Principle Component Analysis on Random Dataset, Iris Dataset and Palmer Penguins Dataset

Rashmi Khadka
Department of Electronics and Computer Engineering
IOE, Thapathali Campus
Kathmandu, Nepal
rashmikhadka6255@gmail.com

Shiwani Shah
Department of Electronics and Computer Engineering
IOE, Thapathali Campus
Kathmandu, Nepal
shahshiwani70@gmail.com

***Abstract*—Principal Component Analysis (PCA) is a widely used statistical technique for dimensionality reduction in datasets while preserving the most significant sources of variation. In this lab report, we investigate the application of PCA in analyzing random number data generated through a normal distribution, as well as the well-known Iris Dataset and Palmer Penguins Dataset. We compare manual PCA implementations with the PCA functionality provided by the Scikit-learn library. The report explores the mathematical foundation of PCA, focusing on how it computes the eigenvectors of the covariance matrix to determine the directions of maximum data variation. These eigenvectors, or principal components, are then used to transform the data into a new coordinate system, effectively reducing its dimensionality.**

***Index Terms*—Eigenvector, Eigenvalue, PCA, Principal Components**

I. Introduction

PCA is a statistical technique used to reduce the dimensionality of a dataset while preserving the most significant sources of variation. By identifying a set of orthogonal axes called principal components, PCA enables us to understand the underlying structure of the data. It has emerged as one of the oldest and most widely used methods for dimensionality reduction. It provides a powerful framework for transforming datasets with a multitude of variables into a lower-dimensional representation while preserving the fundamental characteristics of the data. By identifying and extracting the principal components, which represent the orthogonal axes capturing the largest inherent variation in the data, PCA simplifies the analysis and visualization of the dataset.

The purpose of this lab report is to investigate and analyze the application of Principal Component Analysis (PCA) in data analysis of random number generated through normal distribution, Iris Dataset and Palmer Penguins Dataset from scratch. Also, this lab focuses on comparing the manual Pricipal Component Analysis (PCA) of datasets with the Pricipal Component Analysis (PCA) of datasets using Scikit-learn library. In this report, we explore the concept of PCA and its mathematical foundation. We discuss how PCA computes the eigenvectors of the covariance matrix to determine the directions in which the data varies the most. These eigenvectors, known as the principal components, are used to transform the data into a new coordinate system. By projecting the data onto these principal components, we can effectively reduce the dimensionality of the dataset.

II. Methodology

The methodology for performing Principal Component Analysis (PCA) on various datasets involves importing necessary libraries and defining visualization functions. There are several steps in principle component analysis through which dimensions are reduced and data is visualized.

A. Steps of PCA

To reduce the dimensions of a d-dimensional dataset and to project it onto a k-dimensional subspace where $k < d$ with high computational efficiency and retaining most of the information, the procedures are:

- **Standardization of dataset**
  In this step we standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.
  More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges (for example, a variable that ranges between 0 and 200 will dominate over a variable that ranges between 0 and 1), which will lead to biased results. So, transforming the data to comparable scales can prevent this problem.
  Let $\bar{X}$ be the mean vector (taking the mean of all rows)
  Adjust the original data X by the mean

$$X' = X - \bar{X} \tag{1}$$

- **Compute the covariance matrix A of adjusted X**
  The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information. So, to identify these correlations, we compute the covariance matrix.

$$Cov(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \tag{2}$$

- **Find the eigenvectors and eigenvalues of A**
  Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data.

  Eigenvectors of Covariance matrix A are v such that

$$A.v = \lambda.v \tag{3}$$

$$(A - \lambda).v = 0 \tag{4}$$

where λ is called an eigenvalue of A.
The eigenvectors of the Covariance matrix gives the directions of the axes (Principal Components) where there is the most variance (most information). And eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component.

Principal component
Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that the new variables are uncorrelated and most of the information within the initial variables is compressed into the first components. So, a 10-dimensional data gives 10 principal components, but PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on.

– Principal Component (PC1)

The eigenvalue with the largest absolute value will indicate that the data have the largest variance along its eigenvector, the direction along which there is greatest variation is PC1.

