FINAL PROJECT – ODD SEMESTER 2024/2025

DATA MINING

"Analyzing Medical Costs: Clustering and Regression Approaches"

Created By:

SI-46-INT

Bintang Amelia Wijaya	1202224135
Vakha Adzhany	1202224190
Freska Prisia Putri	1202224226
Vanessa Wiyen Cristie	1202224397

BACHELOR'S PROGRAM IN INFORMATION SYSTEMS

FACULTY OF INDUSTRIAL ENGINEERING

TELKOM UNIVERSITY

2024

PREFACE

This paper, titled *"Analyzing Medical Costs: Clustering and Regression Approaches,"* is the result of our final project for the Data Mining course in the Odd Semester of 2024/2025. The study applies data mining techniques, specifically clustering and regression analysis, to explore the factors influencing medical costs. Our goal is to identify significant cost drivers and provide insights that could contribute to improving healthcare affordability and resource management.

To ensure a systematic approach, we follow the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which guides us through stages such as data understanding, preparation, modeling, and evaluation. This project is a collaborative effort by Bintang Amelia Wijaya, Vakha Adzhany, Freska Prisia Putri, and Vanessa Wiyen Cristie, international students in the Bachelor's Program in Information Systems at Telkom University.

We would like to extend our sincere gratitude to our instructor, Nur Ichsan Utama, for his valuable guidance and support throughout the course and project. The knowledge and resources provided have been crucial to the successful completion of this study. We hope that the findings presented in this paper offer meaningful insights into the complexities of medical cost analysis and highlight the potential of data mining techniques to address real-world challenges in healthcare.

TABLE OF CONTENTS

CHAPTER I: INTRODUCTION

I. 1 Background

The rising costs of healthcare are a significant challenge globally, prompting research into understanding and mitigating these expenses. Medical cost analysis involves identifying the primary drivers of costs, understanding patient demographics, and exploring usage patterns of healthcare services.

Recent studies highlight the role of data mining techniques in revealing insights from complex healthcare datasets (Anderson et al., 2021; Guo & Chen, 2019) . Regression and clustering techniques are particularly effective for identifying key cost determinants and grouping individuals based on similar cost patterns, respectively. This paper employs the CRISP-DM methodology to conduct a comprehensive analysis, starting from data preparation and culminating in deployment, demonstrating the practical utility of these techniques in real-world scenarios.

I. 2 Problem Statement

Medical costs are determined by a multitude of factors, including age, gender, geographic location, pre-existing conditions, and healthcare utilization patterns. However, the volume and complexity of healthcare data make it difficult to extract actionable insights.

This study aims to address the following research questions:

1. Which factors significantly influence medical costs, and to what extent?
2. How can clustering techniques group individuals with similar medical cost profiles to aid targeted interventions?

Addressing these questions is critical for improving healthcare affordability and resource allocation. Past research underscores the importance of leveraging predictive modeling and segmentation techniques to tackle such challenges effectively (Thomas et al., 2024) .

I. 3 Objectives

The primary objectives of this study are:

1. To identify significant predictors of medical costs using regression analysis.
2. To segment individuals into meaningful clusters using advanced clustering algorithms based on medical cost patterns.
3. To deploy a predictive model that integrates regression and clustering results for real-time application.

I. 4 Scope and Limitations

The scope of this study includes:

1. Dataset
   Using a publicly available medical cost dataset, focusing on key variables such as demographics, health conditions, and usage patterns.
2. Techniques
   Linear regression is employed to model cost predictors, and clustering algorithms such as K-Means and hierarchical clustering are employed for segmentation.
3. Deployment
   Demonstrating the final model in a simulated environment, showcasing its predictive and segmentation capabilities.

Limitations:

1. Data Availability

The analysis is restricted to the variables present in the dataset, potentially omitting important factors like socioeconomic status and lifestyle behaviors (Puka et al., 2022).

2. Generalizability

   Results may not apply to other datasets with different population characteristics or healthcare systems.

These limitations underscore the need for careful interpretation and future work to address identified gaps.

I. 2 Crisp DM Overview

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, established as a gold standard in data mining projects, guides this analysis (Wirth and Hipp, 2000). Its phases are:

1. Business Understanding

   This phase involves defining the objectives and framing research questions for analyzing medical costs. It ensures that the data mining project aligns with the business goals and provides a clear understanding of the problem to be solved.

2. Data Understanding

   In this phase, the dataset is explored for quality and relevance. This includes performing descriptive statistics and visualizations to identify patterns and anomalies. Understanding the data is crucial for making informed decisions about data preparation and modeling.

