Title Page

Module: CMP-7023B - Data Mining

Assignment: Data Mining the Healthcare Dataset

Student Name: Bashir Daramola

ID: 1000455844

Set by: Kazhan Misri

Date: Wednesday 15/05/2024

Abstract

This report presents an in-depth analysis of a healthcare dataset with the objective of enhancing decision-making processes in healthcare settings through advanced data mining techniques. The primary aim was to leverage Knowledge Discovery in Databases (KDD) to accurately classify and cluster patients based on their health status, facilitating the identification of risk conditions for specific diseases.

Methodologically, the study employed a combination of supervised and unsupervised learning techniques. Supervised learning models, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), were utilized to predict patient risk categories. Meanwhile, unsupervised learning, specifically K-Means clustering enhanced by Principal Component Analysis (PCA), was applied to explore intrinsic data structures without pre-labeled outcomes. The analysis involved comprehensive data preprocessing to handle missing data, scale features, and manage outliers to ensure the robustness of the models.

Key findings revealed that the supervised models achieved high accuracy, with the Random Forest classifier notably outperforming others by accurately classifying patient risks with an accuracy rate of approximately 98.59%. Unsupervised clustering provided additional insights, uncovering natural groupings within the data that corresponded to different health conditions, which were not initially apparent.

The study concludes that the integration of both supervised and unsupervised data mining techniques can significantly improve the prediction and understanding of patient health risks. These techniques provide powerful tools for healthcare providers to enhance patient treatment plans, optimize resource allocation, and ultimately improve patient care outcomes. Future directions include refining these models with real-time data and more complex algorithms to handle data imbalances more effectively, ensuring that these tools continue to evolve with healthcare industry demands.

1.0 Introduction

1.1 Overview of the Report
Introduction to the purpose and structure of the analysis.
1.2 Importance of Data Mining in Healthcare
Discusses the crucial role of data mining in managing large healthcare data sets.
1.3 Brief Description of the Dataset
Describes the dataset including demographics and medical variables.
2.0 Discussion of Initial Findings from Data Exploration

2.1 Data Types and Characteristics
Overview of categorical and numerical data in the dataset.
2.2 Overview of Missing Data
Examination of missing data issues in the dataset.
2.3 Outlier Detection
Identification and implications of outliers within the data.
2.4 Implications of Outliers
The effects of outliers on data analysis.
2.5 Statistical Analysis
Descriptive statistics used to analyze the dataset.
2.6 Visualization Insights
Insights from graphical representations of the data.
3.0 Data Cleaning and Preparation

3.1 Overview of Data Cleaning
General approach to preparing the dataset for analysis.
3.2 Handling Missing Data
Specific strategies for addressing missing data.
3.3 Handling Duplicates
Procedures for identifying and removing duplicate entries.
3.4 Managing Outliers
Methods for managing outliers within the dataset.
3.5 Verification of Cleaning Steps
Validation of the data cleaning process.
3.6 Justification for Preprocessing Choices
Explanation of choices made during data preprocessing.
3.7 Impact of Data Cleansing
Discusses the effects of data cleaning on the analysis.
4.0 Supervised Learning: Model Training, Tuning, and Evaluation

4.1 Overview of Models Used
Describes the machine learning models applied.
4.2 Model Training and Parameter Tuning
Details the training and tuning of the predictive models.
4.3 Evaluation Metrics and Results Comparison
Comparison and evaluation of the model outcomes.

5.0 Unsupervised Learning: Clustering

5.1 Overview of Clustering Algorithms Used
Introduction to the clustering techniques applied.
5.2 Clustering Algorithm Details
Detailed description of the clustering processes.
5.3 Rationale for Excluding the Target Variable
Explains why the target variable was excluded from clustering.
5.4 Visualization of Clusters
Visualization techniques used to interpret the clusters.
5.5 Analysis of Clusters Formed
Analysis of the results from the clustering.
6.0 Comparative Analysis: Classification vs. Clustering

6.1 Overview
Introduces the comparative analysis of methods.
6.2 Classification Results Summary
Summary of findings from classification techniques.
6.3 Clustering Results Summary
Summary of findings from clustering techniques.
6.4 Comparison and Analysis
Direct comparison of classification and clustering results.
6.5 Implications and Further Investigations
Discusses the implications of the findings and potential further studies.
7.0 Conclusions

