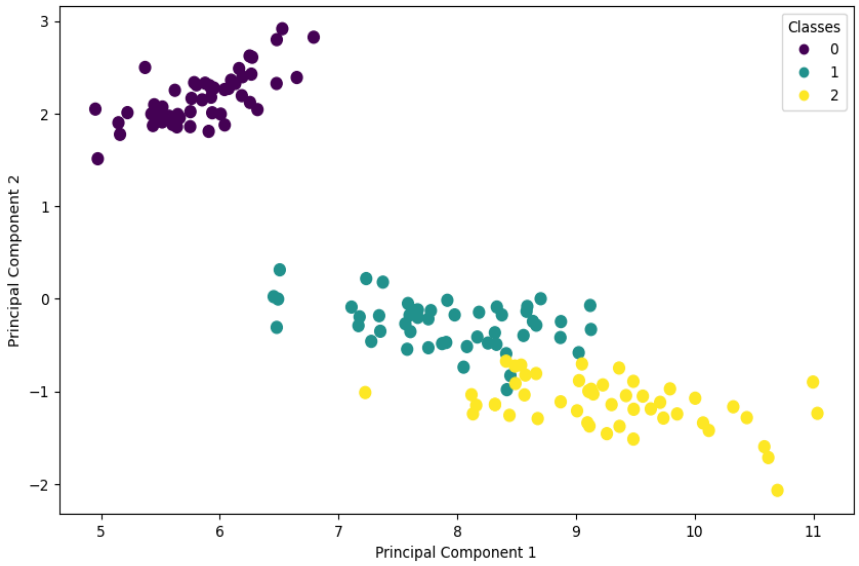

Figure 13. Iris dataset. Principal component 1 and 2 scatter plot after SVD.

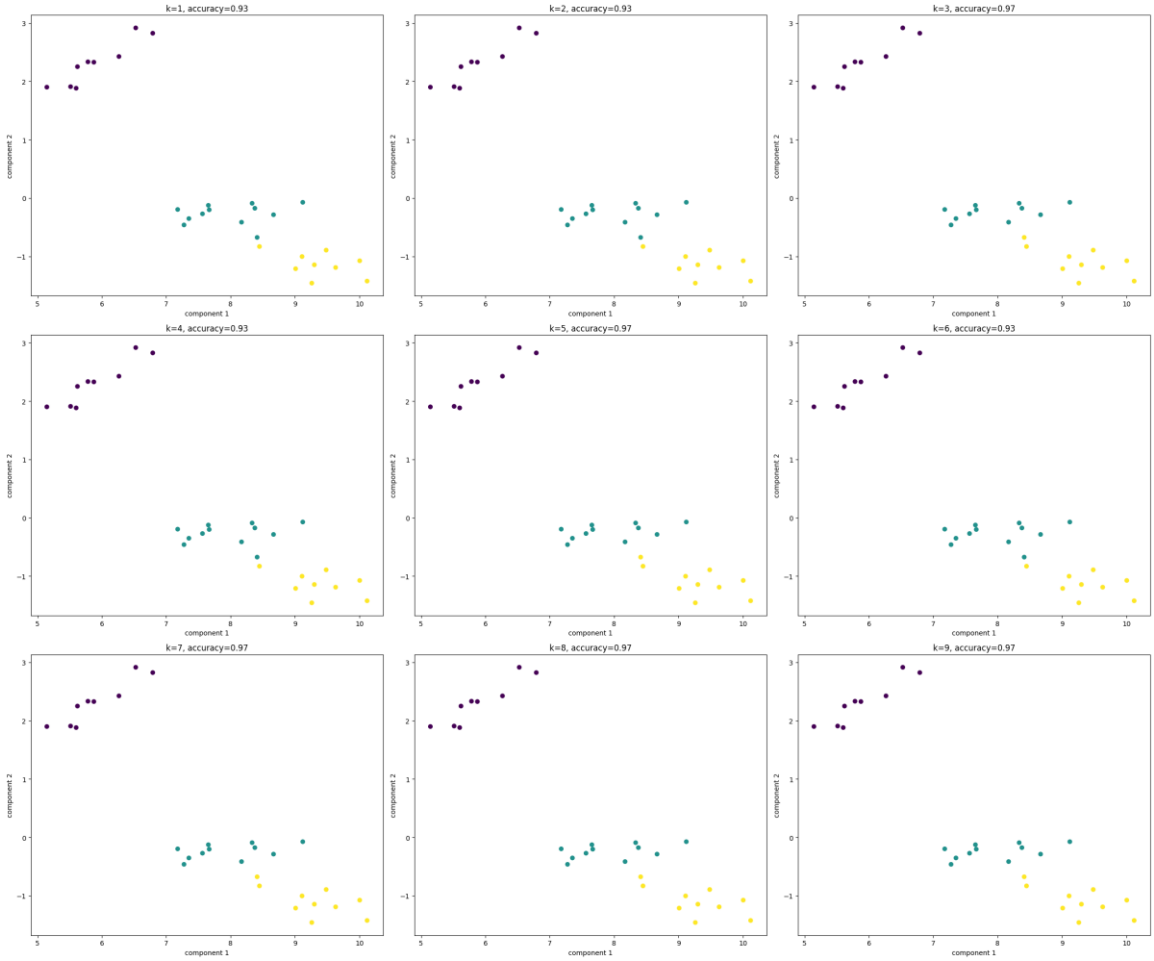

Figure 14. Iris dataset. Principal component 1 and 2 scatter plot after SVD. Applied K-nn with different value of k and their corresponding accuracy.