3. Data Preparation

   This phase involves preprocessing the data to ensure it is suitable for modeling. Tasks include handling missing values, normalizing features, and

encoding categorical variables. Proper data preparation is essential for the accuracy and efficiency of the modeling algorithms.

4. Modeling

   During the modeling phase, various algorithms are applied to the prepared data. For predicting costs, regression models such as linear regression are used. For segmentation, clustering algorithms like K-Means and DBSCAN are employed. The choice of model depends on the specific objectives and the nature of the data.

5. Evaluation

   The evaluation phase assesses the performance of the models using metrics such as R-squared for regression and silhouette scores for clustering. This ensures that the models meet the business objectives and are reliable for deployment.

6. Deployment

   In the final phase, the predictive and segmentation models are implemented in a simulated environment. This includes creating actionable dashboards and reports for stakeholders, ensuring that the insights gained from the data mining process are effectively communicated and utilized.

CHAPTER II: FRAMEWORK

II. 1 Business Understanding

1. Business Goals

   The goal of this study is to analyze and understand the factors influencing medical costs and to provide actionable insights that can help in reducing healthcare expenses. This includes identifying significant predictors of medical costs, segmenting individuals into meaningful clusters based on cost patterns, and deploying a predictive model to aid decision-making in real-world scenarios.

2. Problem Understanding

   Medical costs are influenced by various factors such as age, BMI, and smoking status. The challenge lies in analyzing the high volume and complexity of healthcare data to extract actionable insights. Addressing this issue involves:

   a. Identifying significant cost predictors using regression analysis.
   b. Grouping individuals with similar medical cost profiles using clustering techniques.
   c. Developing an integrated model for predictive and segmentation capabilities.

II. 2 Data Understanding

1. Data Collection

   The dataset used in this study is a publicly available medical costs dataset, which includes the following variables:

   a. Age: Patient age.
   b. Sex: Gender of the patient.
   c. BMI: Body Mass Index.
   d. Children: Number of dependents.

e. Smoker: Smoking status (yes/no).

f. Region: Geographic region.

g. Charges: Medical costs incurred.

Table 1: Insurance Dataset

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.9
18	male	33.7701	1	no	southeast	1725.55
28	male	33.0003	3	no	southeast	4449.46
33	male	22.705	0	no	northwest	21984.5
32	male	28.88	0	no	northwest	3866.86
…	…	…	…	…	…	…

1338 rows × 7 columns

2. Data Exploration

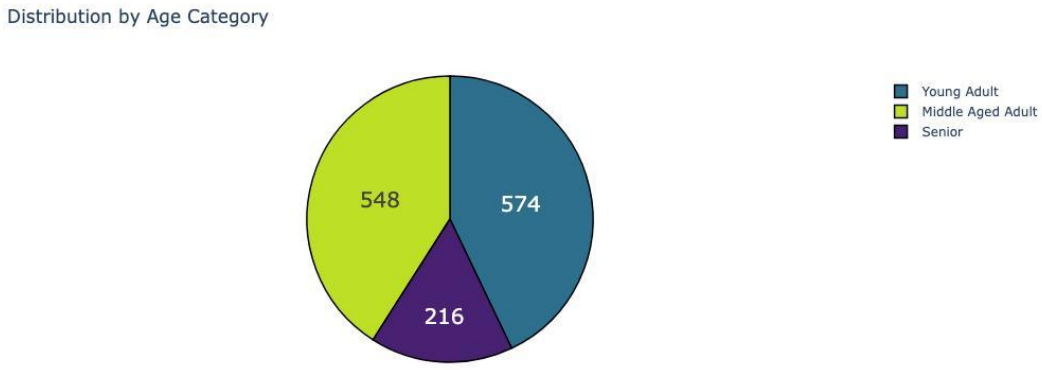

Figure 1: Distribution by Age Category

The pie chart reveals the distribution of individuals across three age categories:

- Young Adults (574 individuals)
- Middle-Aged Adults (548 individuals)
- Seniors (216 individuals)

This balanced distribution for Young and Middle-Aged Adults suggests that any model involving age won't suffer from severe class imbalance within these groups.