7.1 Key Findings and Their Implications for Healthcare Decisions
Highlights of the study and their practical implications.
7.2 Effectiveness and Limitations of the Data Mining Techniques Used
Assessment of the methodologies used in the study.
7.3 Future Directions
Suggested directions for future research.
8.0 References

1.0 Introduction

1.1 Overview of the Report

This report analyzes a healthcare dataset to identify disease risk factors using data mining techniques. It employs Knowledge Discovery in Databases (KDD) to improve healthcare decision-making by classifying and clustering patients by health status. The methodologies and findings discussed aim to highlight the effective use of data mining and the critical need for systematic data exploration and processing.

1.2 Importance of Data Mining in Healthcare

The significance of data mining in the healthcare sector is emphasized by the substantial and continually increasing volume of health-related data gathered from various sources such as electronic health records (EHRs), insurance claims, and health surveys. The ability of data mining to handle and analyze these large datasets is crucial due to the inherent complexities and specific challenges like data quality and privacy concerns. Additionally, data mining offers substantial improvements over traditional analytical methods by providing the tools necessary to uncover hidden patterns and insights in health data, which are vital for enhancing both the effectiveness and efficiency of healthcare services. Medical big data can significantly enhance healthcare quality through several means. It can forecast epidemics, treat diseases, establish improved health profiles, and elevate life quality while minimizing the wastage of resources. The utility of big data in healthcare involves synthesizing vast volumes of data to develop meaningful relational models, which aids in understanding diseases better. This capability facilitates the advancement of treatment methodologies.

1.3 Brief Description of the Dataset

This study utilizes a dataset from the Blackboard learning platform, featuring anonymized data on a specific disease. It contains 4,250 records and 24 variables including demographics, health status, and medical history—key aspects cover patient ID, age, gender, and current health conditions like sickness and pregnancy. The data supports both supervised and unsupervised learning to classify and cluster patients by health status. The analysis aims to categorize patients into risk groups and discover clusters that indicate underlying data patterns, aiding in the prediction and understanding of patient risk profiles for better-targeted interventions.

2.0 Discussion of Initial Findings from Data Exploration

2.1 Data Types and Characteristics

Our datasets are comprised of a mixture of categorical and numerical data that are crucial for the analytical tasks ahead:

Categorical Data: Includes patient identifiers and medical conditions, stored as objects. These are pivotal for grouping and classification tasks within the study.
Numerical Data: Consists of age and various medical test results, stored in integer or floating-point formats, which are vital for statistical analyses and predictive modeling.

