A Practical Investigation of Principal Component Analysis (June 2024)

Kristina Ghimire (THA077BCT023)[1], Punam Shrestha (THA077BCT038)[1]

[1]Department of Electronics and Computer Engineering, IOE, Thapathali Campus, Kathmandu 44600, Nepal
Corresponding author: Kristina Ghimire(ghimirekristina10@gmail.com)

ABSTRACT Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction. Principal Component Analysis is widely used for feature extraction and face recognition process. This practical study of Principal Component Analysis presents various steps performed in the Principal Component Analysis method. Dimensionality reduction using Principal Component Analysis is performed on different types of datasets to explore its application. From normalizing the dimensions of a dataset to calculating eigenfaces or principal components, there are different steps involved in PCA. Eigenfaces or principal components are the projection of eigenvectors having the highest eigenvalues on the datasets. The use of eigenfaces reduces dimension but may lose some information from the dataset. This experiment shows that performing PCA and reducing dimensionality makes the analysis simple, helps in noise reduction, improves computational accuracy, and speeds up the process. This paper provides a foundation for analyzing the use of PCA on multivariate datasets and applying PCA for various applications like face recognition, face expression identification, and pattern recognition.

INDEX TERMS Dimensionality reduction, Eigenfaces, Feature extraction, Principal Component Analysis

I INTRODUCTION

Principal Component Analysis (PCA) is a classical multivariate (unsupervised machine learning) non-parametric dimensionality reduction method that is used to interpret the variation in high-dimensional interrelated datasets (datasets with many variables). PCA reduces the high-dimensional interrelated data to low-dimension by linearly transforming the old variable into a new set of uncorrelated variables called principal component (PC) while retaining the most possible variation.

It offers a sequence of best linear approximations to high dimensional observations. The algorithm for Principal Component Analysis (PCA) is based on finding orthogonal directions that explain the maximum variance in the data. In the context of dimensionality reduction, the objective is to find 'm' orthonormal directions that minimize the representation error. Dimensionality reduction helps in analyzing high-dimensional datasets and makes it easier to capture the essential features within the data while reducing computational complexity. Learning the ways of PCA will enable students to become aware of dimensionality reduction techniques as we guide them to build a fundamental understanding for learning about implementing it in more complex algorithms. High-dimensional datasets make it challenging to analyze and interpret data effectively. Dimensionality reduction techniques, such as Principal Component analysis (PCA), extract only relevant features that capture the critical features within the data by removing low-variance and irrelevant features. Reducing the dimension makes it less complex, and easy to understand. It offers steps to understand the impact of each feature on the target and extract only the features that carry most of the data (approx. 90% variance of proportion).

The Principal Component Analysis (PCA) process is based on finding orthogonal directions that explain the maximum variance in the data. Understanding PCA for dimensionality reduction provides students with a valuable understanding of its process and its practical use. The hands-on experience of working with real or random datasets enables students to dive into the step-by-step process of PCA, analyzing the outputs, and extracting meaningful insights of patterns and structures of the data. Learning the foundational concepts of PCA provides students with a solid foundation that helps them explore more advanced dimensionality reduction algorithms and their practical implementations. This knowledge prepares them with the skills necessary to tackle complex data analysis challenges and strengthens their understanding of optimizing data while keeping critical information.

II RELATED WORK

In dimensionality reduction and feature extraction algorithms, Principal Component Analysis (PCA) has been extensively studied and used.

Liton Chandra Paul et al. [1] performed a comprehensive study on the Principal Component Analysis (PCA) method's methodological analysis and showed stimulation on data using MATLAB. The author listed the detailed steps to perform the PCA on high dimensional data and created a dataset with less dimension representing each data as an eigenface. The eigenface is obtained from the eigenvector projection with the highest eigenvalue on the original data. The original data is preserved if we take all the eigenvectors but if only eigenvectors with the highest eigenvalue are selected, there is dimension reduction.

Sukanya Sagarika Meher et al. [2] analyzed the Principal Component Analysis (PCA) method and experimented with its performance on Face Recognition and Facial Expression Identification. Face recognition is a difficult problem because the feature extracted from the face is multidimensional and its classification is a complex process. The author converted the face data into a two-dimensional problem for easy analysis. Facial expressions were identified from the mouth, eyes, and eyebrows features. Expressions like happy, sad, disgusted, surprised, and neutral were identified. The use of fewer eigenfaces improved the efficiency of the analysis from the PCA method. Their experiment shows that the PCA method achieves a lower error rate for face recognition, provides simple calculations, and improves the execution speed.

These works contribute to the understanding and use of PCA as a powerful tool for dimensional reduction and feature extraction in various areas.