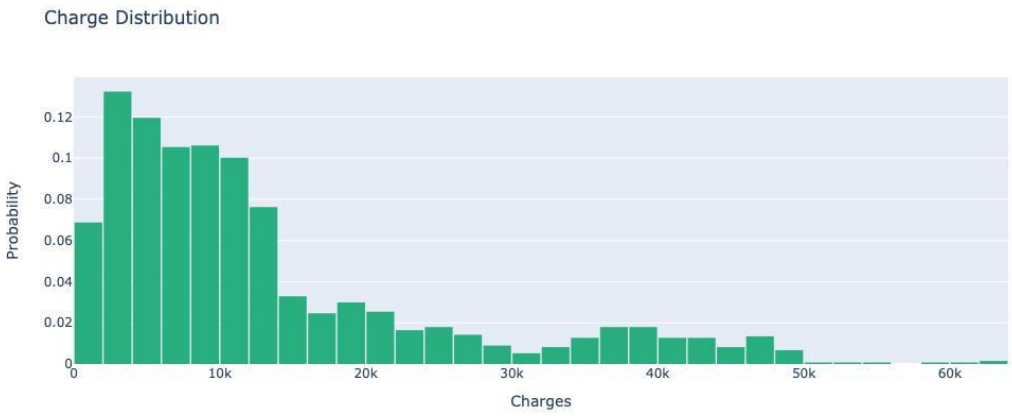

Figure 2: Charge Distribution

The histogram is right-skewed, showing that most charges are clustered below $20,000, with a long tail extending up to $60,000. There's a steep drop in probability beyond $50,000. The skewed nature indicates that models predicting charges might require log transformation or other scaling techniques to stabilize variance.

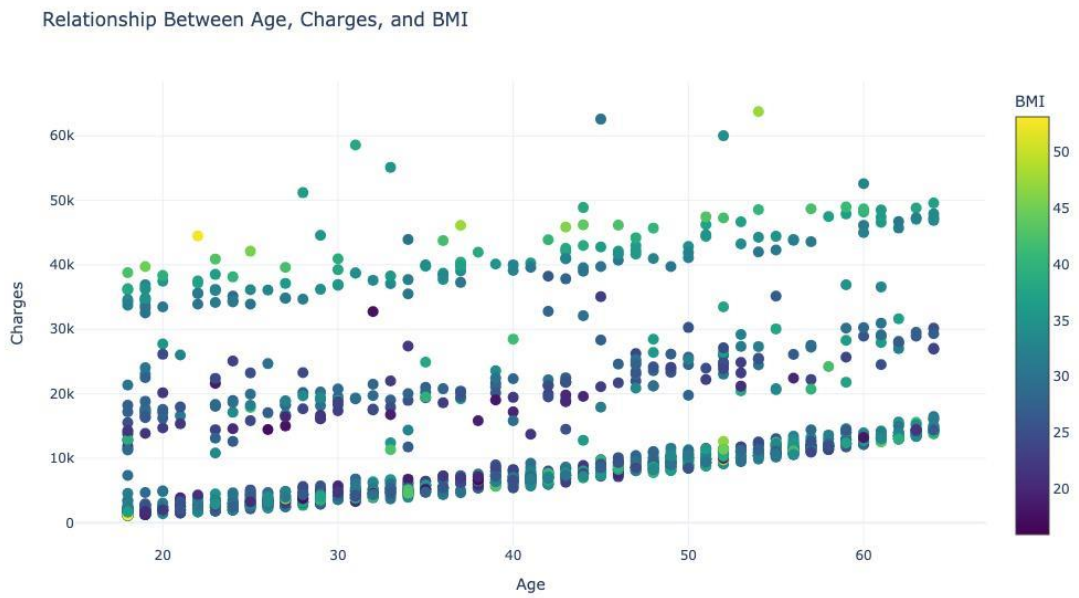

Figure 3: Relationship Between Age, Charges, BMI

There's a clear trend where charges increase with age. Higher BMI (shown via color intensity) correlates with higher charges, especially at the upper end of the age spectrum. There's also stratification, suggesting distinct clusters in charges for given age-BMI combinations.

Figure 4: Correlation Heatmap

This heatmap visualizes the correlation coefficients between age, charges, and BMI. Correlation ranges from -1 to 1, where:

- 1: Perfect positive correlation.
- 0: No correlation.
- -1: Perfect negative correlation.

The correlation between charges and age is moderately high, indicating that as people age, medical charges tend to increase. The correlation between charges and BMI is weaker but still positive, suggesting that individuals with higher BMI may face slightly higher medical costs.