4.2.2. Breast Cancer Wisconsin (Diagnostic) - KNN

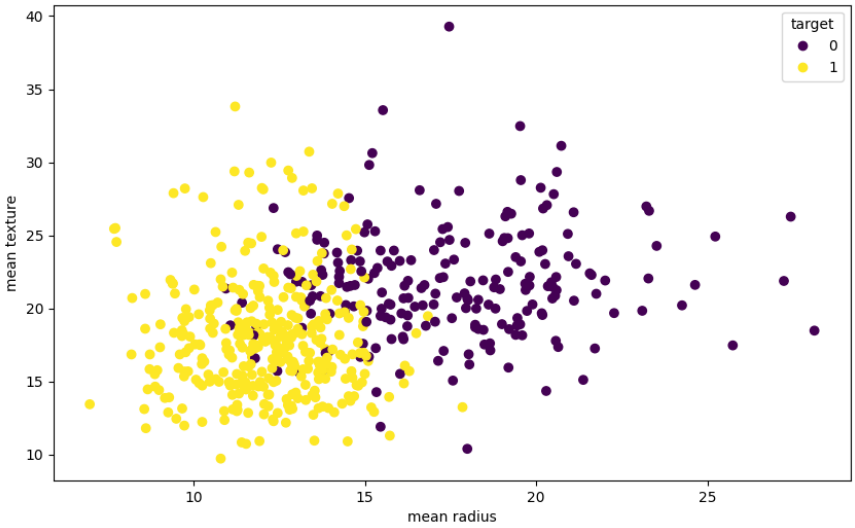

Figure 15. Breast Cancer Wisconsin (Diagnostic) dataset. Mean Texture and Mean Radius scatter plot.

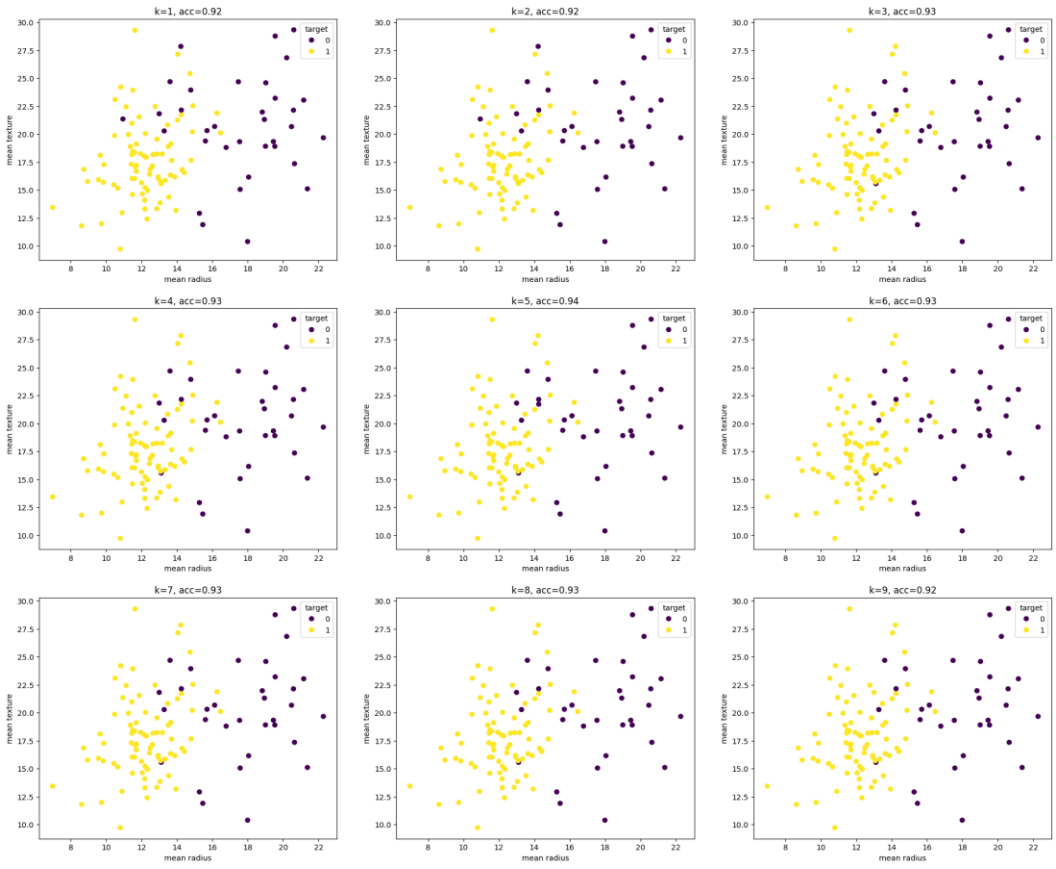

Figure 16. Breast Cancer Wisconsin (Diagnostic) dataset. Mean Texture and Mean Radius scatter plot. Applied K-nn with different value of k and their corresponding accuracy.

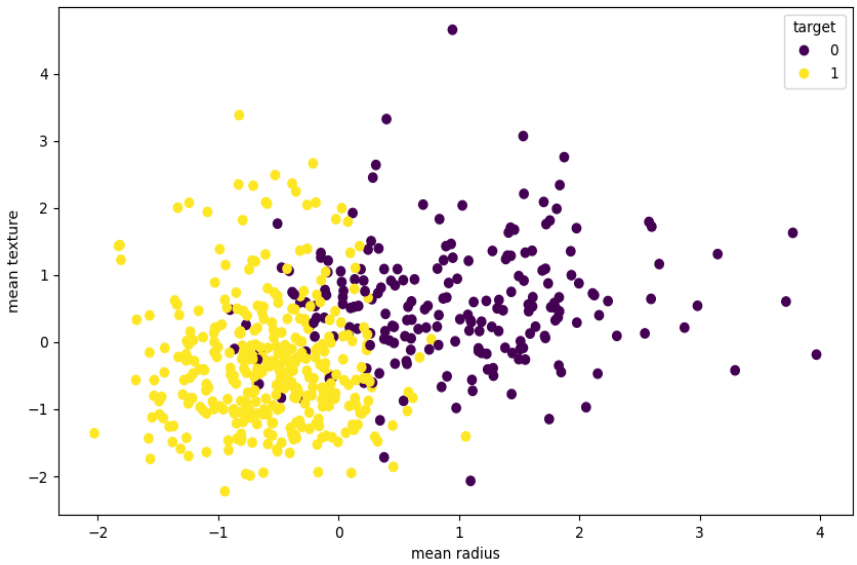

Figure 17. Breast Cancer Wisconsin (Diagnostic) dataset. Mean Texture and Mean Radius scatter plot after normalization.