III METHODOLOGY

A DATASET DESCRIPTION

The performance evaluation of feature extraction algorithms based on Principal Component Analysis (PCA) involved conducting experiments on three datasets. The first dataset was a randomly generated 20x2 array, which served as a synthetic dataset for evaluating PCA's effectiveness in a controlled environment. The second dataset is the Irish dataset [3], which is useful for the classification of the dataset. This dataset contains 3 classes of iris plants: Iris Setosa, Iris Versicolour, or Iris Virginica having 50 instances each. There are 4 features: sepal length, sepal width, petal length, and petal width (in cm unit) to help classify the iris plants. The third dataset is Heart Failure Clinical Records [4], which is useful for classification, regression, and clustering tasks. There are 299 medical records of patients with heart failure 12 features that provide insights for knowing the cause of heart failure and one target that gives information if the patient died during the follow-up period or not. The use of the PCA process on these datasets was done to evaluate the improvement in feature extraction after the reduction of dimension.

B PROPOSED METHODOLOGY

Principal Component Analysis (PCA) is a statistical technique used to transform a dataset's features into a set of uncorrelated variables called principal components. These components are derived through an orthogonal transformation, where the first component captures the highest variance, the second component captures the second highest variance, and so on. In our study, the PCA algorithm was implemented from basic Python libraries to perform this transformation. The first step involves data preprocessing, where the dataset is loaded and irrelevant attributes are removed. To encode the crop labels, a mapping dictionary is created, enabling numerical representation. Next, the dataset is converted into a matrix form, facilitating further computations. Exploratory Data Anal y-axis is conducted through a pair plot, visualizing attribute relationships. Standardization is performed by subtracting attribute means, and the resulting standardized dataset is stored. The covariance matrix is computed, revealing attribute relationships. Eigenvalues and eigenvectors of the covariance matrix are calculated and sorted. By multiplying the standardized dataset with the sorted eigenvectors, a change of basis is achieved. The proportion of variance 2 explained by each eigenvalue is determined. Finally, PCA is applied to project the transformed data onto different combinations of principal components, generating scatter plots and 3D visualizations. Through this proposed method, the crop recommendation dataset can be effectively analyzed and visualized using PCA.

C MATHEMATICAL FORMULAE

Let's assume we have a matrix or dataset with dimensions m×n. The formula for the mean () is:

$$\mu = x_{ij}mn \tag{1}$$

Where x_{ij} represents the element at the i^{th} row and j^{th} column of the matrix.

The formula for the covariance matrix (Σ) is:

$$\sum = \frac{1}{m-1}(X^T X) \tag{2}$$

For a square matrix A, the eigenvalues can be obtained by solving the characteristic equation:

$$\left|A - \lambda * I\right| = 0 \tag{3}$$

Where λ is the eigenvalue and I is the identity matrix of the same size as A. Given its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, the eigenvectors can be found by solving the equation:

$$(A - \lambda_i I)v_i = 0 \tag{4}$$

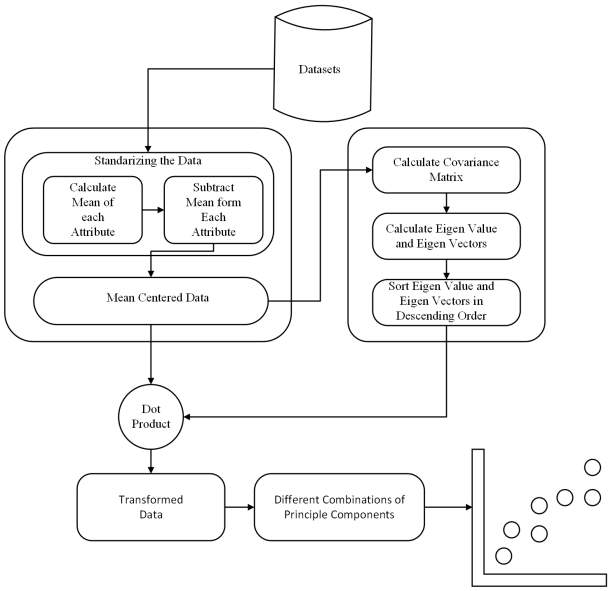

Figure 1: Block Diagram

The proportion of variance (PoV) for each eigenvalue λ_i is given by:

$$PoV_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \tag{5}$$

Where $\sum_{j=1}^{n} \lambda_j$ represents the sum of all eigenvalues.

D INSTRUMENTATION DETAILS

The experimentation was conducted using Jupyter Notebook. Python programming language is used to perform various steps of PCA methods. Many Python libraries were utilized for data analysis and visualization. The numpy library was imported to perform mathematical operations. It also supports matrices and other functions. For generating a random dataset, numpy.random was used which can also create data in different distribution forms like normal distribution. The pandas library was imported to operate on tabular format datasets and can manipulate values within the table. For data visualization, matplotlib.pyplot library was used to visualize pair plots, and 2D and 3D scatterplots on different combinations of features.