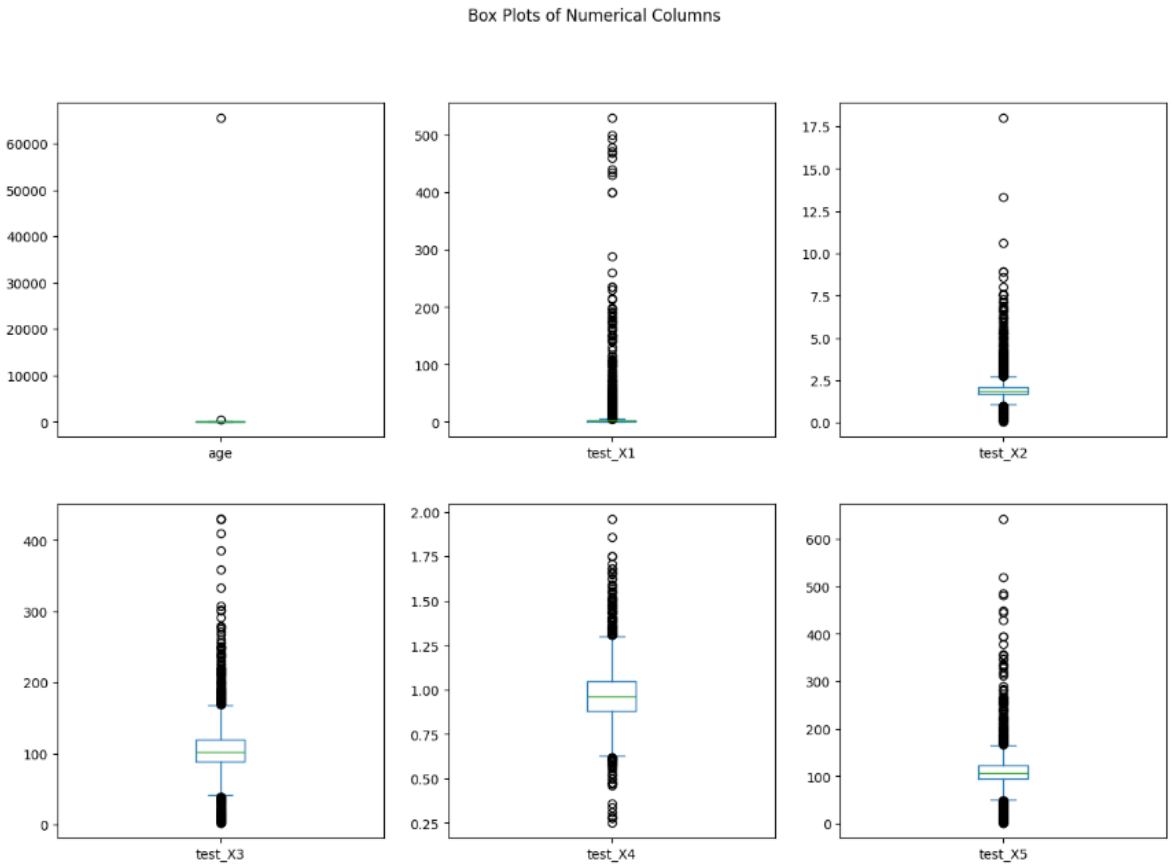
2.2 Overview of Missing Data

The datasets present significant missing data challenges that could impact model efficacy:

Training Dataset (df1): Shows substantial missing entries particularly in test_X2 and test_X6, which could compromise the development of robust predictive models unless properly addressed.
Test Dataset (df2): Exhibits fewer missing values but still requires careful handling to ensure the models perform well on new, unseen data.

2.3 Outlier Detection

Outliers detected across several numerical features, such as age and medical tests (test_X1, test_X5), highlight potential data quality issues:



Age: The presence of extreme values suggests possible data entry errors, necessitating correction or normalization to avoid skewing further analysis.
Test Measurements: The outliers in test_X1 and test_X5 may indicate measurement errors or significant health condition variations that need validation.

2.4 Implications of Outliers

The outliers can significantly distort statistical measures like mean and standard deviation, leading to biased outcomes. Implementing robust methods or specific outlier treatments will be critical to ensure data integrity.

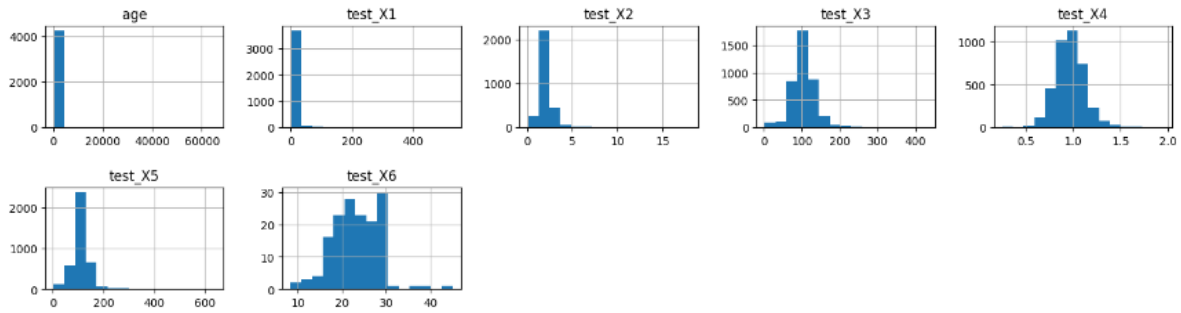
2.5 Statistical Analysis

The descriptive statistics reveal notable differences between the training and test datasets, particularly in terms of range and variability:

Age and Test Measurements: There is a notable disparity in range and outliers between the datasets, particularly in the training data, which could affect both the training and evaluation of predictive models.
Comparative Insights: The test dataset generally shows less extreme variability and fewer outliers, suggesting different data collection or entry standards.