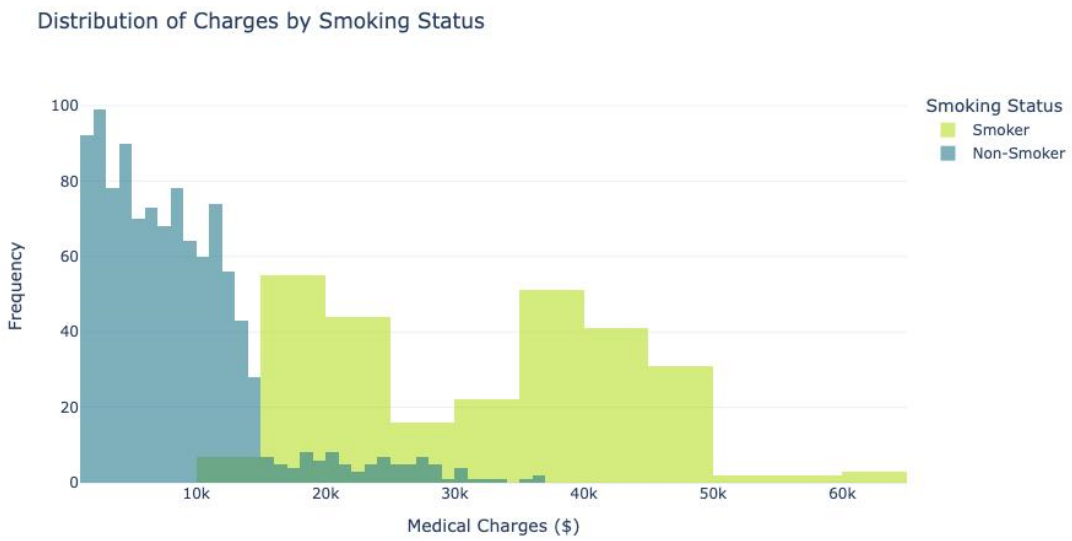

Figure 5: Distribution of Charges by Smoking Status

This histogram shows the distribution of medical charges for smokers and non-smokers.

- Smokers: Distribution is right-skewed, with most charges above $20,000.
- Non-Smokers: Charges are concentrated below $20,000.

  Insights:

Smokers generally have much higher charges compared to non-smokers, suggesting smoking status is a strong predictor of medical expenses.

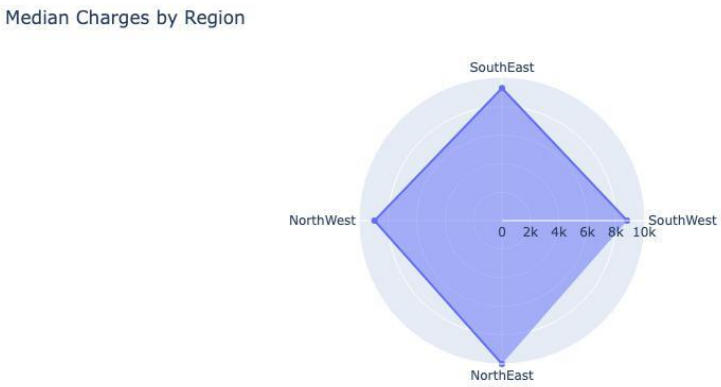

Figure 6: Median Charges by Region

This radar chart represents the median medical charges across four regions: Southeast, Southwest, Northwest, and Northeast. The Southeast region has the highest median charges, which could be due to regional factors such as lifestyle, healthcare costs, or demographics. Other regions have similar and lower median charges, suggesting regional disparities that might affect prediction models.

II. 3 Data Preparation

1. Data Cleaning

   The code cleans the data frame by removing rows with missing values (NaN) and dropping duplicate rows. Here's what happens step by step:

   a. Drop rows with NaN values:

      The dropna() method removes any row that contains at least one NaN value.

   b. Drop duplicate rows:

      The drop_duplicates() method removes rows that are entirely identical to another row in the DataFrame.

   c. Count of rows reduced:

Initially, the df has 1338 rows. After cleaning, the new df_cleaned has 1337 rows. This indicates that only one row was removed, and it could have been removed due to either:

- Containing a NaN value.
- Being a duplicate of another row.

Table 2: Cleaned Insurance Dataset

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.9
18	male	33.7701	1	no	southeast	1725.55
28	male	33.0003	3	no	southeast	4449.46
33	male	22.705	0	no	northwest	21984.5
32	male	28.88	0	no	northwest	3866.86
…	…	…	…	…	…	…

1337 rows × 7 columns

2. Data Splitting
   a. Shuffling the dataset:

      Before splitting, the dataset is randomly shuffled to eliminate any order or bias in the data that could affect training and evaluation.

      - sample(frac=1): Randomly rearranges all rows in the dataset. The parameter frac=1 ensures all rows are included.
      - random_state=1: Ensures the shuffle is consistent and reproducible every time the code runs.

   b. Separating features and target:

      The dataset is then divided into features (X) and the target variable (y):

- X: Contains all the columns except "charges", which are the inputs the model will use for prediction.
- y: Contains only the "charges" column, which is the value the model will learn to predict.