E SYSTEM BLOCK DIAGRAM

IV EXPERIMENTAL RESULTS

A PROBLEM 1: PCA ON RANDOM DATA

PCA on random data of size 20 by 2 showed that Principal Component Analysis (PCA) effectively reduced the dimensionality of random data from 2-dimensional to 1-dimensional. Several 2D scatterplots were generated to visualize the data features before and after PCA.

Then principal component is calculated by multiplying the data with the eigenvector of the covariance matrix.

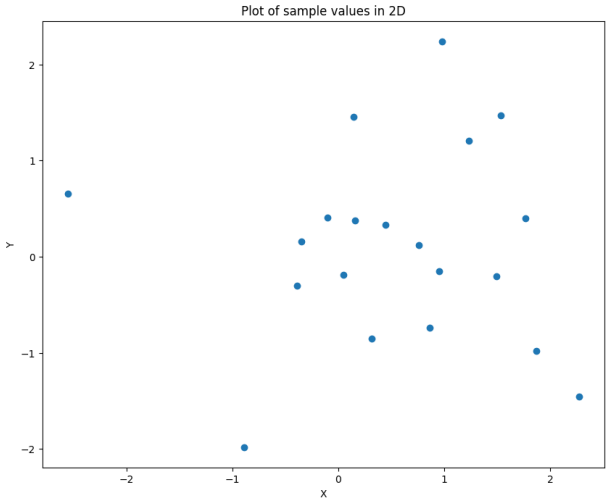

Figure 2: Plot of Random Data in Normal distribution

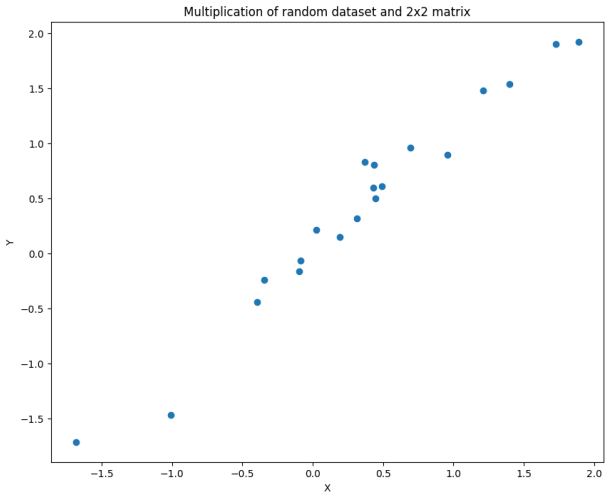

Figure 3: Plot of data multiplied with 2x2 matrix

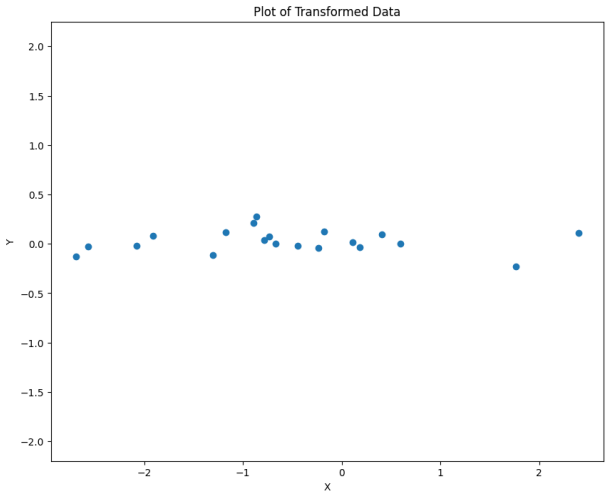

Figure 4: Plot of Principal component

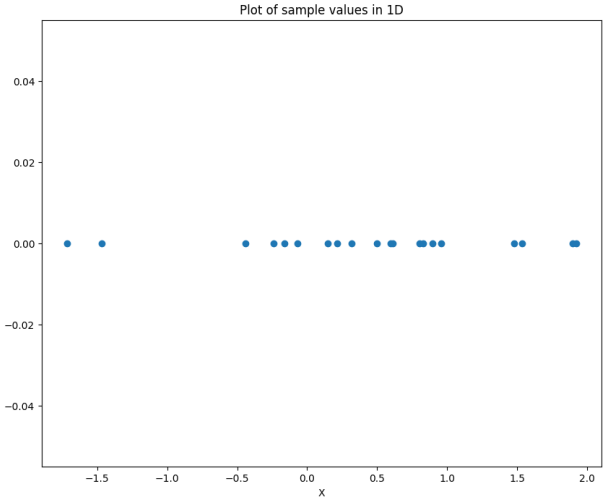

Figure 5: Plot of sample values in 1D using PCA1

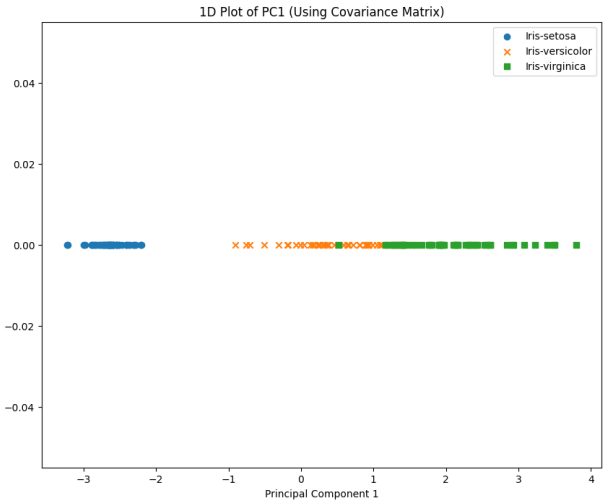

Figure 6: Plot of Principal component

Using the eigenvector having the highest eigenvalue, 2-dimension data is reduced to 1-dimension data which is visualized below:

B PROBLEM 2: PCA ON IRIS DATASET

1 1D PLOT OF DATASET

The principal component is calculated by multiplying the data with the eigenvector of the covariance matrix having the highest eigenvalue. So, there is only one principal component that reduces 4-dimensional data to 1-dimensional data.

Using the principal components, several 2D and 3D scatterplots were generated to visualize the data features.