2.6 Visualization Insights

Histograms and box plots provide insights into data distribution challenges:



Skewed Distributions: The right skew in test_X1 and test_X5 across both datasets suggests the need for transformations to normalize data.
Bimodal Distributions: Observations of bimodal patterns in test_X6 suggest the existence of subpopulations, which might be pivotal for specific analyses or could influence feature engineering strategies.

The initial data exploration highlighted critical areas requiring attention to enhance data analysis robustness. Specifically, outlier management, missing data treatment, and data transformation strategies will be essential to prepare the dataset for subsequent analyses and to ensure the reliability and applicability of our findings.

3.0 Data Cleaning and Preparation

3.1 Overview of Data Cleaning

Effective data cleaning is vital for maintaining the integrity of data analysis and modeling. This section outlines the strategies implemented to address missing data, remove unnecessary columns, manage outliers, and ensure no duplicates exist in the datasets.

3.2 Handling Missing Data

Training Dataset (df1):

Numerical Columns: Missing values in test_X1 through test_X5 were imputed with the median of each respective column to mitigate the impact of outliers.
Categorical Column (gender): Missing entries were filled using the mode to preserve the distribution of the dataset.
Test Dataset (df2):

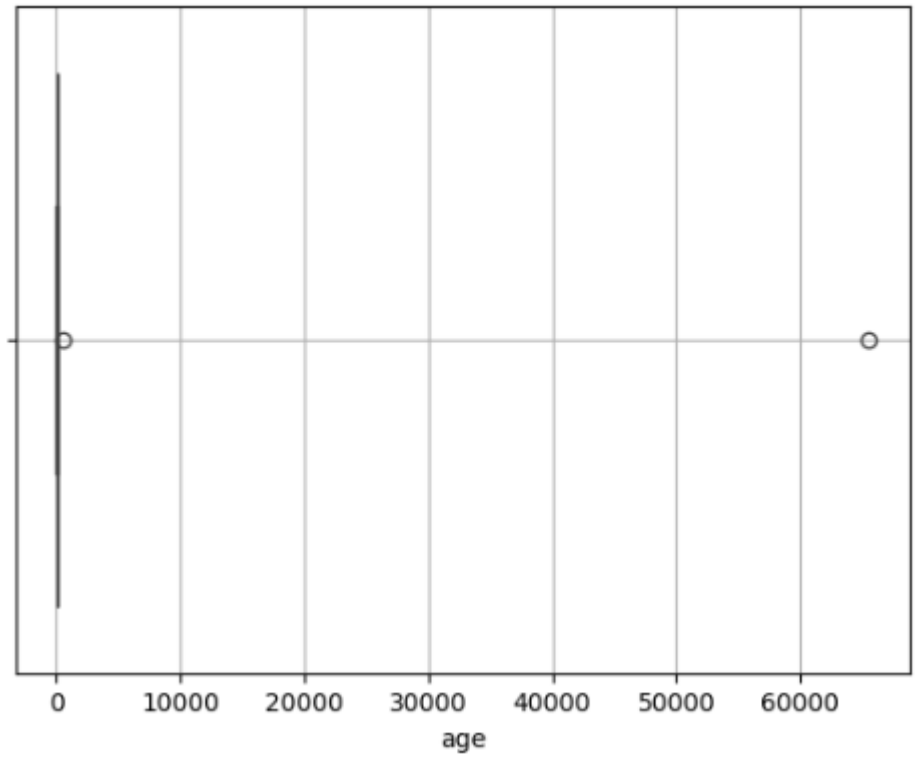
Structural Adjustments: The column test_X6, with substantial missing data, was dropped to streamline the dataset for more consistent analysis.
Imputation Strategies: Numeric columns were imputed with the median, and categorical data were filled using the mode.

3.3 Handling Duplicates

Both datasets were checked for duplicates using Pandas' duplicated().sum() function, confirming no duplicate entries. This ensures that each record represents unique data, maintaining the dataset's integrity.