– Principal Component (PC2)

It is the direction with maximum variation left in data, orthogonal to the PC1.
Organizing information in principal components this way will allow to reduce dimensionality without losing much information, and this by discarding the components with low information and considering the remaining components as new variables.

- **Sort eigenvalues**
  Eigenvalues are sorted in descending order and k eigenvectors that correspond to the largest k eigenvalues are chosen.

- **Create the projection matrix**
  The projection matrix P is created from selected k eigenvectors as column vectors. Each eigenvector corresponds to a principal component.

- **Transformation of dataset**
  The original dataset X of d-dimension is transfered via P to obtain a k -dimensional subspace Y.

B. System architecture

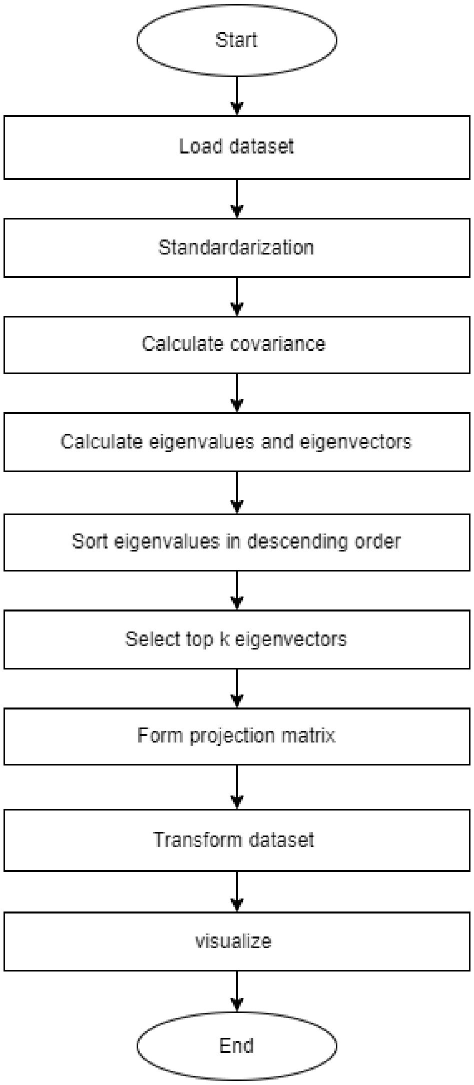

Fig. 1. System flow diagram

C. Instrumentation

The Principle Component Analysis(PCA) involved the use of several powerful Python libraries:

- **NumPy**: A fundamental library for scientific computing, providing efficient numerical operations and array manipulation capabilities.
- **Pandas**: A versatile library for data manipulation and analysis, offering convenient data structures and functions to handle datasets effectively.
- **Seaborn**: A high-level data visualization library that creates visually appealing statistical graphics to gain insights from the datasets.
- **Matplotlib** : A comprehensive plotting library used to generate various types of plots and figures, aiding in the visualization of the PCA results.
- **Scikit-learn**: A popular machine learning library used for data preprocessing and modeling. The StandardScaler class standardized the dataset, while the PCA class performed the Principal Component Analysis and extracted the principal components.

D. Analysis of PCA on various datasets

Random dataset

In this PCA analysis, we begin by initializing a pseudo-random number generator with a random seed. This ensures that the random numbers generated throughout the analysis can be reproduced. Next, we sample from a univariate ”normal” (Gaussian) distribution to create an array of 20 by 2 samples. These samples represent the original dataset that we will analyze using PCA. To gain a visual understanding of the dataset, we plot the sample values, allowing us to observe the distribution and structure of the data.

To perform the transformation, we created a 2 by 2 transformation matrix with random values from a uniform distribution between 0 and 1. This matrix determined how the original dataset would be transformed. By multiplying the sampling and transformation matrices, we obtained a new set of transformed sample values. Visualizing these transformed values, we looked for emerging patterns or clusters.