2 2D PLOTS OF DATASET

3 3D PLOTS OF DATASET

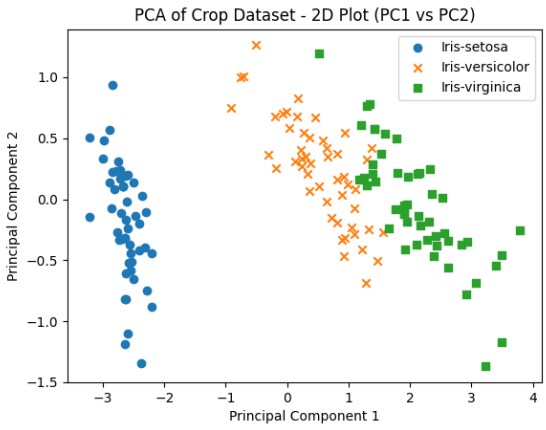

Figure 7: Plot using PC1 and PC2

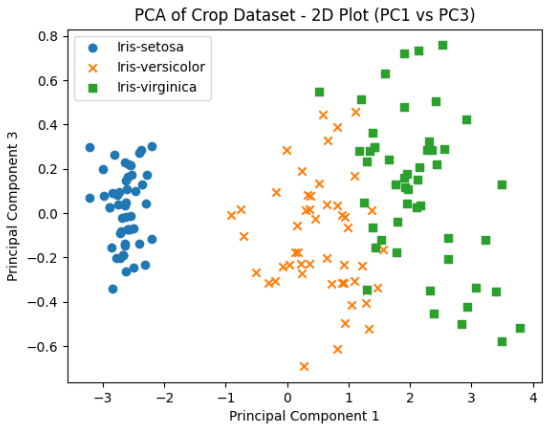

Figure 8: Plot using PC1 and PC3

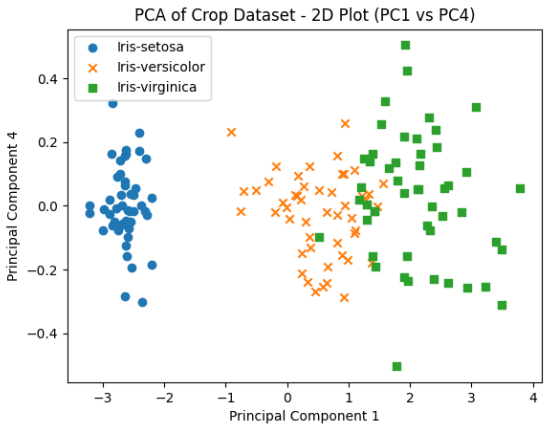

Figure 9: Plot using PC1 and PC4

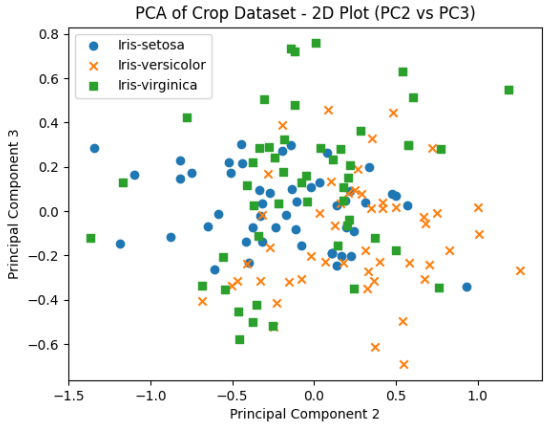

Figure 10: Plot using PC2 and PC3

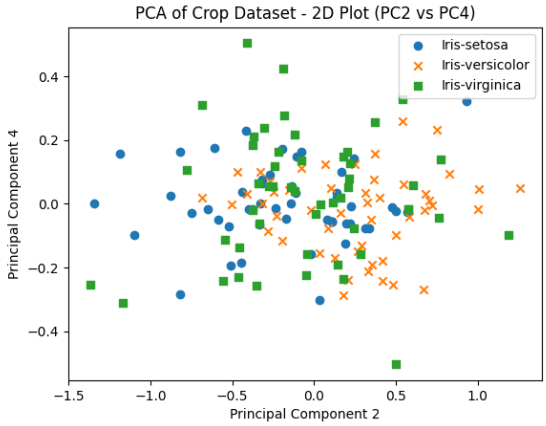

Figure 11: Plot using PC2 and PC4

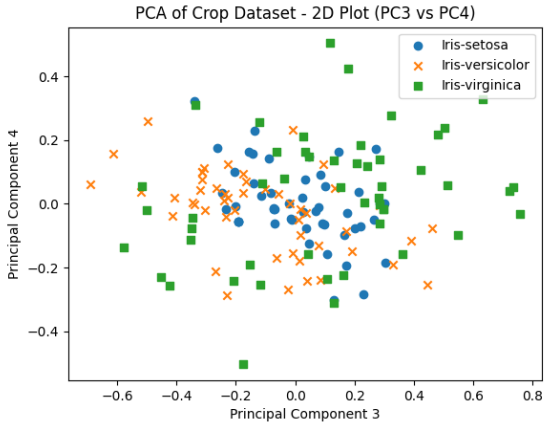

Figure 12: Plot using PC3 and PC4

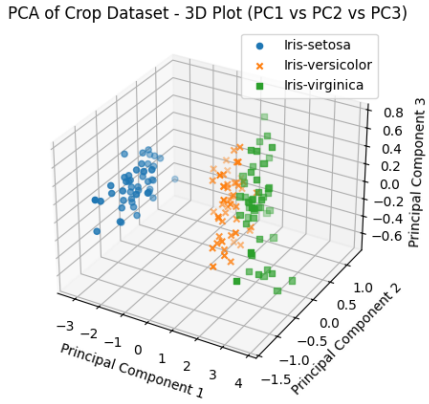

Figure 13: Plot using PC1, PC2 and PC3

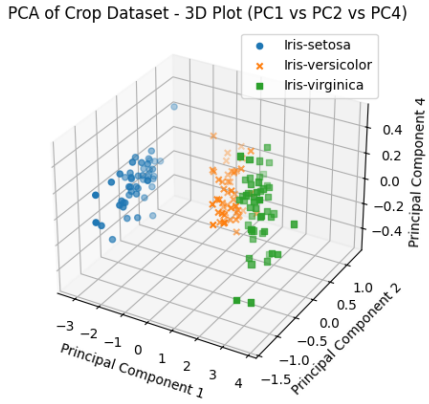

Figure 14: Plot using PC1, PC2 and PC4

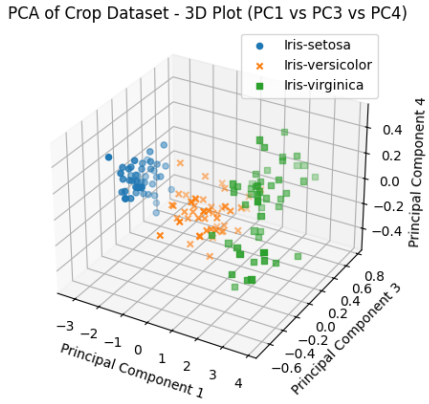

Figure 15: Plot using PC1, PC3 and PC4

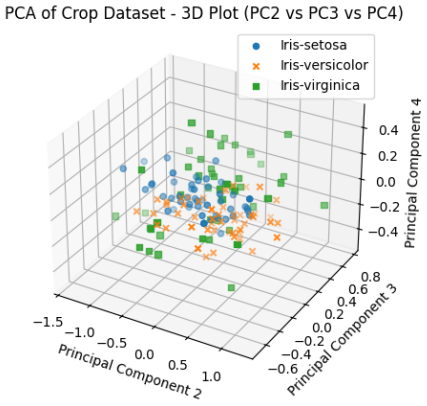

Figure 16: Plot using PC2, PC3 and PC4

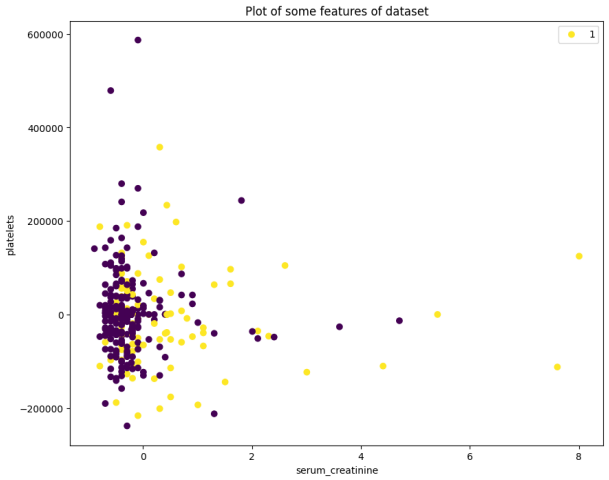