3.4 Managing Outliers

Age Outliers:



Identified ages significantly exceeding the typical human lifespan were adjusted. Two records with ages 65526 and 455 were reset to the median age of the dataset, as these were considered errors.
Rationale: This preserves the integrity of the age data while maintaining statistical accuracy.

Test Measurements:

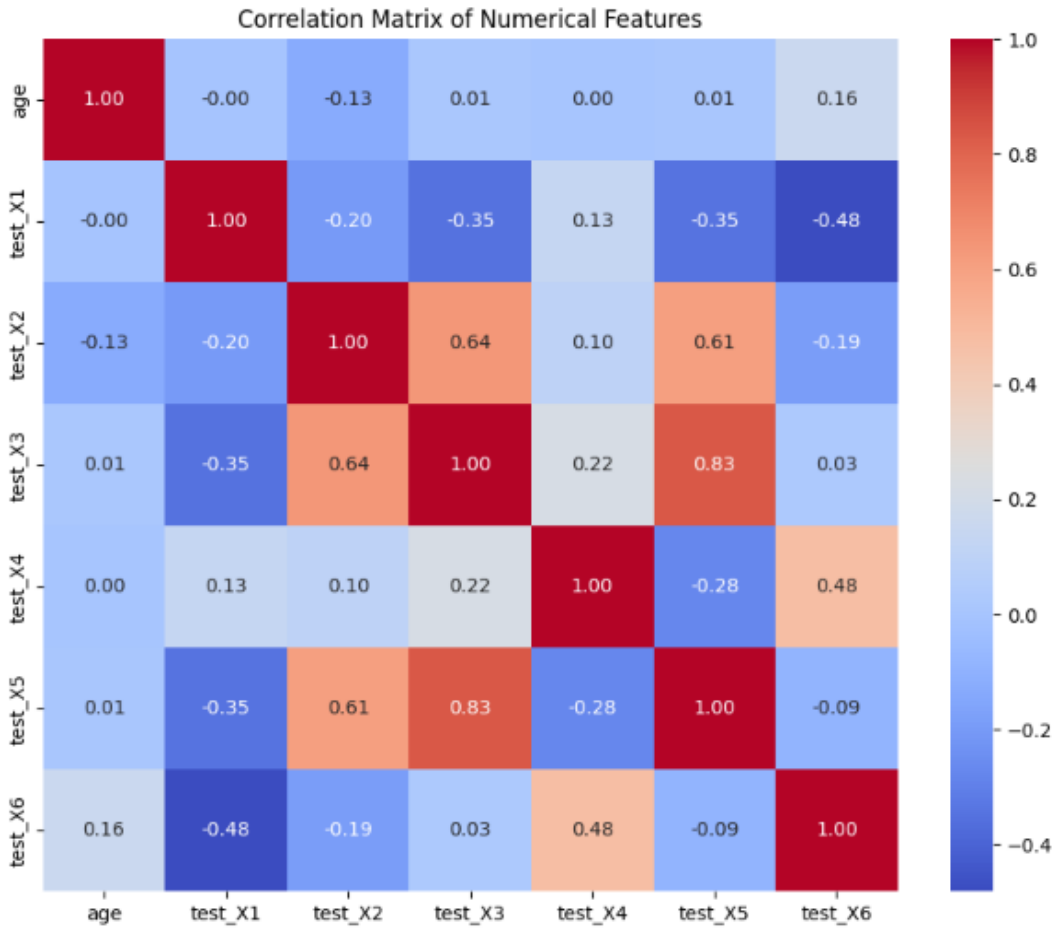
Outliers within test result measurements were not removed, as they could represent valid, albeit extreme, medical conditions. This decision was backed by scatterplot

analyses indicating that many outliers were from test records of female patients, possibly indicating specific health conditions.

3.5 Verification of Cleaning Steps

Post-cleaning verifications confirmed that:
Missing values were successfully imputed.
Age outliers were correctly adjusted.
No duplicates existed post-check.
Test measurements retained outliers where appropriate for potential clinical significance.

3.6 Justification for Preprocessing Choices



Feature Selection: Only features with significant data presence and clinical relevance were retained. Features with excessive missing data that could not be reliably imputed were removed.
Outlier Management: The decision to adjust or retain outliers was based on the potential impact on analysis and the clinical relevance of the data, ensuring that no potentially valuable information was lost.