This separation is crucial for supervised learning, where the goal is to predict the target (y) based on the features (X).

c. Splitting into training and testing sets:

The features (X) and target (y) are split into:

- Training set: Used to train the machine learning model.
- Testing set: Used to evaluate the model's performance on unseen data.

test_size=0.2: Specifies that 20% of the data will be used for testing, while 80% will be used for training.

3. Data Transformation

a. Custom categorical encoder:

1) Initialization:

The CategoricalEncoder class provides a custom implementation for encoding categorical variables.

- encoding: Specifies the encoding type ('onehot' or 'ordinal').
- categories: Defines possible categories for each feature ('auto' lets the encoder infer them).
- dtype: The data type of the output.
- handle_unknown: Handles unseen categories ('error' raises an exception, or they can be ignored).

2) Fit Method:

- Validates input X (ensures it’s valid and in the correct format).

- Creates a LabelEncoder for each feature to encode categories as integers.
- Handles unknown categories according to the handle_unknown parameter.

3) Transform Method:
   - Converts categorical values into numerical format.
   - Supports both onehot encoding (returns a sparse matrix or dense array) and ordinal encoding.

b. DataFrameSelector:

The DataFrameSelector class is a utility to handle data selection. Scikit-learn doesn't natively support pandas DataFrames, so this class selects specific columns from a DataFrame. It works seamlessly within pipelines.

1) Initialization:

   Takes the column names to be selected.

2) Fit and Transform:

   Fits the selector (no operation) and transforms the input by selecting specified columns.

c. ColumnTransformer:

The ColumnTransformer is used to combine multiple preprocessing pipelines for numerical and categorical data.

1) Identify Data Types:
   - Numerical columns: Exclude categorical (object) data.
   - Categorical columns: Include only object-type data.
2) Set Up Preprocessing Pipelines:
   - Numerical Pipeline: Uses StandardScaler to scale numerical features (mean = 0, std = 1).
   - Categorical Pipeline: Uses OneHotEncoder to encode categorical features as binary vectors.

- sparse_output=False ensures the result is a dense array.

3) Fit and Transform:

Applies the preprocessing steps to the training data (X_train).

II. 4 Modeling

1. Model Selection

a. OLS

OLS is a natural choice for predicting medical charges due to the following insights from data exploration:

1) Figure 3 shows that charges increase with age, and higher BMI correlates with higher charges. This trend indicates a linear relationship between the predictors (age, BMI) and the target variable (charges), making OLS regression an appropriate model for predicting continuous outcomes like charges.
2) The correlation heatmap (Figure 4) reveals moderate to strong positive correlations between charges and predictors like age and BMI, affirming that these features can explain variance in charges.
3) The task aims to predict charges, a continuous variable, which aligns well with the objective of OLS regression.

b. K-Means Clustering

K-means clustering was chosen for segmenting individuals into meaningful groups based on their characteristics (e.g., age, BMI, charges):

1) Figure 3 shows stratification of charges for age-BMI combinations, suggesting natural clusters of individuals with

similar characteristics. K-means clustering can identify these groups and enhance understanding of the dataset.

2) The age distribution (Figure 1) and smoking-status charge distribution (Figure 5) highlight distinct patterns. Clustering can group individuals based on such factors, revealing demographic or behavioral patterns.
3) K-means can handle multiple features simultaneously, such as age, BMI, smoking status, and region, providing a more comprehensive grouping than visual inspection alone.

2. Model Training
    a. OLS Regression for Predicting Charges
        1) Dataset Preparation
            - Training data (X_train, y_train) and testing data (X_test, y_test) are combined into separate DataFrames for better handling during the analysis.
            - The training set is randomly shuffled using sample(frac=1) to ensure the model is not biased by the order of rows in the dataset.
            - sm.add_constant(scaled_xtrain) adds a constant column to the feature matrix to account for the intercept in the OLS regression model.
            - y_train is converted to an array format to fit the regression model.
        2) Model Training
            - Initializes the OLS regression model using scaled features (X_train) and target variable (y_train).
            - results = model.fit():
            - Fits the regression model and computes the parameters that minimize the sum of squared residuals.

- Prints the detailed statistical summary of the regression model, including coefficients, p-values, and R-squared.