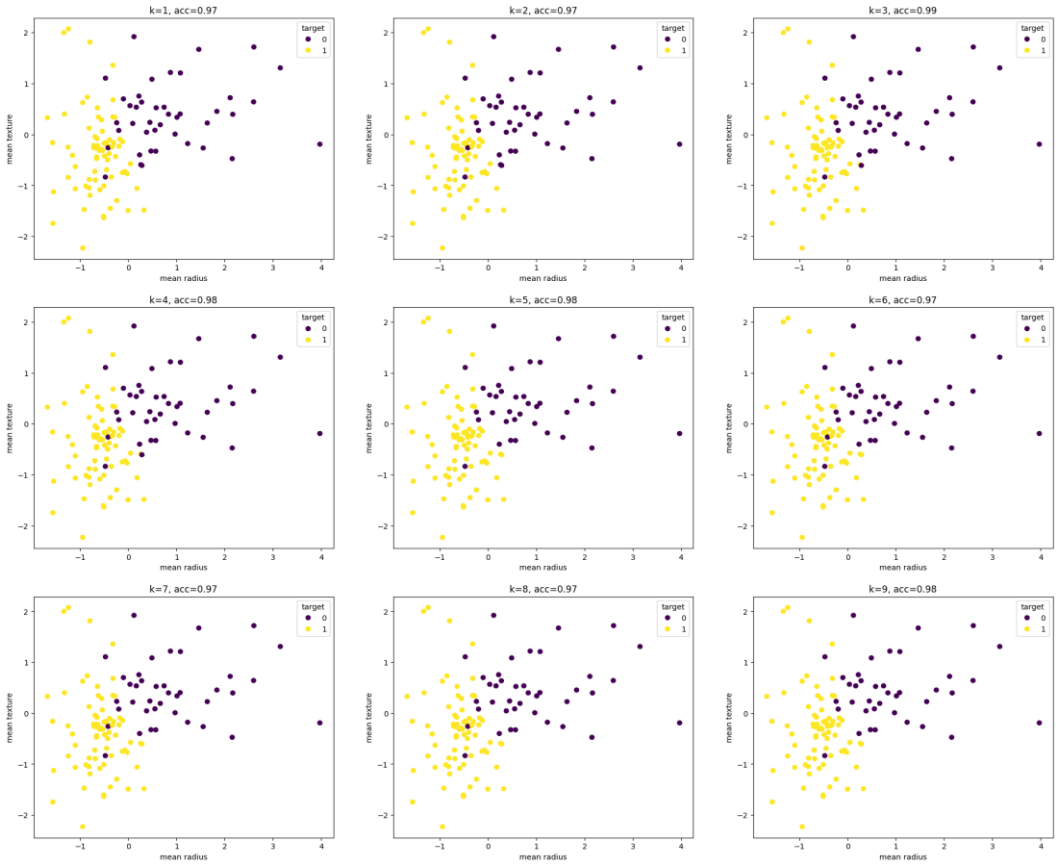

Figure 18. Breast Cancer Wisconsin (Diagnostic) dataset. Mean Texture and Mean Radius scatter plot with normalization. Applied K-nn with different value of k and their corresponding accuracy.

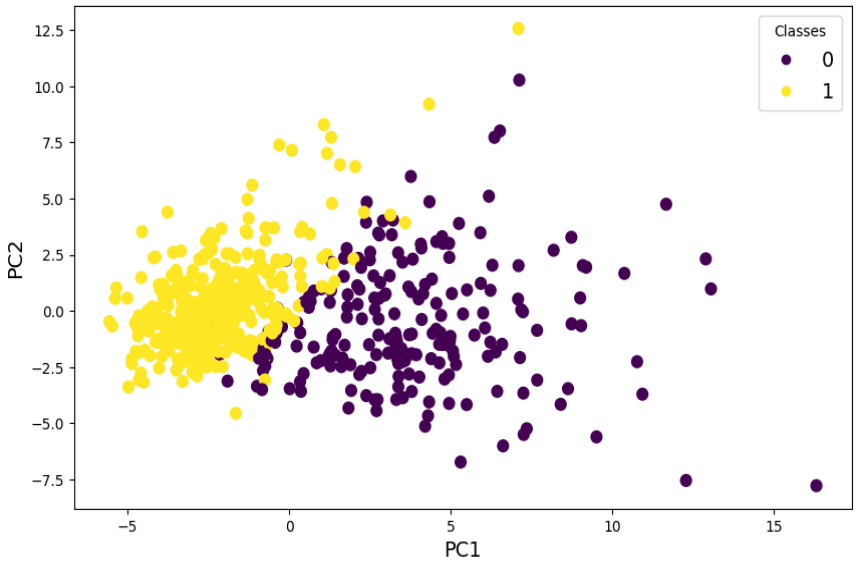

Figure 19. Breast Cancer Wisconsin (Diagnostic) dataset. Principal component 1 and 2 scatter plot after PCA.