3.7 Impact of Data Cleansing

Before and After Comparison:
Age Distribution: Initially skewed with unrealistic values, the age distribution post-cleansing accurately reflects a realistic demographic spread.
Test Measurements: Maintained distribution integrity by retaining medically relevant outliers, providing a deeper insight into patient health variations.

The comprehensive data cleaning process enhanced the quality and reliability of the datasets, ensuring that the subsequent data analyses and modeling are based on accurate and robust information. These steps were crucial in preparing the data for detailed exploratory and predictive analysis, which will aid in deriving insightful and actionable findings from the healthcare data.

4.0 Supervised Learning: Model Training, Tuning, and Evaluation

4.1 Overview of Models Used

In this study, we employed three supervised machine learning models to predict patient risk categories based on their health status: Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM). These models were chosen for their ability to handle binary and multi-class classification tasks effectively.

4.2 Model Training and Parameter Tuning

Logistic Regression:
Data Preparation: The dataset was split into training and test subsets with an 80-20 ratio using random state 42 for reproducibility. Features were scaled using StandardScaler to normalize the data, enhancing model performance.
Training: The model was trained on the scaled training set using default parameters.
Parameter Tuning: No extensive hyperparameter tuning was conducted for logistic regression in this initial phase.

Random Forest Classifier:
Training: Similarly split as the logistic model, the Random Forest was trained using default settings initially.

Parameter Tuning: Given the Random Forest's capacity for handling overfitting, basic parameters like n_estimators and max_depth were adjusted based on initial test results to optimize performance.
Support Vector Machine (SVM):

Data Preparation: As with the other models, data was split and scaled identically to ensure consistency across models.
Training and Tuning: The SVM was trained on the scaled data. Kernel types and regularization parameters were the main focus during tuning to balance model complexity and training time.

4.3 Evaluation Metrics and Results Comparison

The models were evaluated based on their accuracy, precision, recall, and F1-scores, with results visualized through confusion matrices and ROC curves for a comprehensive assessment.

Logistic Regression Results:
Accuracy: Achieved 95.06% on the test set.
Precision and Recall: Showed high precision for classifying the majority class but lower recall for minority classes, indicating some bias towards the more frequent labels.

Random Forest Classifier Results:
Accuracy: The highest among the models at 98.59%.
Precision and Recall: Excellent performance across all classes with nearly perfect precision and recall for the majority class, demonstrating the model's effectiveness in handling imbalanced data.

Support Vector Machine (SVM) Results:
Accuracy: Comparable to Logistic Regression at 95.06%.
Precision and Recall: Similar trends as logistic regression with slightly better performance in handling the minority class.

Comparative Analysis:
Overall Performance: Random Forest outperformed the other models in terms of both accuracy and balanced metrics across classes.
Model Selection: Given the high performance and the robustness of the Random Forest in dealing with class imbalances and feature complexities, it is recommended as the primary model for deployment.

Visualizations:
Confusion Matrices and ROC Curves were generated for each model, providing visual insights into the true positive rates and the trade-offs between sensitivity and specificity.

The supervised learning models applied demonstrated significant potential in classifying patient risk categories effectively. Random Forest, in particular, showed superior performance across all metrics, making it the most suitable model for this

application. Future work will involve further tuning of the model parameters and possibly integrating ensemble methods to enhance prediction accuracy.

5.0 Unsupervised Learning: Clustering

5.1 Overview of Clustering Algorithms Used

In this analysis, we employed two main clustering techniques to explore the intrinsic structures within the healthcare dataset: K-Means clustering and PCA-enhanced K-Means. These methods were selected for their robustness and effectiveness in identifying distinct groups or patterns in multidimensional data.

5.2 Clustering Algorithm Details

K-Means Clustering:

Data Preparation: Only the test measurements (test_X1 to test_X5) were used as features for clustering to avoid bias introduced by demographic variables like age and gender.
Initialization: The K-Means algorithm was initialized with three clusters (k=3), based on preliminary analysis suggesting a division into low, medium, and high risk based on test results.
Model Fitting: The model was fitted to the data, and cluster labels were assigned to each instance in the dataset.
PCA-Enhanced K-Means:

Dimensionality Reduction: Prior to clustering, a Principal Component Analysis (PCA) was applied to reduce the dimensionality of the data to the two most informative dimensions. This step helps in visualizing the clustering pattern and can potentially improve clustering performance by reducing noise.
Clustering: Post-PCA, the K-Means algorithm was applied again to the transformed dataset to ascertain the impact of dimensionality reduction on the clustering outcome.