Figure 17: Plot of some features of dataset

C PROBLEM 3: PCA ON HEART FAILURE CLINICAL RECORDS

Experimentation on Heart Failure Clinical Records and implementation of PCA methods provided less-dimension data from 12-dimension data.

1 VISUALIZATION OF DATA FEATURES BEFORE AND AFTER MEAN CENTERING

The features were mean-centered. The 2D plot after mean centering is visualized for the above two features.

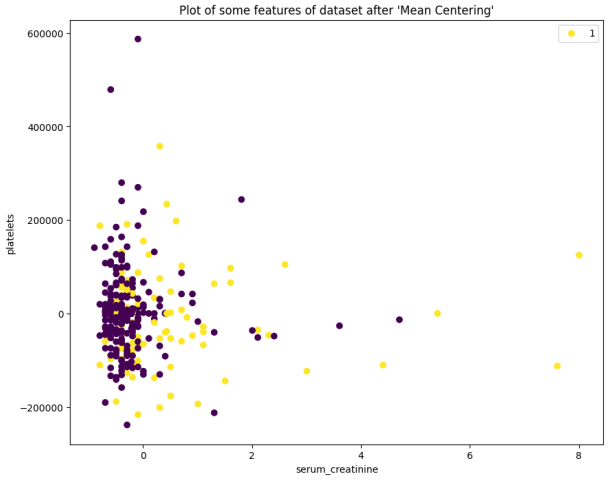

Figure 18: Plot of some features of the dataset after 'Mean Centering'

2 1D PLOT OF DATASET

The principal component is calculated by multiplying the data with the eigenvector of the covariance matrix having the highest eigenvalue. So, there is only one principal component which reduces 12-dimension data to 1-dimension data.

Using the top 4 principal components, several 2D and 3D scatterplots were generated to visualize the data features.