3) Key Metrics:
   - results.params: Lists the coefficients for each predictor, showing their contribution to the target variable.
   - results.rsquared: Indicates the proportion of variance in charges explained by the predictors. In this case, an R-squared of 0.7598 shows the model explains ~76% of the variance.

b. K-Means Clustering for Segmentation

1) Elbow Method for Optimal Clusters
   - Only BMI and Charges are used for clustering, as these features showed interesting patterns during exploration.
   - For each k (number of clusters), a K-Means model is trained, and inertia (sum of squared distances of samples to their closest cluster center) is recorded.

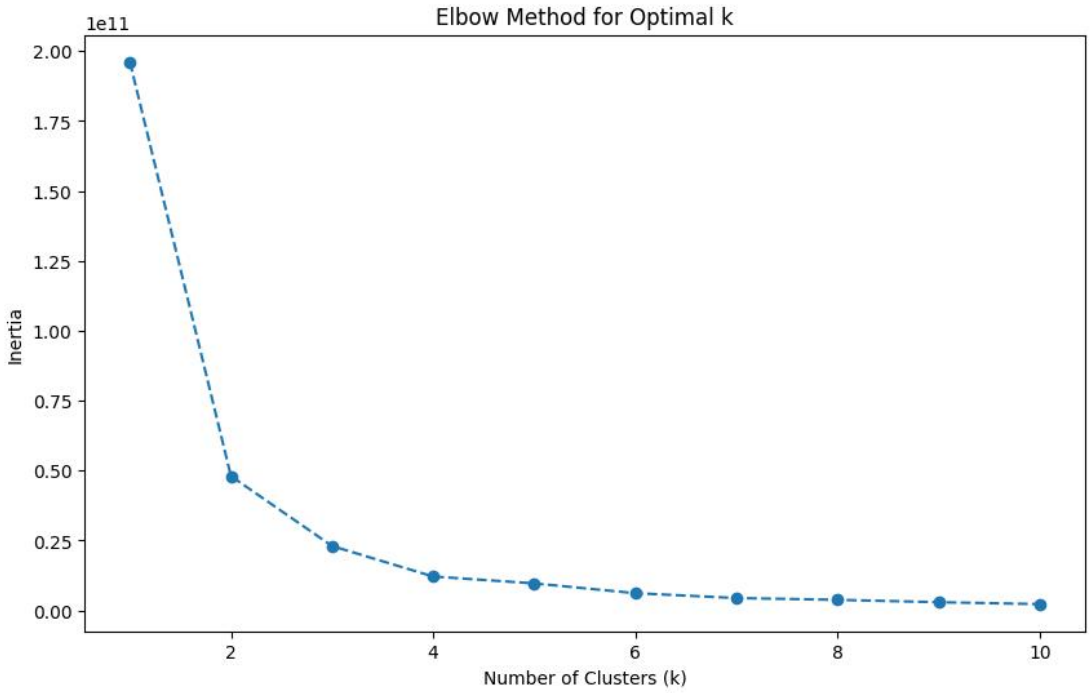

Figure 7: Elbow Method for Optimal k

- A plot of k vs. inertia reveals the "elbow point," where adding more clusters results in diminishing returns. This helps determine the optimal number of clusters.

2) Clustering
   - A K-Means model with n_clusters=3 is trained on BMI and Charges.