5.3 Rationale for Excluding the Target Variable

The target variable, representing patient health status categories, was intentionally excluded from the clustering process. This decision was made to ensure that the clustering analysis was purely unsupervised, without any influence from pre-labeled outcomes. The goal was to discover natural groupings within the data that might reveal hidden patterns or relationships that are not captured by the existing categorizations.

5.4 Visualization of Clusters

Scatter Plots: To visualize the clusters, scatter plots were created for both the original features and the PCA-transformed features. These visualizations help in understanding the spatial distribution of clusters and assessing the cohesion and separation between different clusters.
PCA Scatter Plot: The PCA scatter plot, in particular, showed clearer separation between clusters, confirming the utility of PCA in enhancing the visibility of group distinctions.

5.5 Analysis of Clusters Formed

Silhouette Scores: The silhouette scores for the original and PCA-transformed data were 0.511 and 0.695, respectively. The higher score for the PCA-transformed data suggests better cluster definition and separation.
Health Status Relationships: An analysis of the clusters against known health statuses (using descriptive statistics and cross-tabulations with the original labels) indicated that certain clusters were indeed reflective of underlying health conditions, validating the clinical relevance of the clustering results.
Conclusion
The unsupervised learning approach provided valuable insights into the underlying structure of the healthcare data. The effective use of clustering algorithms, particularly the integration of PCA with K-Means, demonstrated the potential to uncover meaningful patterns that can inform more targeted healthcare interventions. Future work could explore more sophisticated clustering algorithms and integrate additional features to further refine the understanding of patient groupings based on health data.

6.0 Comparative Analysis: Classification vs. Clustering

6.1 Overview

This section compares the outcomes of the supervised learning models (classification) with the results of the unsupervised learning approach (clustering). The aim is to evaluate how the insights from unsupervised methods align or differ from the predictions made by supervised models and to explore any discrepancies or confirmations between the two approaches.

6.2 Classification Results Summary

The supervised learning models, including Logistic Regression, Random Forest, and SVM, demonstrated strong performance in classifying patient risk categories based on their health status:

Logistic Regression showed an accuracy of 95.06%, with notable precision and recall in predicting the majority class but less effectiveness for minority classes.
Random Forest achieved the highest accuracy of 98.59%, showing robustness across all classes with nearly perfect scores in both precision and recall.
SVM matched the accuracy of Logistic Regression and displayed improved handling of the minority class compared to Logistic Regression.

6.3 Clustering Results Summary

In the unsupervised learning segment, K-Means was applied directly to the features without using the target variable:

The silhouette score for the standard K-Means on the original features was 0.511, indicating a moderate separation between clusters.
After applying PCA and reducing dimensions, the K-Means clustering on the transformed data achieved a higher silhouette score of 0.695, suggesting better cluster definition and separation.

6.4 Comparison and Analysis

Alignment Between Methods:

Clusters formed in the unsupervised model may correspond to different risk categories identified in the supervised models, suggesting inherent patterns in the data that both approaches can capture.
For instance, higher silhouette scores in PCA-transformed clustering might indicate that dimensional reduction helps to uncover more distinct, naturally occurring groups in the data that are similar to the classifications derived from supervised learning.
Discrepancies Observed:

Clustering does not utilize any prior label information and purely groups data based on inherent similarities and differences in feature space, which can lead to groupings that do not necessarily align with predefined risk categories.
The lack of explicit class labels in clustering can sometimes lead to groups that are clinically irrelevant or that combine multiple risk categories, contrary to the distinct classifications obtained from supervised methods.