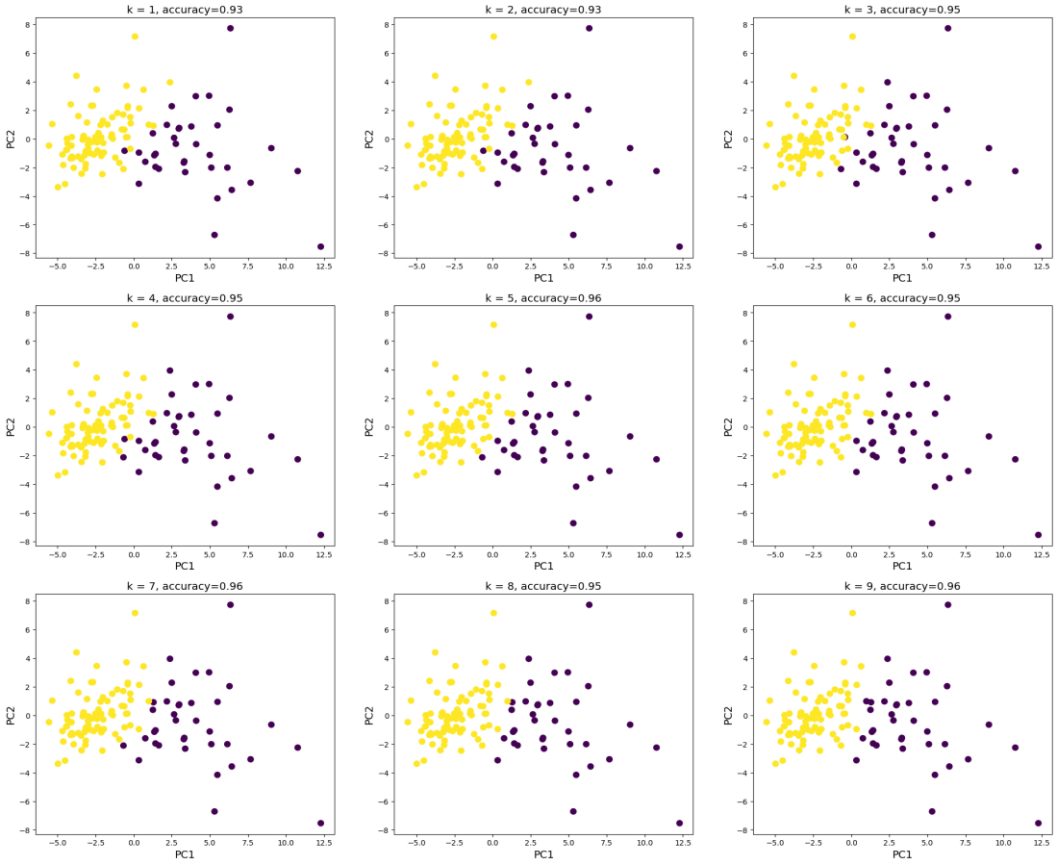

Figure 20. Breast Cancer Wisconsin (Diagnostic) dataset. Principal component 1 and 2 scatter plot after PCA. Applied K-nn with different value of k and their corresponding accuracy.

4.2.3. Wine - KNN

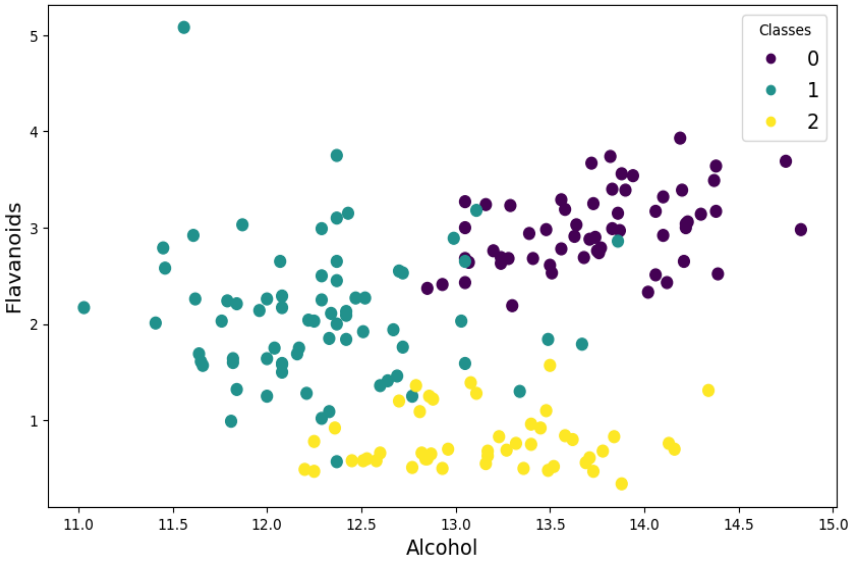

Figure 21. Wine dataset. Alcohol and Flavonoids scatter plot.

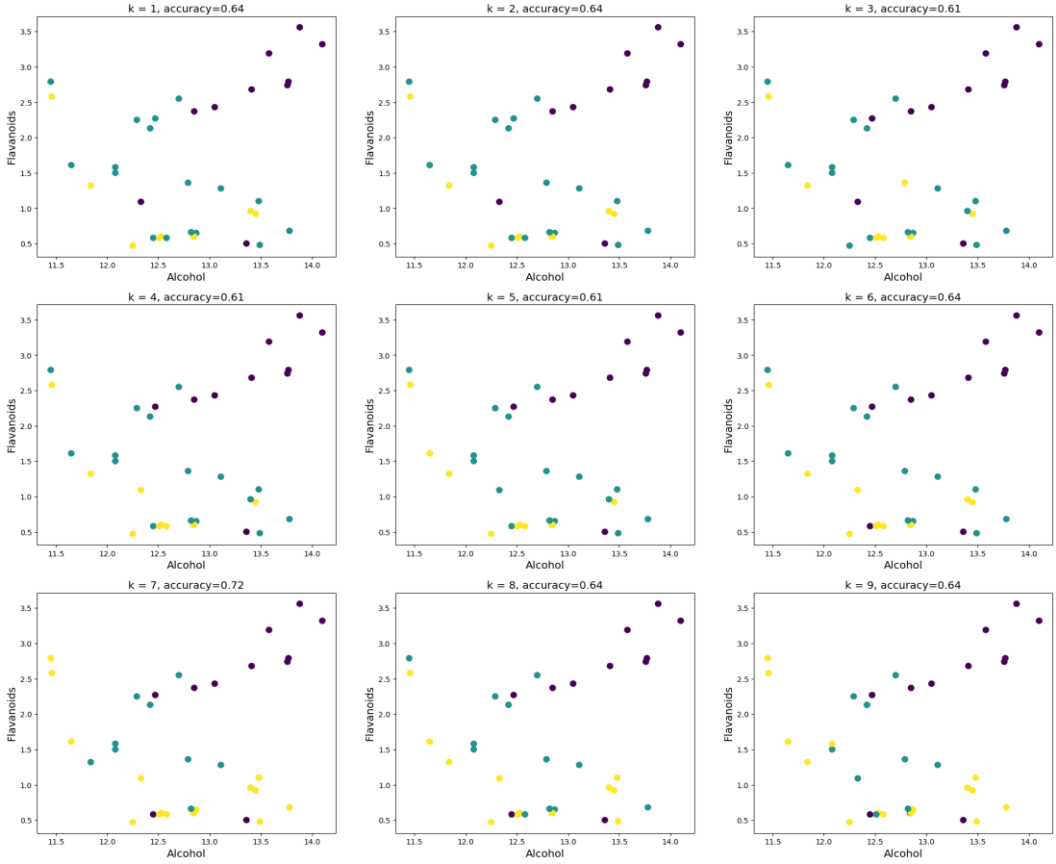

Figure 22. Wine dataset. Alcohol and Flavonoids scatter plot. Applied K-nn with different value of k and their corresponding accuracy.