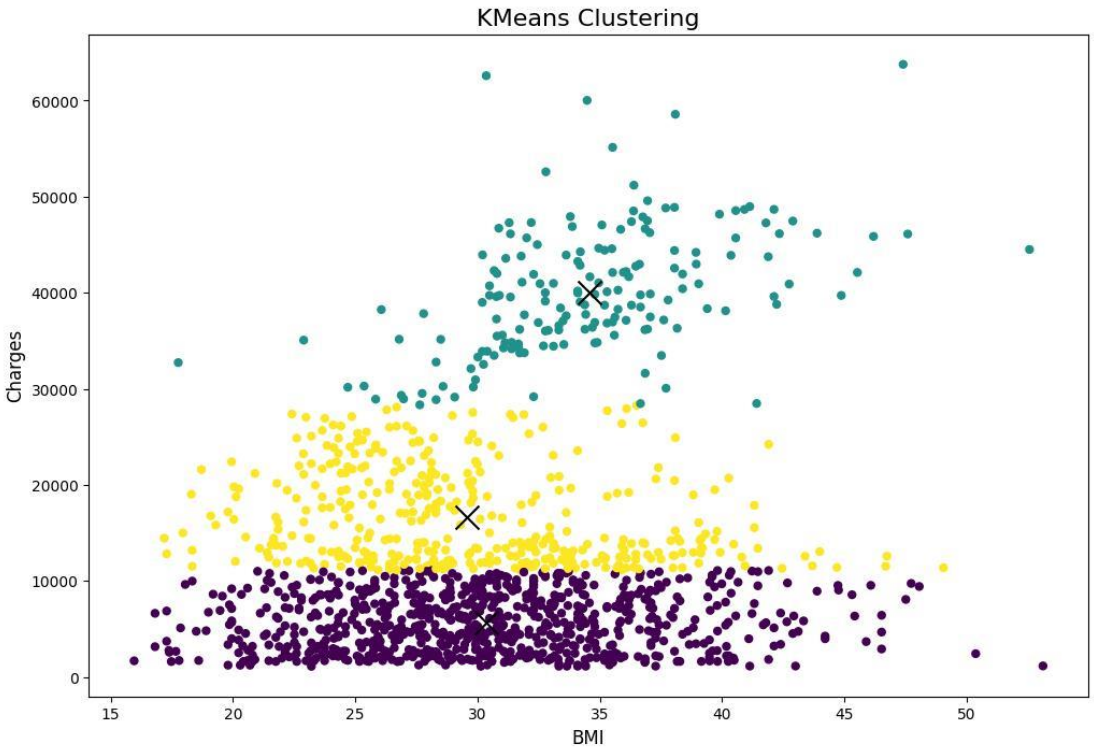

Figure 8: K-Means Clustering

- A scatter plot of BMI vs. Charges shows individuals grouped into clusters, color-coded by cluster labels. Cluster centers are marked with black 'x' markers.
- Silhouette Score

  Measures how well-separated the clusters are, ranging from -1 to 1. A Silhouette Score of 0.593 indicates good clustering performance, where individuals in the same cluster are relatively similar, and clusters are distinct.

CHAPTER III: RESULTS

III. 1 Evaluation

1. Evaluating Distribution of charges

Figure 9: Patient Charges Skewness

a. Skewness Analysis

1) Initial Skewness (not_normalized):

Measures asymmetry in the distribution of charges. A right-skewed distribution indicates many low values and few high values.

2) Normalized Skewness (normalized):

Log transformation is applied using np.log() to make the distribution more symmetric, reducing skewness.