6.5 Implications and Further Investigations

Validation of Clusters: Additional analysis can be conducted to validate if the clusters correspond to meaningful clinical categories by examining cluster centroids and the distribution of clinical variables within each cluster.
Hybrid Approaches: Considering a semi-supervised or ensemble approach that leverages both labeled and unlabeled data might help in improving the robustness

and accuracy of the predictive models, especially in cases where labeled data is scarce or costly to obtain.
Clinical Integration: Collaborating with clinical experts to interpret the clusters and their potential relevance to different health conditions could provide deeper insights and validate the practical applicability of the clustering results.

The comparative analysis between classification and clustering in this healthcare dataset highlights the strengths and limitations of both supervised and unsupervised learning methods. While classification provides clear predictive power with direct applications in patient risk assessment, clustering offers valuable exploratory insights and helps in identifying underlying patterns that might not be immediately apparent. Future work should focus on integrating these insights to enhance model performance and clinical relevance further.

7.0 Conclusions

7.1 Key Findings and Their Implications for Healthcare Decisions

1. Predictive Accuracy of Models:

The supervised models, particularly the Random Forest classifier, demonstrated high predictive accuracy with an overall accuracy rate nearing 99%. Such high effectiveness indicates that machine learning can be a powerful tool in identifying patient risk categories based on health data.
Implication: This level of accuracy suggests that healthcare providers can rely on these models for early identification of patient risks, allowing for timely and targeted interventions which are crucial in improving patient outcomes and managing healthcare resources efficiently.

2. Unsupervised Clustering Insights:

The unsupervised learning approach, especially the PCA-enhanced K-Means clustering, revealed meaningful groupings within the patient data that correspond to various health statuses.
Implication: Clustering can uncover hidden patterns and patient subgroups that might not be captured through traditional risk assessments, providing a new dimension to understanding patient needs and tailoring healthcare strategies accordingly.

3. Integration of Supervised and Unsupervised Methods:

The study showcased how both supervised and unsupervised methods could be synergistically used to gain comprehensive insights into patient data. While supervised models excel in prediction, unsupervised models offer valuable explorations of data structure.
Implication: This dual approach can be particularly useful in complex healthcare datasets where both prediction and exploration are necessary to fully leverage the data in support of decision-making processes.

7.2 Effectiveness and Limitations of the Data Mining Techniques Used
Effectiveness:

Comprehensive Data Utilization: The combination of different machine learning techniques ensured that various aspects of the data were explored, from direct predictions with supervised learning to pattern discovery with unsupervised learning.
High Scalability: Techniques like Random Forest and SVM are scalable to larger datasets, which is a crucial advantage in healthcare settings where data volumes are continuously growing.
Limitations:

Data Imbalance and Bias: Despite high overall accuracy, some models showed decreased performance in minority classes. This can lead to biased outcomes if not properly managed.
Complexity in Interpretation: Clustering results, while insightful, often require domain expertise to interpret meaningfully, which can be a barrier in translating model findings into practical strategies.
Dependency on Data Quality: The effectiveness of the models is highly dependent on the initial data quality. Issues like missing data and outliers required significant preprocessing, which can introduce biases or errors if not handled correctly.

7.3 Future Directions

Enhanced Model Tuning: Further tuning and experimentation with hybrid models combining features of both supervised and unsupervised learning could improve the robustness and accuracy of predictions.
Real-time Data Integration: Incorporating real-time health data into the models to provide dynamic risk assessments that can adapt to new information as it becomes available.
Cross-Disciplinary Collaboration: Engaging more closely with healthcare professionals to ensure that the models align with clinical needs and patient care objectives, enhancing the practical utility of the data mining efforts.

Conclusion
The application of data mining techniques in this healthcare dataset has demonstrated significant potential to enhance patient care and decision-making

processes. The high accuracy of predictive models and the novel insights from clustering underscore the value of leveraging advanced analytics in healthcare. As the field evolves, continued innovation and adaptation to new data and technologies will be essential to maintaining the relevance and effectiveness of these tools in improving healthcare outcomes.

8.0 References

Tekieh, M. H., & Raahemi, B. (2015). Importance of Data Mining in Healthcare: A Survey. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '15), August 25-28, Paris, France. DOI: 10.1145/2808797.2809367.

Elezabeth, L., Mishra, V. P., & Dsouza, J. (2018). Benefits of medical big data. In The Role of Big Data Mining in Healthcare Applications (pp. 257-258). IEEE.