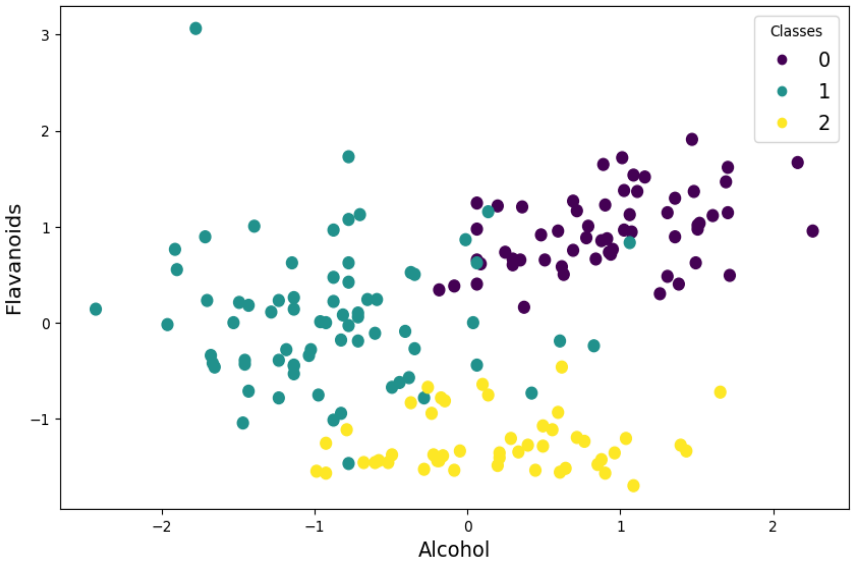

Figure 23. Wine dataset. Alcohol and Flavonoids scatter plot with normalization.

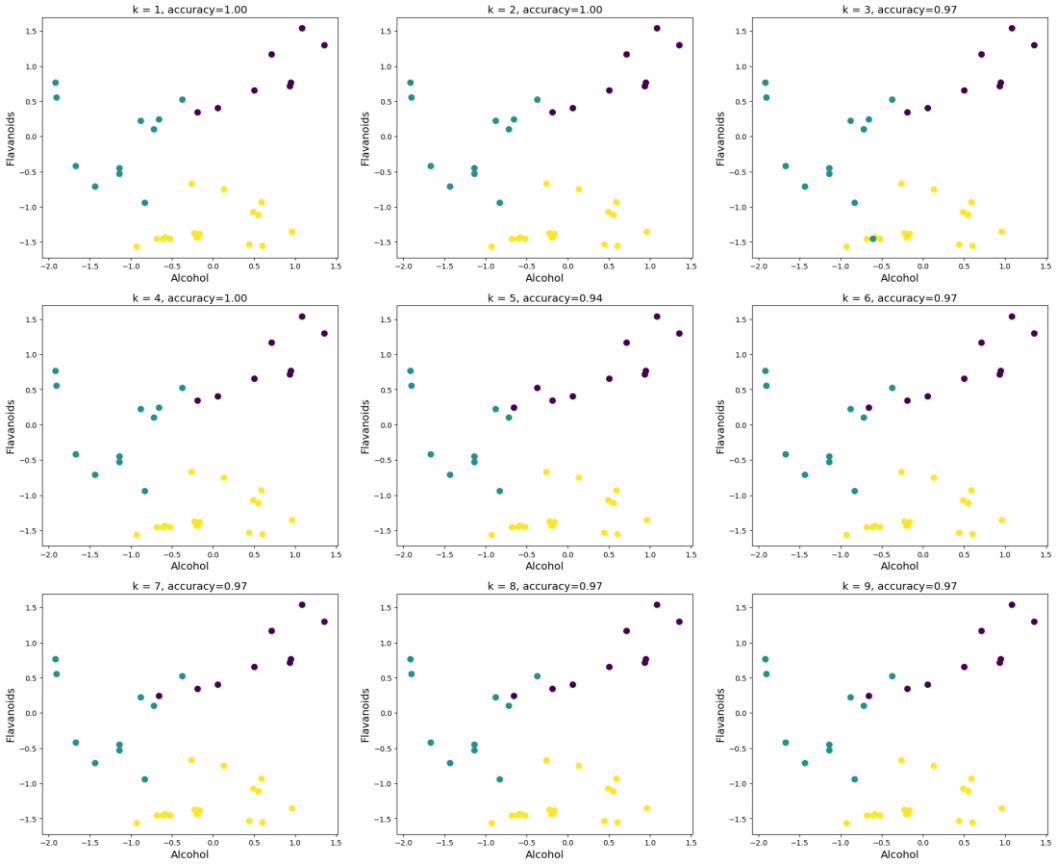

Figure 24. Wine dataset. Alcohol and Flavonoids scatter plot with normalization. Applied K-nn with different value of k and their corresponding accuracy.

