**Performance Assessment**

OFM4 — OFM4 Task 2: Dimensionality Reduction Methods

Bader Ale

Department of Information Technology, Western Governors University

D212 Data Mining II

2024

Table of Contents

[Part I: Research Question 3](#_Toc183855699)

[Part II: Method Justification 3](#_Toc183855700)

[Part III: Data Preparation 4](#_Toc183855701)

[Part IV: Analysis 6](#_Toc183855702)

[Determining the Principal Component Matrix 6](#_Toc183855703)

[Identifying the Total Number of Principal Components 7](#_Toc183855704)

# Part I: Research Question

The purpose of this report is to explore the following research question: *"How can Principal Component Analysis (PCA) help identify key patterns and simplify a hospital's medical dataset to improve the understanding of patient characteristics and resource utilization?"* By using Principal Component Analysis (PCA), the goal is to identify the most important variables that influence differences in readmission rates among patients. Understanding these key factors can help hospital staff make informed decisions to improve patient care and lower readmission rates. This question is relevant to real-world healthcare settings, where efficient resource management and focused interventions are essential for better patient outcomes and smoother hospital operations.

The goal of this analysis is to simplify the hospital's medical dataset while retaining the most important information. Specifically, PCA will be used to reduce the original set of variables to a smaller set of principal components that capture the majority of the variance in the data. The aim is to identify the components that have the most influence on patient readmissions, helping healthcare providers focus on these factors for targeted interventions. This approach helps make the dataset more manageable, allowing for more effective analysis and actionable insights to reduce readmission rates.

# Part II: Method Justification

Principal Component Analysis (PCA) systematically transforms a dataset by identifying the directions of maximum variability, known as principal components. The process begins with standardizing the data to ensure variables measured on different scales contribute equally to the analysis. Next, the covariance matrix is calculated to capture relationships between variables, followed by eigenvalue and eigenvector decomposition of the matrix. Eigenvectors define the directions of the principal components, while eigenvalues quantify the amount of variance each component captures. The principal components are then ranked based on their eigenvalues, with those contributing the most variance selected for further analysis. By projecting the original data onto this reduced set of principal components, PCA simplifies the dataset while retaining its most significant information. This dimensionality reduction facilitates deeper insights into complex datasets, allowing researchers to uncover underlying patterns and relationships in a more interpretable form (scikit-learn developers , n.d.).

Standardizing the data before applying PCA is particularly crucial in this context because the hospital dataset contains variables measured in different units. For example, 'Age' is measured in years, 'TotalCharge' is in monetary units, and 'Vitamin D levels' have their own scale. These differences in scales can lead to disproportionate influences on the principal components if not addressed. Without standardization, features with larger scales would dominate the PCA, resulting in skewed or misleading insights. By using StandardScaler to give each feature a mean of 0 and a standard deviation of 1, we ensure that all variables contribute equally to the PCA, allowing us to accurately capture the underlying variance patterns in the data (scikit-learn developers , n.d.). The expected outcome of using PCA is a smaller number of components that still explain the majority of variance in the original dataset. This helps identify key factors driving patient readmission, making it easier for hospital administrators to prioritize areas for improvement and allocate resources effectively (Fonseca, 2023).

One fundamental assumption of Principal Component Analysis (PCA) is the linearity of the relationships among variables. PCA operates under the premise that the dataset's underlying structure can be effectively summarized by linear combinations of the original variables. This assumption is integral to the method, as PCA identifies directions (principal components) that maximize the variance in the data within this linear framework. If the relationships among variables are predominantly non-linear, the algorithm may fail to capture the true structure of the data, potentially resulting in the loss of critical information and an inaccurate representation of the dataset (scikit-learn developers , n.d.).

Additionally, the Kaiser criterion was used to determine the number of principal components to retain, selecting components with eigenvalues greater than one. By doing this, PCA reduces the dimensionality of the data while retaining as much information as possible. This process allows us to simplify complex datasets and focus on the components that contribute most significantly to differences in patient readmission rates (Displayr, 2023).

# Part III: Data Preparation

The continuous variables selected for this analysis are:

* Age: Patient's age. Age is a critical determinant of health outcomes and readmission likelihood, as older patients may have more complex health needs.
* Income: Income level of the patient. Financial status can impact a patient's access to healthcare resources, medications, and follow-up care, influencing readmission rates.
* VitD\_levels: Vitamin D levels measured in the patient. Vitamin D is known to play a role in immune function and bone health, potentially affecting recovery and readmission.
* Doc\_visits: Number of doctor visits made by the patient. Frequent doctor visits may indicate underlying health issues that could lead to higher readmission risk.
* Full\_meals\_eaten: Number of full meals eaten daily by the patient. Nutritional intake is crucial for recovery, and inadequate nutrition can lead to complications and readmission.
* vitD\_supp: Use of vitamin D supplements by the patient. Supplementation may indicate a proactive approach to health, potentially affecting readmission outcomes.
* Initial\_days: Length of the patient's initial hospital admission. The duration of the initial stay can reflect the severity of the condition, influencing the probability of readmission.
* TotalCharge: Total cost incurred during the patient's initial hospital stay. Higher costs may be associated with more intensive treatments, which could correlate with readmission likelihood.
* Additional\_charges: Additional costs incurred during the patient's stay. These costs could reflect complications or additional care required, providing insight into readmission risks

These variables are crucial for understanding patient characteristics, resource usage, and potential financial impacts, which may relate to readmission.

The continuous variables were standardized to ensure that each contributes equally to the analysis. Standardization was performed using the StandardScaler from the scikit-learn library, which scales the data so that each variable has a mean of 0 and a standard deviation of 1. This is a necessary step because PCA is sensitive to the scale of input features (scikit-learn developers , n.d.).