To analyze the transformed data further, we computed the variance along the data axes, providing insights into the spread and distribution. We also calculated the covariance matrix, revealing relationships between different features. From this covariance matrix, we derived eigenvalues and eigenvectors, representing the principal components of the data. These components guided a change of basis, projecting the transformed data onto a new coordinate system. By considering both eigenvectors, we determined the proportion of variance explained by each component, highlighting their importance. Plotting the data in the new basis, we observed its distribution and potential clustering patterns. We also calculated a new covariance matrix, offering additional information about the transformed dataset [**Research Article 3**].

Finally, we repeated the process using only one eigenvector to understand the impact of a single principal component on the data representation. Through these steps, we conducted a comprehensive PCA analysis, unraveling the underlying structure and gaining valuable insights into the transformed data.

Iris dataset

The Iris flower data set is a multi-variable dataset introduced by Ronald Fisher in his 1936 as an example of linear discriminant analysis .The Iris dataset is a well-known and frequently used dataset in machine learning and statistics. The dataset contains measurements of four features (sepal length, sepal width, petal length, and petal width) from 150 iris flowers, with 50 samples from each of three different species: setosa, versicolor, and virginica [**scikit-learn**].

First, we need to load the Iris dataset using the scikit-learn library. We split the dataset into data and class files, where the data-file is a 150x4 matrix, with each row representing a sample and each column representing a feature. Next, we standardize the dataset by scaling it onto the unit scale, with zero mean and unit variance. This step is important because PCA is sensitive to the scales of the input variables. Standardization allows us to give equal importance to all the features during the analysis. Once the dataset is standardized, we calculate the covariance matrix and perform Singular Value Decomposition (SVD) to determine the eigenvalues and eigenvectors. The eigenvalues represent the amount of variance explained by each principal component, while the corresponding eigenvectors represent the direction of the principal components in the original feature space. In the third step, we select the principal components by ranking the eigenvalues in descending order. By choosing the top K eigenvectors, we can retain the most informative components.

In the last step, we project the original data onto the new feature space defined by the selected principal components. The projection matrix is constructed by concatenating the top K eigenvectors. In the case of the Iris dataset, the original 4-dimensional feature space is reduced to a 2-dimensional subspace by selecting the top two eigenvectors with the highest eigenvalues. This projection transforms the Iris data from a 150x4 matrix into a new 150x2 matrix, where each sample is represented by its corresponding values in the new feature space which can be visualized through scatter plots.

Palmer Penguin dataset

It contains measurements of “344” Penguins dataset from “3” different species, Adelie 152, Gentoo 124, Chinstrap 68. The seven features are species, island, culmen_length_mm, culmen_depth_mm, flipper_length_mm, body_mass_g, sex. Species, island and sex are string values while others are float [**Palmer-Penguin-dataset**]. The approach starts with loading the dataset and splitting it into data and class files. The data-file is 344*7 in which the columns represent the different features, and each row represents a specific Penguin species. Here we drop the 3 features column that are in string and take species as target. The data has some null values in features, so we drop them also. Now it is 342*4 dimensional.

For optimal response, the second step of the algorithm is standardizing the dataset onto the unit scale (zero-mean and unity-variance). The algorithm determines the eigenvalues and the corresponding eigenvectors from the covariance matrix.

In the third step, the approach selected the principal components by ranking the eigenvalues in descending order and selecting the top 3 eigenvectors (as these contain most of the information).

After having the principal components, to compute the percentage of variance (information) accounted for by each component, we divide the eigenvalue of each component by the sum of eigenvalues. For this dataset, we find that PC1, PC2, PC3 carry respectively about 68 percent, 19 percent, 9 percent and so on of the variance of the data. The calculated variance from the eigenvalues informs how much information is attributed to each principal component.

The last step in the approach is projecting the original data onto the new feature space. The four-feature space of data is reduced to a 3-dimensional feature subspace by selecting the top three eigenvectors with the highest eigenvalues. Thus, we use the 4*3 projection matrix (selected eigenvector) to transform the data into a new 342*3 matrix.

III. Result

A. Random plot

The dataset comprises 20 randomly sampled data points from a Gaussian distribution, which are scattered in a random pattern across the plot depicted in Fig. 2. These samples provide an appropriate basis for applying Principal Component Analysis (PCA) to extract valuable insights.