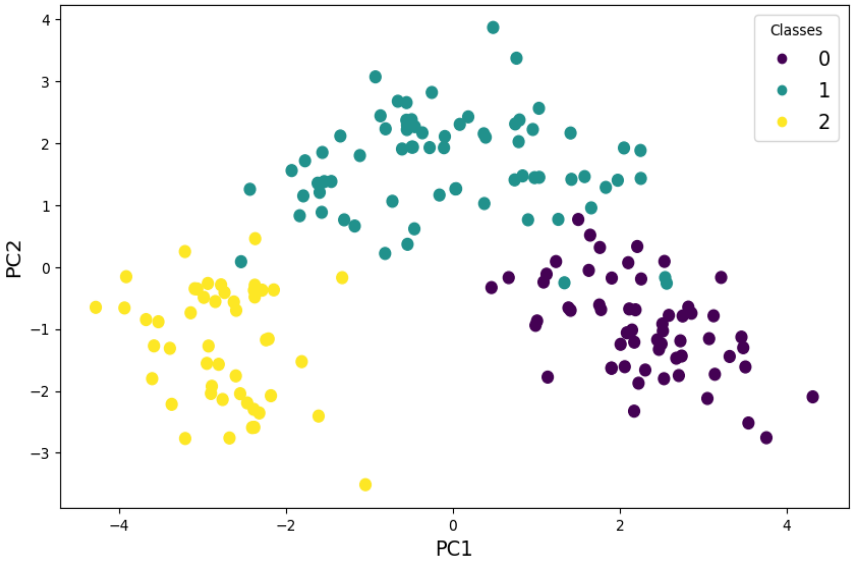

Figure 25. Wine dataset. Principal Components 1 and 2 scatter plot after PCA.

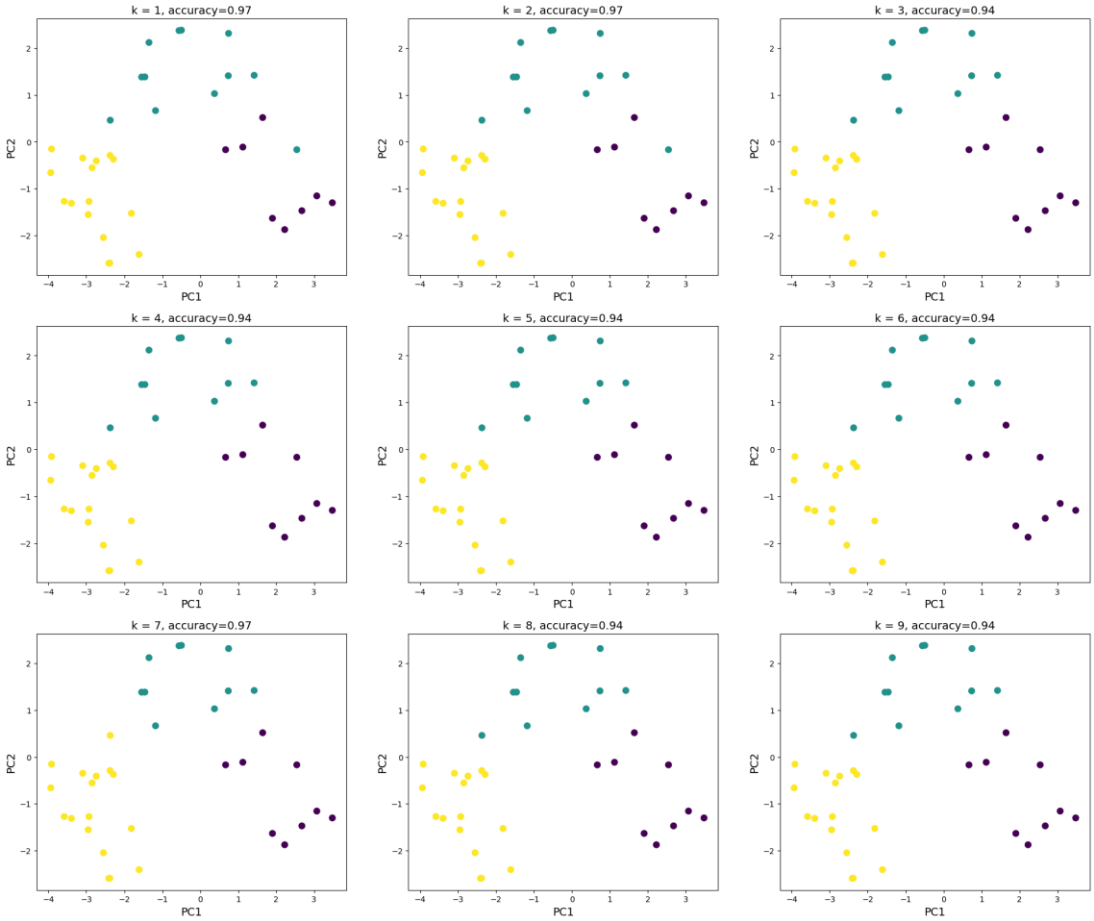

Figure 26. Wine dataset. Principal Components 1 and 2 scatter plot after PCA. Applied K-nn with different value of k and their corresponding accuracy.