3 2D PLOTS OF DATASET

4 3D PLOTS OF DATASET

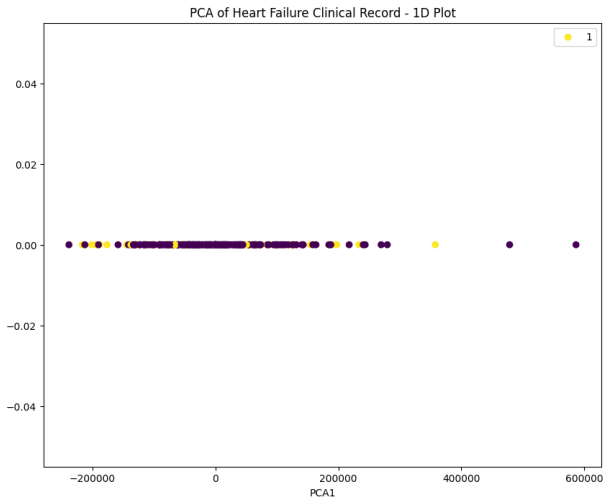

Figure 19: Plot of Principal component

V DISCUSSION AND ANALYSIS

In IV A random data points having two features were generated, and visualized using a scatterplot. After applying PCA, the data points were transformed into 1-dimensional with the principal component.

In IV B Iris dataset, there are four dimensions. By applying PCA, the data were transformed into 1-dimension, 2-dimension, and 3-dimension data using principal components. The dimensionality reduction helped in better visualization of data features. Using the PCA method the data separately. The clusters of data were visualized properly in 2-dimension and 3-dimension datasets. The plot showcased the separation and grouping of similar classes.

