**CSE-564 - Visualization**

**Lab 1 – Assignment: Data Visualization with PCA, K-Means Clustering & Scatterplot Analysis on Kidney Disease Dataset**

**PCA, K-Means, and Scatterplot Visualization Dashboard**

**Introduction**

This project aims to visualize the principal component analysis (PCA) and K-means clustering on the dataset of chronic kidney disease (CKD). It provides an interactive dashboard to explore the complexities of high-dimensional CKD data by projecting it into a two-dimensional space. This allows for easy identification of patterns and insights that could potentially assist in the medical diagnostics process.

**Source of the Dataset**

The dataset used in this project is obtained from Kaggle, which is a platform for predictive modeling and analytics competitions. The specific dataset for chronic kidney disease can be accessed through the following link:[**CK Disease Dataset**](https://www.kaggle.com/datasets/mansoordaku/ckdisease) on Kaggle. It comprises various medical attributes that are essential in diagnosing CKD in patients.

**Attributes Description**

The dataset contains multiple attributes, including both numerical and categorical data, such as age, blood pressure, specific gravity, albumin, sugar, red blood cells, and more. Each attribute provides critical information regarding the patient's health status, contributing to the diagnosis of CKD.

**Interesting Insights in the Data**

Upon analysis of the dataset, several interesting insights were revealed, such as correlations between certain attributes and the presence of CKD. For instance, higher levels of albumin in the urine (proteinuria) showed a significant association with the disease. Visualization of data clusters helped identify groups of patients with similar symptoms and conditions.

The PCA and K-Means visualizations reveal several interesting insights:

* The scree plot suggests a point of diminishing returns after a certain number of components, indicating the number of principal components that capture the most variance.
* The biplot highlights the loading of each variable and how they contribute to the components, providing an intuitive understanding of which variables have the most significant effect.
* The scatter plot matrix allows for the visualization of relationships between pairs of attributes, showing how certain variables correlate with each other.
* The K-Means MSE plot demonstrates the process of identifying the optimal number of clusters by locating the elbow point, which signifies where adding more clusters does not significantly improve the model.

**Noteworthy Implementation Aspects**

The dashboard utilizes D3.js for dynamic data visualization, Flask as the web server framework, and Python for data processing and PCA computation. Special attention was given to the scalability and responsiveness of the visualizations, ensuring a seamless user experience across different devices.

The implementation of this PCA, K-Means & Scatter Plot Visualization Dashboard represents a multifaceted approach to data analysis, showcasing a blend of analytical rigor and interactive visualization. Some of the noteworthy aspects include:

* **Adaptive Visualization:** The dashboard provides a user-interactive environment where selections made in one plot dynamically influence the display of others. For instance, selecting the number of clusters in the K-Means MSE plot directly updates the scatter plot matrix to reflect this choice. This adaptability not only enhances user engagement but also facilitates a deeper exploration of the dataset's structure.
* **Automated Elbow Detection:** In the K-Means MSE plot, an algorithmic approach is implemented to automatically detect the elbow point—a key step in determining the optimal number of clusters for K-Means analysis. The use of the KneeLocator function from the kneed library streamlines this process, providing a more objective and reproducible method for identifying the point of diminishing returns in variance reduction.
* **Biplot with Contextual Insights:** The PCA biplot is particularly noteworthy for its capacity to concurrently display the original data points and the principal component vectors (loadings). This overlay provides immediate visual feedback on how each variable influences the principal components, offering intuitive insights into the relative importance of each attribute in the dataset.
* **Responsive Design:** The dashboard is responsive, ensuring that the visualization remains clear and coherent across different devices and screen sizes. The media queries and flexible grid layouts help maintain the usability of the dashboard, making the analysis accessible to users regardless of their device.
* **Advanced Data Preprocessing:** The implementation includes a robust preprocessing pipeline that standardizes the data, ensuring that the PCA is not biased by the varying scales of the original attributes. This step is crucial for meaningful dimensionality reduction and is often overlooked in less rigorous analyses.
* **Intuitive User Interface:** The interface design is clean and minimalistic, directing focus to the visualizations themselves. It avoids clutter and unnecessary distractions, providing an intuitive user experience that caters to both novice and expert users.
* **Technological Integration:** The project seamlessly integrates various technologies such as Flask for web server management, D3.js for creating dynamic visualizations, and Pandas for data manipulation. This harmonious integration showcases how different libraries and frameworks can be combined to create a sophisticated data analysis tool.

**Conclusion**

The PCA, K-Means & Scatter Plot Visualization Dashboard stands as a testament to the power of combining data visualization with interactive web technologies. The successful implementation of this dashboard has resulted in a tool that is not only analytically robust but also highly intuitive and user-friendly.

By leveraging the automatic detection of the optimal number of clusters and enabling dynamic interaction with the PCA biplot, the dashboard empowers users to gain deeper insights into their data. The automated features minimize the need for manual intervention, making advanced data analysis more accessible to a wider audience.

Furthermore, the responsive design ensures that these insights are accessible across various devices, highlighting the importance of adaptability in modern data analysis tools. The project's modular design and the integration of multiple technologies pave the way for future enhancements and features.

**Running My Code**

To launch the dashboard, navigate to the project directory and run the following command in the terminal:

**python app.py**

The Flask server will start, and the dashboard will be accessible on http://localhost:5000. Ensure that all dependencies are installed, and the data file is placed in the correct directory as shown in the project structure.

**Snapshots of Dashboard UI:**

Integrated PCA, K-Means, and Scatterplot Dashboard showcasing a Scree Plot and Biplot Analysis of Multidimensional Data

A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated