**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](#_Toc183704376)

[Part II: Method Justification 3](#_Toc183704377)

[Part III: Data Preparation 4](#_Toc183704378)

[Part IV: Analysis 5](#_Toc183704379)

[Determining the Principal Component Matrix 5](#_Toc183704380)

# Part I: Research Question

The purpose of this report is to explore the following research question: *"What factors in a hospital's medical dataset explain the biggest differences in patient readmission outcomes?"* 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

PCA analyzes the dataset by transforming the original features into a new set of uncorrelated variables, called principal components, which capture the maximum variance in the data. 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.

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).

# Part III: Data Preparation

The continuous variables selected for this analysis are:

* Children: Number of children a patient has.
* Age: Patient's age.
* Income: Income level of the patient.
* VitD\_levels: Vitamin D levels measured in the patient.
* Doc\_visits: Number of doctor visits made by the patient.
* Full\_meals\_eaten: Number of full meals eaten daily by the patient.
* vitD\_supp: Use of vitamin D supplements by the patient.
* Initial\_days: Length of the patient's initial hospital admission.
* TotalCharge: Total cost incurred during the patient's initial hospital stay.
* Additional\_charges: Additional costs incurred during the patient's stay.

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.

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

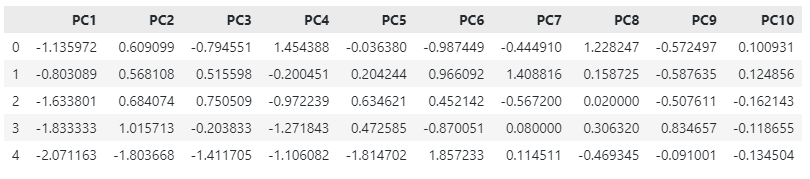
6. df1\_pca = pd.DataFrame(data=principal\_components, columns=['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7', 'PC8', 'PC9', 'PC10'])

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 elbow rule, visualized through a scree plot. The scree plot showed an "elbow" point at three components, suggesting that these components capture most of the dataset’s variance (Awan et al., 2019). The principal components retained represent the majority of information in the dataset.

Works Cited

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