To normalize the data, which is initially distributed in a Gaussian pattern, a 2D matrix transformation is applied. This transformation aligns the random samples into a normal distribution, as illustrated in Fig. 3.

Subsequently, by computing the eigenvalues and eigenvectors, the principal component (PC1)and (PC2) are derived and visualized in Fig. 4. After that data is visualized taking ony one principle component as shown in Fig. 5. In this representation, all data points are aligned in a single direction, reflecting the choice of a single principal component to transform the original samples.

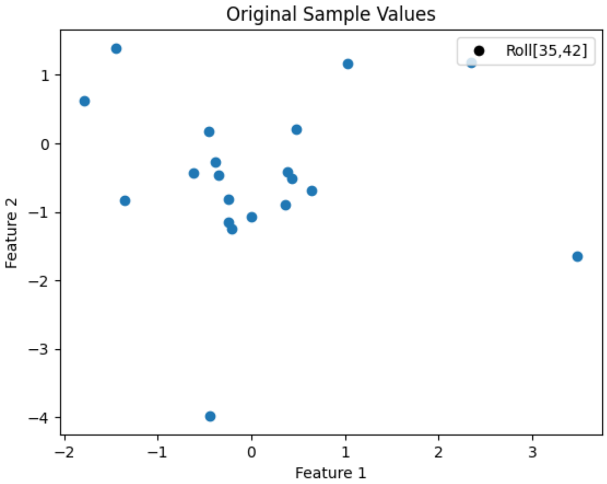

Fig. 2. Plot of random numbers generated

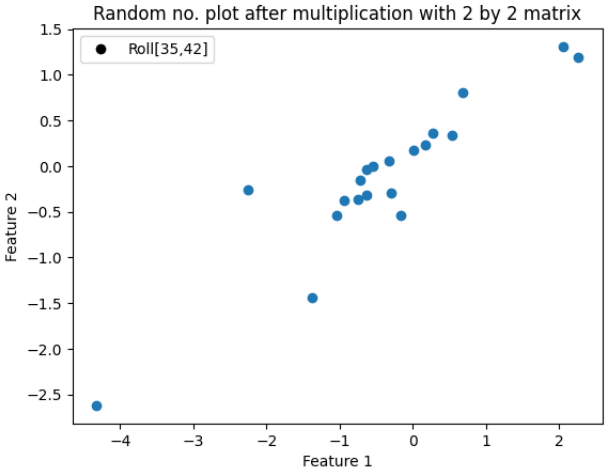

Fig. 3. Random number plot after 2D transformation

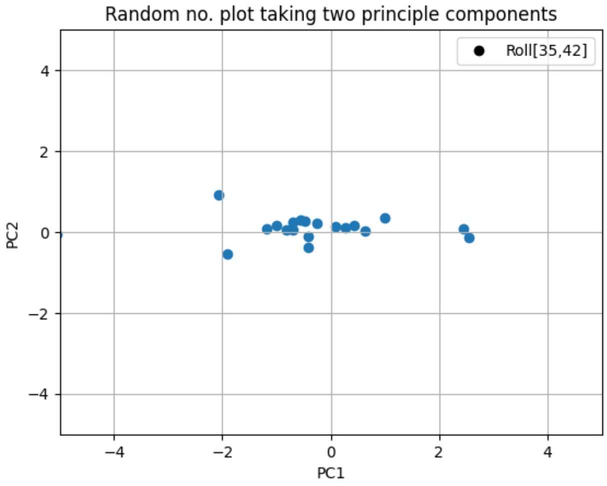

Fig. 4. Random number plot with two principle components

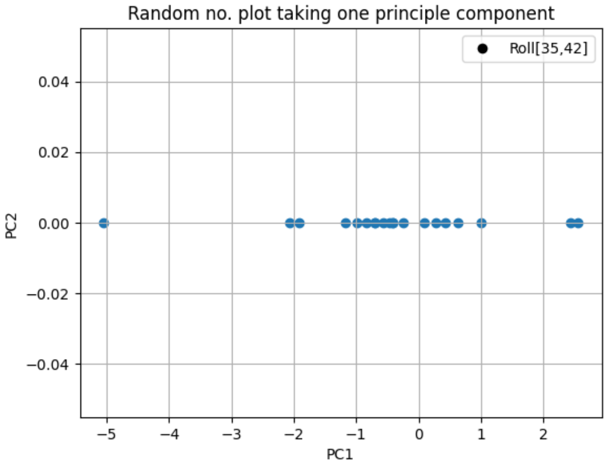

Fig. 5. Random number plot with one principle component

B. *Iris Dataset plot*

The Iris Dataset, obtained from the scikit-learn library, is composed of four features: Sepal Length, Sepal Width, Petal Length, and Petal Width. These features are associated with three target classes: Setosa, Versico, and Virginica. To gain a deeper understanding of the dataset and reduce its dimensionality, Principal Component Analysis (PCA) was applied. The results of PCA are visualized in Fig. 6, showcasing the transformed dataset.

In PCA, the selection of eigenvectors is crucial. In this analysis, the top two eigenvectors with the highest eigenvalues were chosen as the principal components. These components can be observed in Fig. 7. Additionally, the top three principal components were selected and visualized in a three-dimensional plot, as shown in Fig. 9. The application of PCA using the Scikit-learn library is also presented, with the dataset visualized in 2D (Fig. 8) and 3D (Fig. 10). These visualizations provide valuable insights into the dataset's structure and patterns.