4.2.4. Iris – Perception

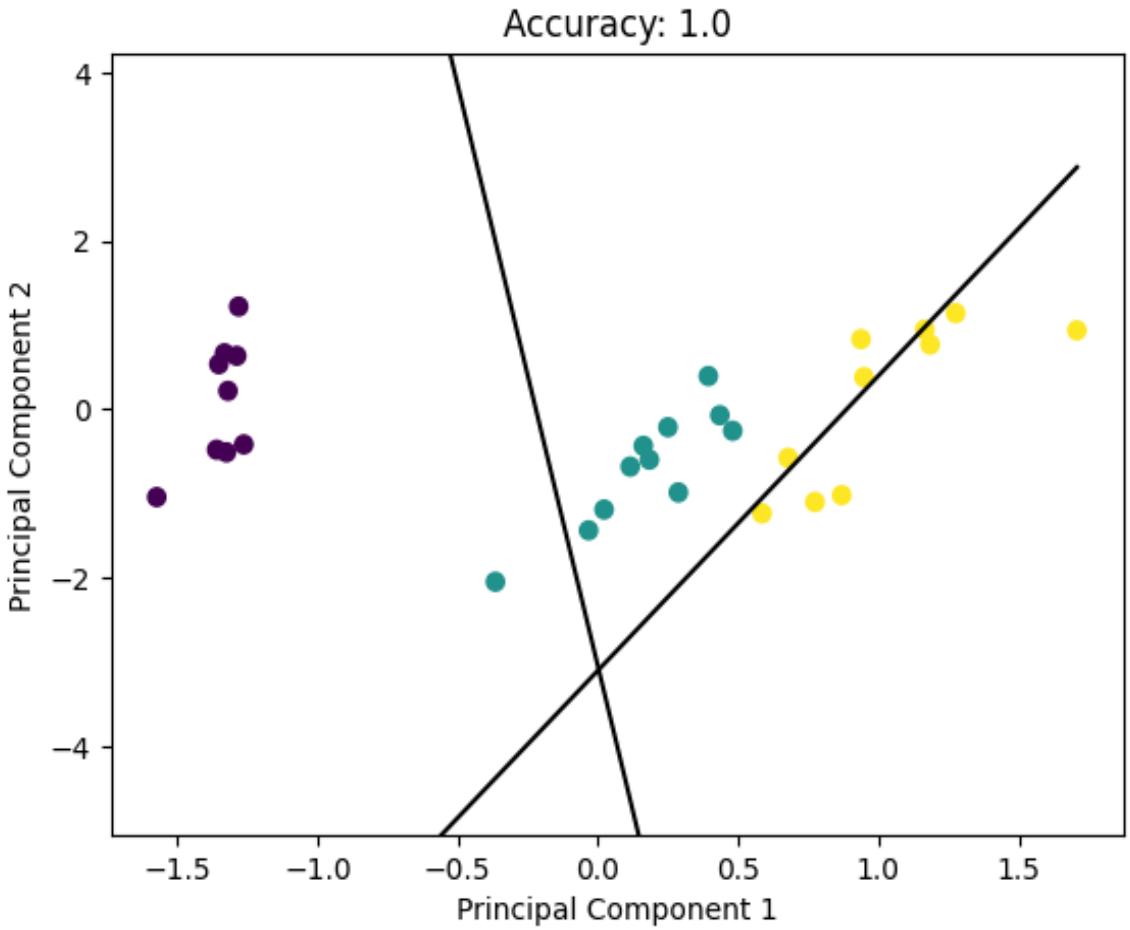

Figure 27. Iris dataset. Principal Components 1 and 2 scatter plots after PCA. Applied Perception Classifier with learning rate = 0.001 and number of iterations = 100.

4.2.5. Breast Cancer Wisconsin (Diagnostic) - Perceptron

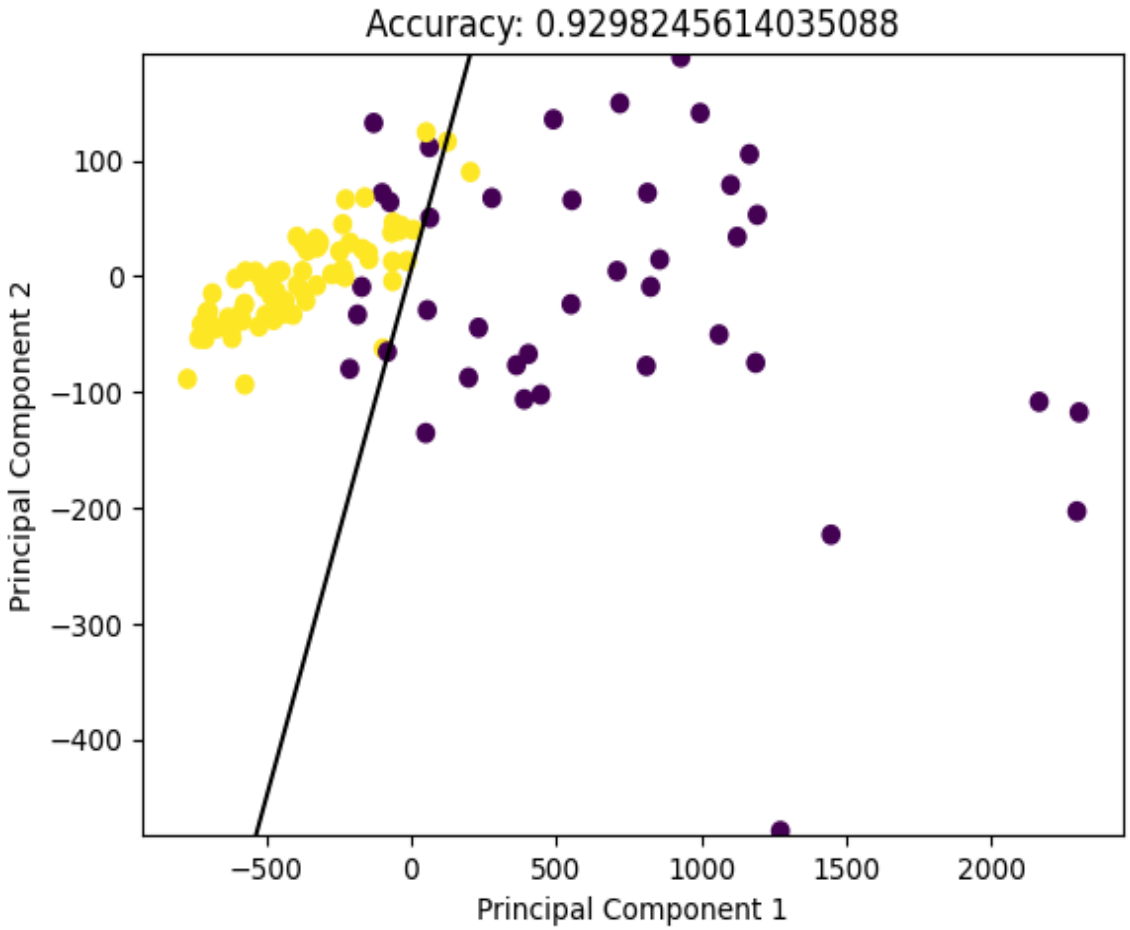

Figure 28. Breast Cancer Wisconsin (Diagnostic) dataset. Principal Components 1 and 2 scatter plots after PCA. Applied Perception Classifier with learning rate = 0.001 and number of iterations = 100.