In IV C we reduced the 12 dimensions of the dataset to 1

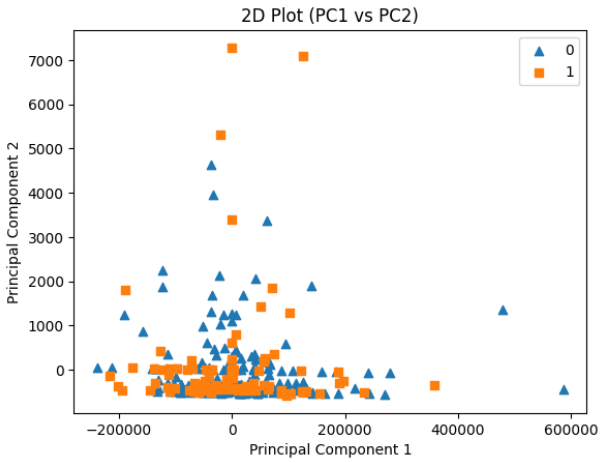

Figure 20: Plot using PC1 and PC2

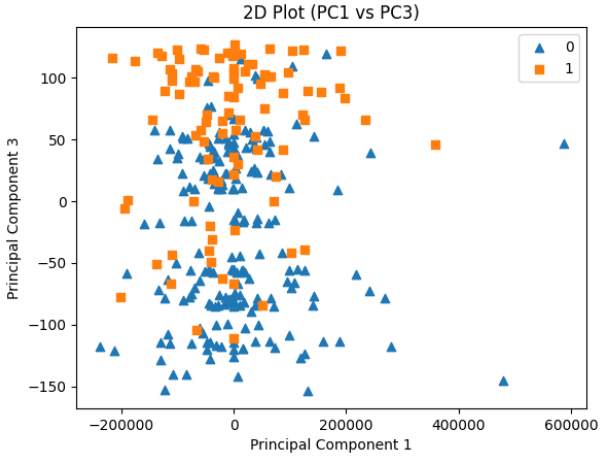

Figure 21: Plot using PC1 and PC3

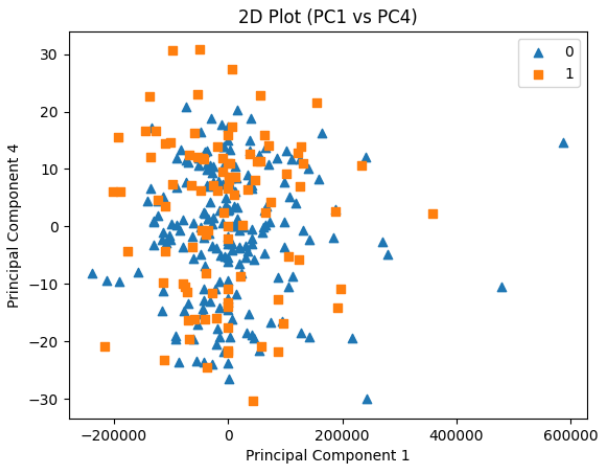

Figure 22: Plot using PC1 and PC4

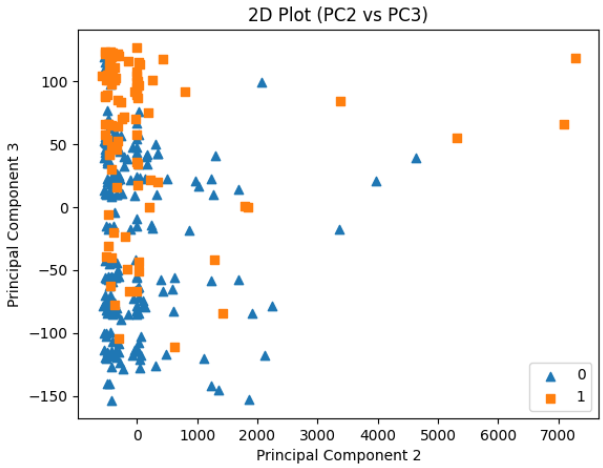

Figure 23: Plot using PC2 and PC3

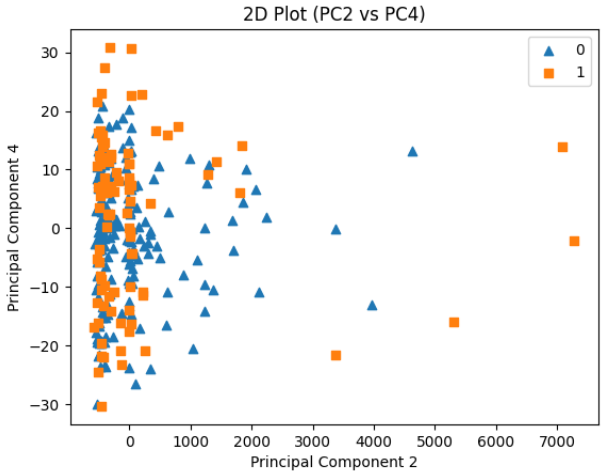

Figure 24: Plot using PC2 and PC4

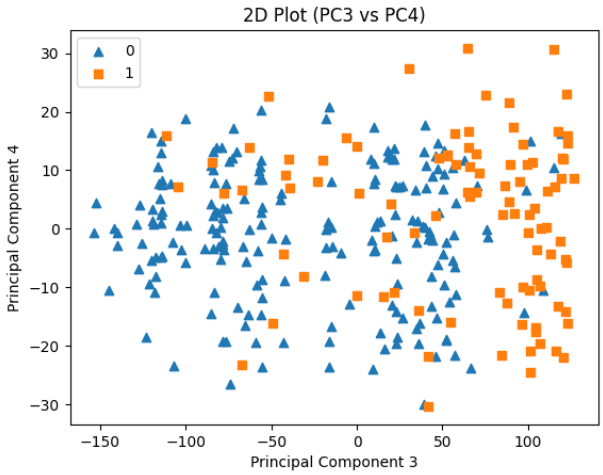

Figure 25: Plot using PC3 and PC4

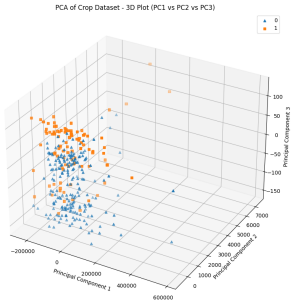

Figure 26: Plot using PC1, PC2 and PC3

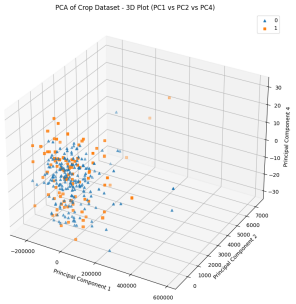

Figure 27: Plot using PC1, PC2 and PC4

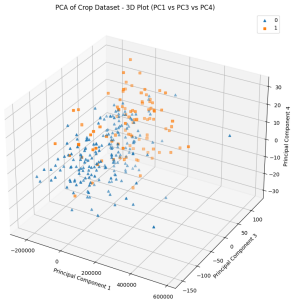

Figure 28: Plot using PC1, PC3 and PC4

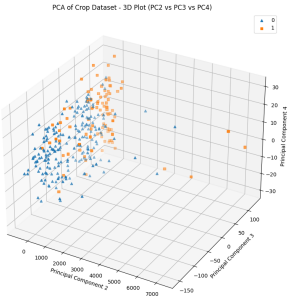

Figure 29: Plot using PC2, PC3 and PC4

dimension using the principal component. Similarly, the data were transformed into 1-dimension, 2-dimension, and 3-dimension data using principal components.

VI CONCLUSION

The results and analysis from the visualization of data prove PCA to be an effective technique for feature extraction and dimensionality reduction. Principal Component Analysis (PCA) shows great efficiency in all types of datasets. In the first problem, it managed to uncover the underlying structure of our random data, showing how it can find patterns even in random-looking information. Moving on to the second problem, PCA did a great job at condensing the iris dataset, making it easier to see the differences between the different types of iris flowers in 2D and 3D plots. This shows us how PCA can make data easier to understand and might even help us tell things apart more accurately. Lastly, in the third problem, PCA simplifies the dataset dimension to different dimensions for easy visualization.