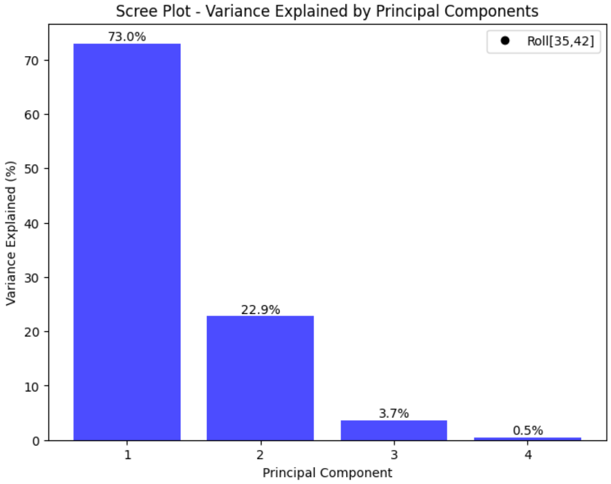

Fig. 6. plot of proportion of variance of principal component

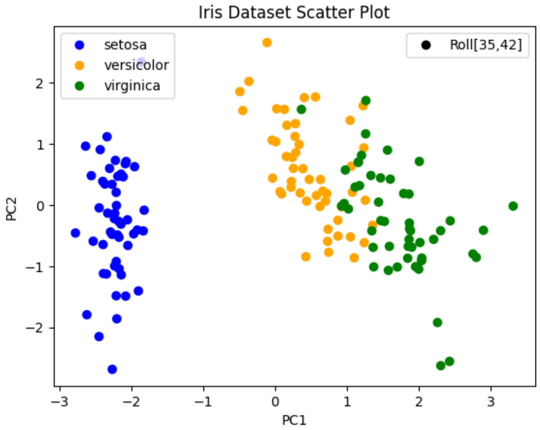

Fig. 7. PCA on Iris dataset in 2D

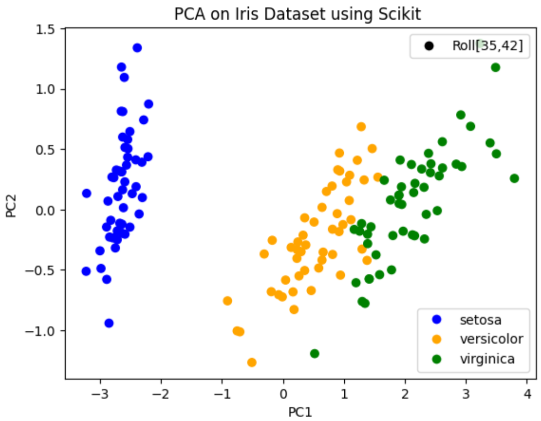

Fig. 8. PCA on Iris dataset in 2D using Scikit

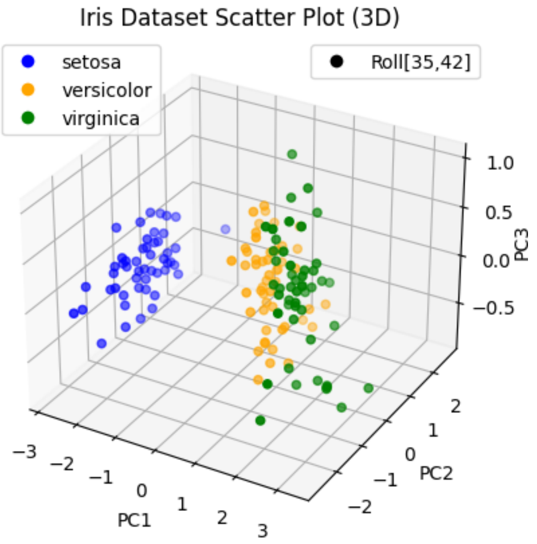

Fig. 9. PCA on Iris dataset in 3D

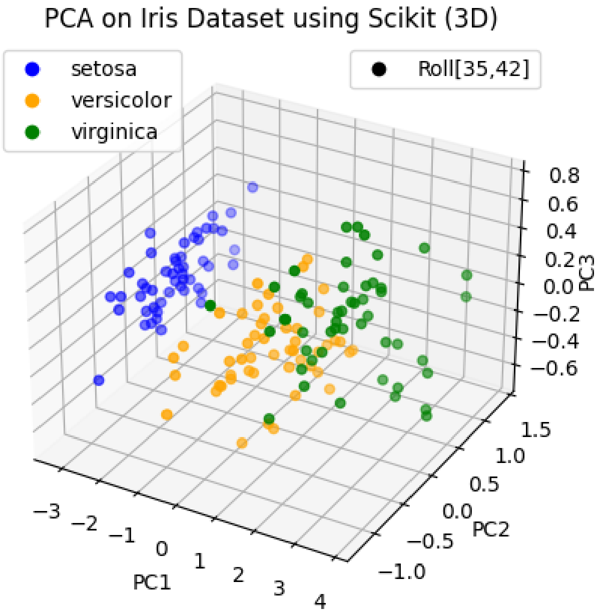

Fig. 10. PCA on Iris dataset in 3D using Scikit

C. Palmer Penguins Dataset plot

The Palmer Penguin Dataset consists of several features such as bill length, bill depth, flipper length, and body mass, associated with three penguin species: Adelie, Gentoo, and Chinstrap. To explore the dataset and uncover underlying patterns while reducing its dimensionality, Principal Component Analysis (PCA) was applied. The transformed dataset after PCA is visualized in Fig. 11, illustrating the reduced dimensionality.

In PCA, the selection of eigenvectors plays a vital role. In this analysis, the principal components were determined by choosing the top two eigenvectors with the highest eigenvalues, as depicted in Fig. 12. Moreover, the top three principal components were selected and visualized in a three-dimensional plot, as presented in Fig. **??**. Additionally, the application of PCA using the Scikit-learn library was utilized to analyze the Palmer Penguin Dataset. The dataset was visualized in 2D (Fig. 13) and 3D (Fig.15) using the Scikit-learn library. These visualizations offer valuable insights into the structure and patterns present within the dataset.