4.2.6. Wine - Perceptron

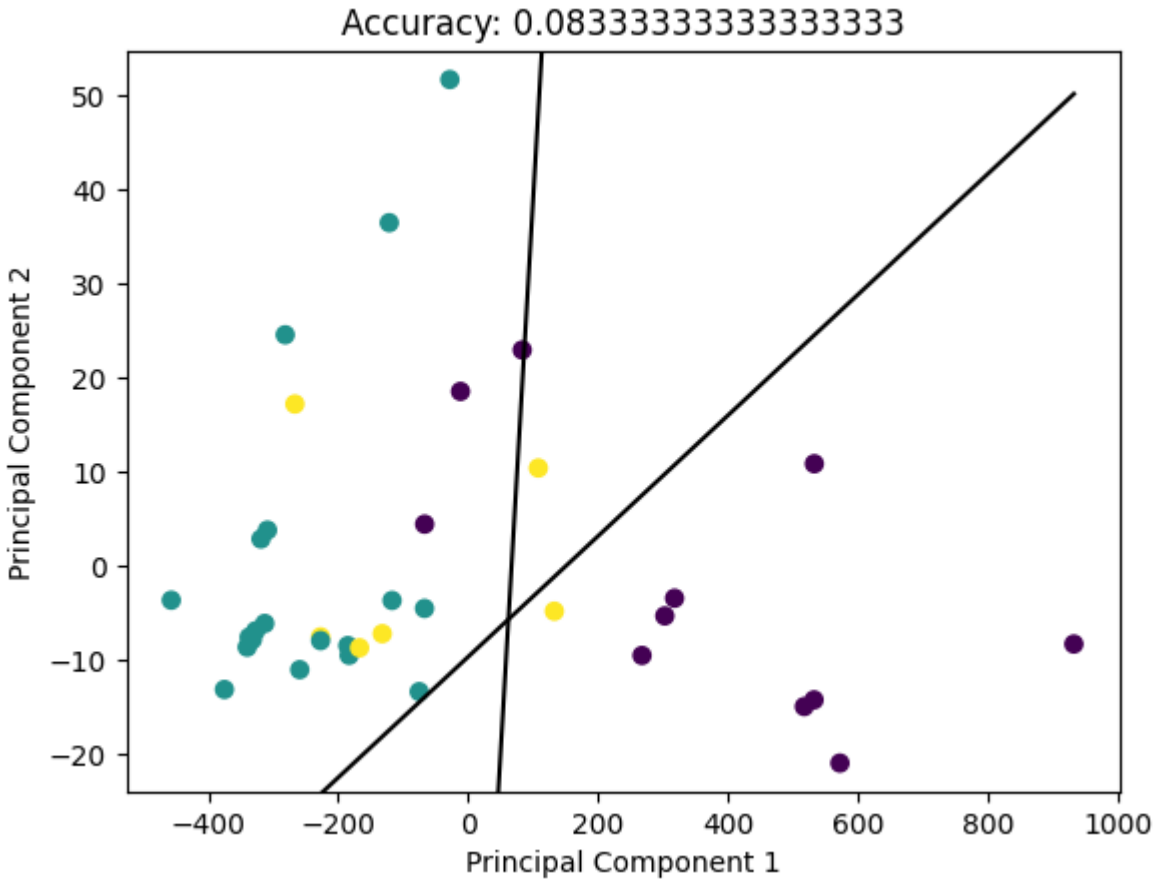

Figure 29. Wine dataset. Principal Components 1 and 2 scatter plots after PCA. Applied Perception Classifier with learning rate = 0.001 and number of iterations = 1000.

5. Evaluation

5.1. KNN

- After normalization can help improve KNN performance and indeed from figures 10, 17, 24 of the three different datasets. And after PCA or SVD help improve KNN performance significantly and indeed from figures 12, 20. 26 And the reasons are:
    1. Uniform Scale: KNN is a distance-based algorithm, meaning it calculates the distance between points to make predictions. If features are on different scales, the feature with a larger scale can dominate when calculating distances. By bringing all features to a uniform scale, normalization ensures that no feature dominates due to its scale.
    2. Speed: Normalization can also speed up training. If features are on different scales, update at different rates for each feature, which can cause the algorithm to take longer to find the optimal model.

3. Convergence: In some cases, normalization can help the algorithm converge at all. If the scales of the features are very different, larger scale features can explode, causing the algorithm to fail to converge.
4. Avoiding the Curse of Dimensionality: In high dimensional spaces, points tend to be far away from each other. This can make the distance calculations in KNN less meaningful. PCA helps to mitigate this problem by reducing the dimensionality of space.

-

Table 6. Iris dataset. Accuracy for k-folds with k=5.

Data	Accuracy
Original	0.9534
Normalized	0.9465
PCA	0.9667
SVD	0.9667

Table 7. Iris dataset. Accuracy for Leave-P-Out with p=1

Data	Accuracy
Original	0.9667
Normalized	0.9467
PCA	0.9667
SVD	0.9667

Table 8. Breast Cancer Wisconsin (Diagnostic) dataset. Accuracy for k-folds with k=5.

Data	Accuracy
Original	0.9385
Normalized	0.9683
PCA	0.9385
SVD	0.9332

Table 9. Breast Cancer Wisconsin (Diagnostic) dataset. Accuracy for Leave-P-Out with p=1

Data	Accuracy
Original	0.9332
Normalized	0.9701
PCA	0.9402
SVD	0.9402

Table 10. Wine dataset. Accuracy for k-folds with k=5.

Data	Accuracy
Original	0.7190
Normalized	0.9719
PCA	0.9549
SVD	0.9550

Table 11. Wine dataset. Accuracy for Leave-P-Out with p=1

Data	Accuracy
Original	0.7191
Normalized	0.9719
PCA	0.9606
SVD	0.9606

References

[1] Fisher,R. A.. (1988). Iris. UCI Machine Learning Repository. https://doi.org/10.24432/C56C76.

[2] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

[3] Aeberhard,Stefan and Forina,M.. (1991). Wine. UCI Machine Learning Repository. https://doi.org/10.24432/C5PC7J.

[4] Freund, Y.; Schapire, R. E. (1999). "Large margin classification using the perceptron algorithm" (PDF). Machine Learning. 37 (3): 277–296. doi:10.1023/A:1007662407062. S2CID 5885617.