In conclusion, PCA has proven to be a valuable support for simplifying complex data and uncovering underlying patterns. However, it's important to acknowledge its limitations, such as its assumption of data behaving in a specific manner and its sensitivity to outliers. Nevertheless, despite these constraints, PCA has demonstrated its versatility in various domains, including data analysis and pattern recognition. Moving forward, it would be beneficial to explore alternative methods for data simplification and experiment with larger datasets. Doing so could expand our knowledge and proficiency in utilizing PCA effectively in our future work.

REFERENCES

[1] L. Paul, A. Suman, and N. Sultan, "Methodological analysis of principal component analysis (pca) method," *International Journal of Computational Engineering Management*, vol. 16, no. 2, pp. 32–38, 2013.

[2] S. S. Meher and P. Maben, "Face recognition and facial expression identification using pca," in *2014 IEEE International Advance Computing Conference (IACC)*, 2014, pp. 1093–1098.

[3] A. Unwin and K. Kleinman, "The iris data set: In search of the source of virginica," *Significance*, vol. 18, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244763032

[4] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211018036

Kristina Ghimire is currently pursuing her undergraduate degree in Computer Engineering at IOE, Thapathali Campus. Her research interests encompass various areas, including data mining, machine learning, and deep learning.(THA077BCT023)

Punam Shrestha is currently pursuing her undergraduate degree in Computer Engineering at IOE, Thapathali Campus. Her research interests encompass various areas, including data mining, and web development.(THA077BCT038)