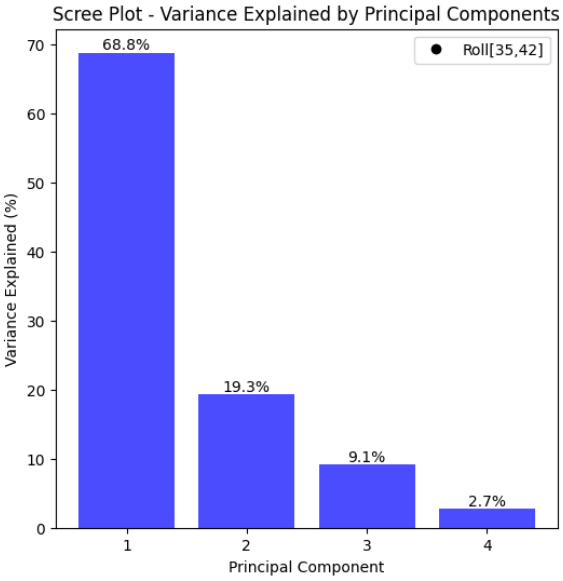

Fig. 11. plot of proportion of variance of principal component

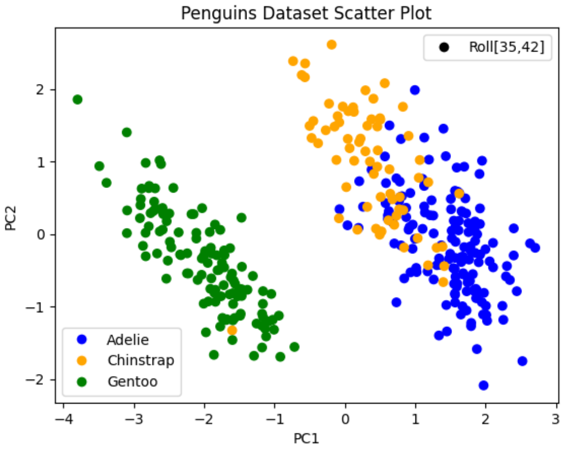

Fig. 12. PCA on penguins dataset in 2D

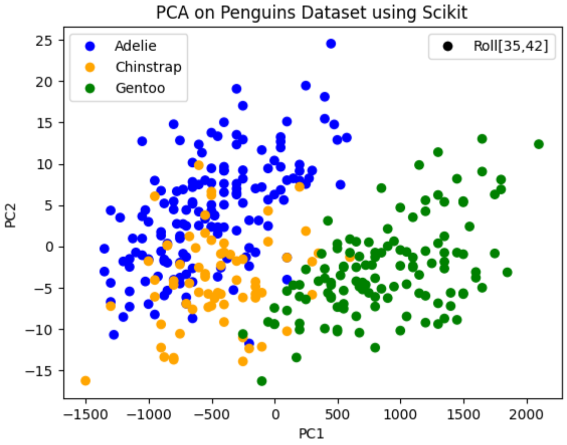

Fig. 13. PCA on penguins dataset in 2D using Scikit

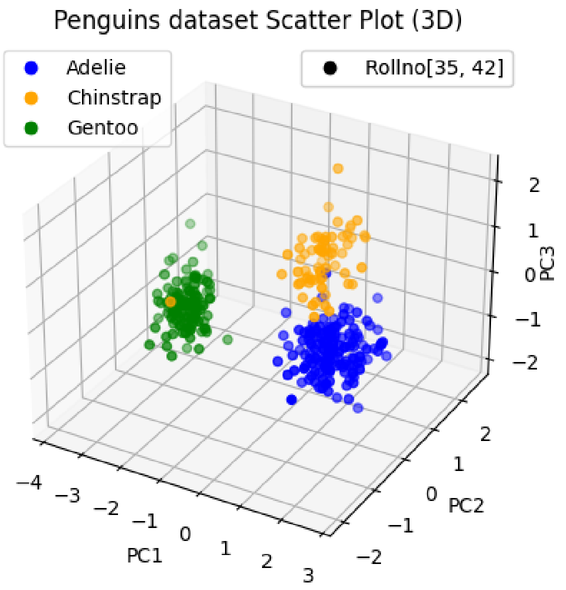

Fig. 14. PCA on penguins dataset in 3D

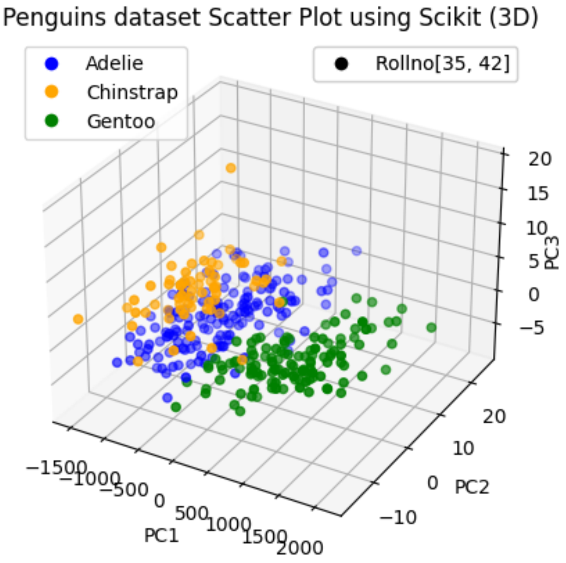

Fig. 15. PCA on penguins dataset in 3D using Scikit

IV. Discussion and Analysis

A. Random Plot

The analysis begins with a random dataset consisting of 20 numbers generated from a gaussian distribution. They are spread randomly across all directions so PCA cannot be applied to it. They undergo a transformation using a randomly selected 2D matrix from a uniform distribution. The transformation resulted in the data points is aligned in a specific direction. After applying PCA, all the data points are now aligned in one horizontal line as only one principal component was selected to transform the data.

B. Iris and Penguin dataset

Both data consist of multi-variable and multi featured dataset. By dropping irrelevant columns, handling null values, and standardizing the data, the dataset is prepared for analysis. The top three principal components are chosen based on their eigenvalues, which capture the majority of the information in the dataset. The variance analysis reveals the percentage of variance explained by each principal component, indicating their relative importance. The transformation of the dataset using the projection matrix results in a reduced-dimensional feature space, enabling easier visualization of datasets in 2D and 3D.

V. Conclusion

Principal Component Analysis (PCA) is a powerful tool for dimensionality reduction and exploring the underlying structure of complex datasets. By identifying and extracting the principal components, PCA enables us to reduce the dimensionality while retaining the essential information. In this report, we applied PCA to three datasets: random numbers generated through a normal distribution, the Iris Dataset, and the Palmer Penguin Dataset.