1. # Dataframe before scaling

2. df1.head()

3.

A screenshot of a menu

Description automatically generated

1. # Using Standard Scaler to scale the dataframe df1

2.

3. scaler = StandardScaler()

4. df1\_scaled = scaler.fit\_transform(df1) # First scaling data

5. df1\_scaled = pd.DataFrame(df1\_scaled, columns=df1\_columns) # Converting scaled data to dataframe

6.

7. #Svaing scaled dataframe 'df1' to CSV

8. df1\_scaled.to\_csv('Medical Clean-Task2', index=False)

9.

1. # Dataframe after Scaling

2. df1\_scaled.head()

3.

A screenshot of a graph

Description automatically generated

# Part IV: Analysis

## Determining the Principal Component Matrix

PCA was performed on the standardized dataset using scikit-learn’s PCA module. The analysis transformed the original data into a matrix of principal components, which represent the directions of maximum variance in the dataset. This transformation provides a new set of variables (principal components) that are linear combinations of the original features (scikit-learn developers , n.d.).

1. # Performing PCA

2. pca = PCA()  # You can change the number of components as needed

3. principal\_components = pca.fit\_transform(df1\_scaled)

4.

5. # Creating a DataFrame with the principal components

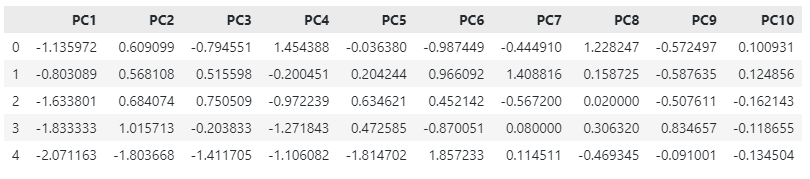
6. df1\_pca = pd.DataFrame(data=principal\_components, columns=['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7', 'PC8', 'PC9', 'PC10'])

7.

8. # Displaying the first few rows of the PCA result

9. df1\_pca.head()

10.



The loading matrix was calculated by using the *pca.components* attribute of the PCA module. The code is shown below:

1. loading\_matrix = pd.DataFrame(pca.components\_, columns=df1\_scaled.columns, index=df1\_pca.columns)

2.

A table of numbers and symbols

Description automatically generated

## Identifying the Total Number of Principal Components

The number of principal components was determined using the Kaiser criterion, which retains components with eigenvalues greater than one. The retained components capture most of the dataset's variance

1. eigenvalues = pca.explained\_variance\_

2. num\_components = sum(eigenvalues > 1)

3. print("Number of components (Kaiser Criterion):", num\_components)

4.



1. # Create a DataFrame for components

2. components\_df = pd.DataFrame({

3.     'Principal Component': [f'PC{i+1}' for i in range(len(pca.explained\_variance\_))],

4.     'Eigenvalue': pca.explained\_variance\_,

5.     'Explained Variance Ratio (%)': pca.explained\_variance\_ratio\_ \* 100

6. })

7.

8. # Filter components based on Kaiser Criterion

9. retained\_components\_df = components\_df[components\_df['Eigenvalue'] > 1]

10.

11. print("Retained Components (Kaiser Criterion):")

12. print(retained\_components\_df)

13.

A white text with black text

Description automatically generated

A graph with a line graph

Description automatically generated

The variance explained by the principal components is as follows: The variance of each of the retained principal components is as follows:

* PC1: 1.9946
* PC2: 1.7148
* PC3: 1.0413
* PC4: 1.0183
* PC5: 1.0043
* PC6: 0.9935
* PC7: 0.9749
* PC8: 0.9684
* PC9: 0.2836
* PC10: 0.0117

The variance values show the importance of each principal component, with components having an eigenvalue greater than one contributing most significantly to the dataset's variance. The Kaiser Criteria dictates that PCs with values greater than 1 be retained and as such only PC 1 through PC 5 were retained.

1. # Create a DataFrame for components

2. components\_df = pd.DataFrame({

3.     'Principal Component': [f'PC{i+1}' for i in range(len(pca.explained\_variance\_))],

4.     'Eigenvalue': pca.explained\_variance\_,

5.     'Explained Variance Ratio (%)': pca.explained\_variance\_ratio\_ \* 100

6. })

7.

8. # Filter components based on Kaiser Criterion

9. retained\_components\_df = components\_df[components\_df['Eigenvalue'] > 1]

10.

11. print("Retained Components (Kaiser Criterion):")

12. print(retained\_components\_df)

13.

A white screen with black text

Description automatically generated

## Summary

The PCA analysis revealed that five principal components were retained based on the Kaiser criterion, which selects components with eigenvalues greater than one. The retained components (PC1 to PC5) capture most of the dataset's variance, accounting for a cumulative variance of 6.7733, which is approximately 67.73%. By focusing on these five components, healthcare providers can better understand the key factors influencing patient readmission outcomes, which could help in improving resource allocation and targeted interventions. The remaining components were not retained as their eigenvalues were below one, indicating they contribute minimally to the dataset's total variance and would add complexity without significant informational gain. The remaining components were not retained as their eigenvalues were below one, indicating they contribute minimally to the dataset's total variance and would add complexity without significant informational gain.

Works Cited

Displayr. (2023, July 24). *Kaiser Rule*. Retrieved November 2024, from Kaiser Rule: https://docs.displayr.com/wiki/Kaiser\_Rule

Fonseca, M. (2023, October 19). *editage insights*. Retrieved November 2024, from An introduction to Principal Components Analysis for biomedical researchers: https://www.editage.com/insights/an-introduction-to-principal-components-analysis-for-biomedical-researchers

scikit-learn developers . (n.d.). *sklearn.decomposition*. Retrieved from PCA: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html