Through PCA, we successfully transformed the random number dataset, gaining insights into its variation and patterns.Through the extraction of principal components, we achieved a lower-dimensional representation of the data while retaining its essential characteristics. Additionally, PCA on the Iris Dataset allowed us to visualize the data in a lower-dimensional space and understand the relationships between the features and target classes. The Palmer Penguin Dataset analysis demonstrated how PCA can uncover hidden patterns and species classifications, facilitating a deeper understanding of the dataset.Comparing manual PCA implementation with the Scikit-learn library demonstrated the effectiveness and convenience of using established tools for PCA analysis.

VI. References

[1] Author, Ledisi G. Kabari, Believe B. Nwamae, *Principal Component Analysis (PCA) - An Effective Tool in Machine Learning*, pp.4, 2019.

[2] Author, Nema Salem, Sahar Hussien, *Data dimensional reduction and principal components analysis*, pp.8, 2019.

[3] Author, Jonathon Shlens, *A Tutorial on Principal Component Analysis*, pp.1, 2014

[4] scikit-learn, [Online]. Available: https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html.

[5] Palmer-Penguin-dataset,[Online]. Available: https://www.kaggle.com/datasets/parulpandey/palmer-archipelago-antarctica-penguin-data.

Rashmi Khadka is a driven individual currently pursuing a Bachelor's degree in Computer Engineering at Thapathali Campus. She possesses a strong passion for machine learning and data mining, consistently seeking to delve into the latest advancements in these domains. While Rashmi may not have amassed notable achievements at this point, her enthusiasm and determination to learn and apply state-of-the-art technologies make her a promising and ambitious figure in the field of computer engineering.

Shiwani Shah is a dedicated individual currently studying Bachelor's in Computer Engineering at Thapathali Campus. With a strong passion for research and innovation, Shiwani actively engages in various projects and practical applications to apply his knowledge. Her continuous quest for learning drives her to stay updated with the latest advancements and trends in the field. Equipped with expertise in machine learning and data science, Shiwani's relentless determination, inquisitive mindset, and commitment to technology position her to make significant contributions to the ever-evolving landscape of the industry.