b. Visualization

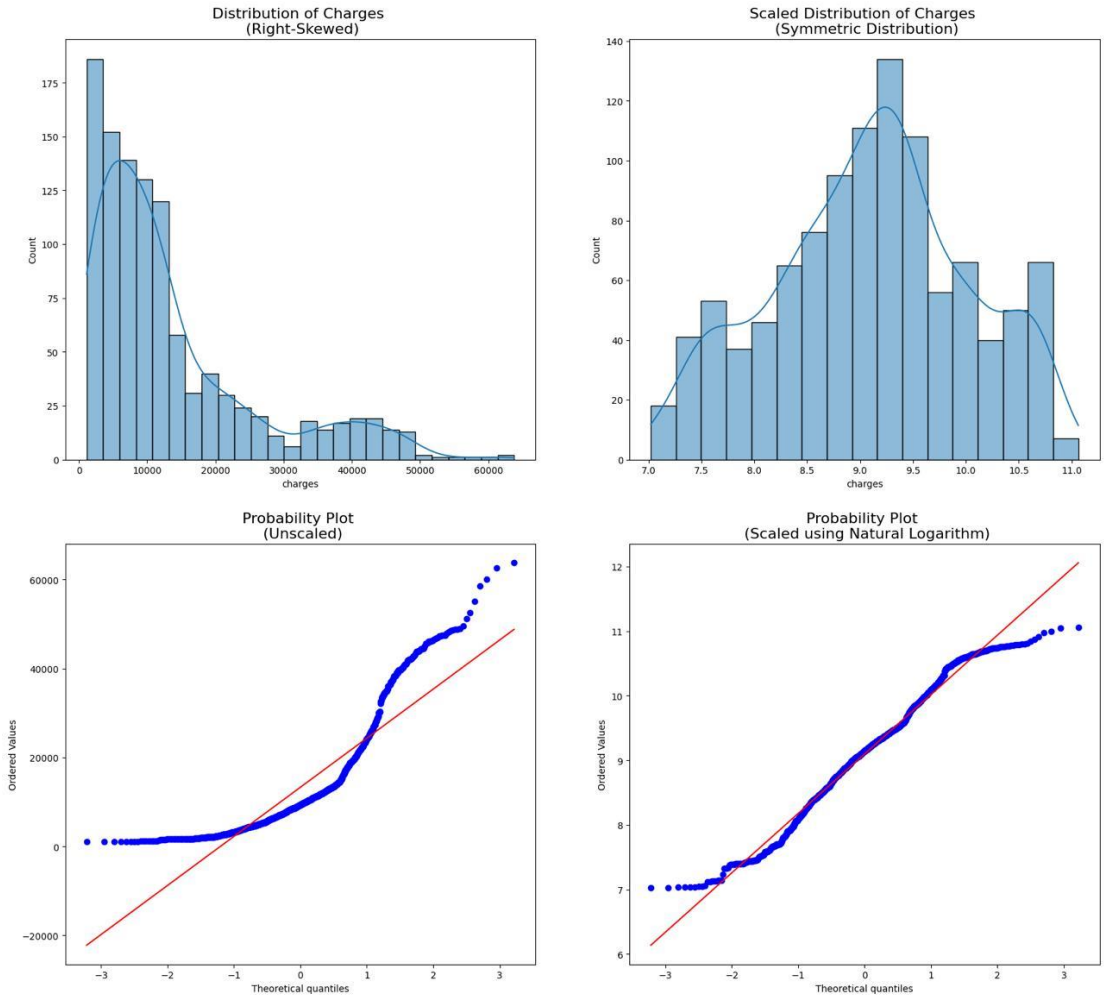

Figure 10: Distribution of Charges and Probability Plot

Four plots are used to assess the transformation effect:

1) Histogram (Unscaled):
   sns.histplot(train['charges'], kde=True) shows the original right-skewed distribution.
2) Histogram (Log-Scaled):
   sns.histplot(np.log(train['charges']), kde=True) shows the symmetric distribution after log transformation.
3) Quantile Plot (Unscaled):
   stats.probplot(train["charges"], plot=ax3) checks if unscaled charges follow a normal distribution.

4) Quantile Plot (Log-Scaled):
stats.probplot(np.log(train["charges"]), plot=ax4) evaluates normality post-transformation.

2. Logarithmic Transformation for Regression

Regression models perform better when target variables are normally distributed. Log transformation reduces the impact of outliers and improves the linear relationship between predictors and the target variable and enhances model accuracy.

Steps

a. Transform Target Variable:
y_train_log = np.log(y_train): Converts charges to its natural log scale.

b. Prepare Features:
X_train_log = sm.add_constant(scaled_xtrain): Adds an intercept term for OLS regression.

c. Fit the Model:
   1) model_log = sm.OLS(y_train_log, X_train_log) initializes the OLS regression using log-transformed charges.
   2) results_log = model_log.fit() fits the model.

d. Evaluate the Model:
   1) results_log.summary() provides model statistics.
   2) **R-squared: Improved from 0.759 to 0.764**, indicating better fit.

III. 2 Deployment

This deployment process involves saving a trained regression model, creating a web interface for user interaction, and enabling insurance charge predictions.

1. Save the Trained Model

To reuse the trained model without retraining it every time.

a. The joblib.dump function saves the trained model_log as a file named "log_regression_model.pkl".
b. The saved file can be loaded using joblib.load, enabling predictions on new data.

2. Create a Web Application with Streamlit

   Streamlit is used to build a user-friendly interface for the model:

   a. User Input: The interface collects user-provided values for features such as age, bmi, children, sex, smoker, and region.
   b. Preprocessing: A preprocess_input_data function ensures user inputs are transformed to match the format used during model training (e.g., scaling and encoding).
   c. Prediction: The preprocessed data is passed to the model, which outputs the predicted insurance charges.
   d. Flow:
      - User inputs values via the interface.
      - Upon clicking "Predict", inputs are preprocessed.
      - The model predicts charges.
      - Results are displayed, formatted in USD.

3. Application Deployment

   The app is defined in a script (app.py), which integrates the model and Streamlit.

III. 3 Interface

Figure 11: Usage Example 1

Figure 12: Usage Example 2

The interface in the image is a sophisticated web-based tool designed to estimate medical expenses using an Ordinary Least Squares (OLS) regression model.

1. Interface Overview
    a. Input Fields
        1) Age: An interactive slider allowing users to specify the individual's age within a range of 18 to 80 years, capturing the relationship between age and healthcare costs.
        2) BMI (Body Mass Index): A slider for precise BMI input, spanning values from 15 to 40, reflecting the influence of body weight and height on medical expenses.
        3) Number of Dependents: A slider enabling users to define the number of dependents (0–5), accounting for familial healthcare needs.
        4) Gender: A dropdown menu for selecting the individual's gender (Male or Female), recognizing potential differences in healthcare utilization.
        5) Smoker: A dropdown to indicate smoking status (Yes or No), a critical factor known to significantly impact medical costs.
        6) Region: A dropdown for selecting the geographical region (e.g., Northwest), addressing regional cost disparities.
    b. Predict Button

        The "Predict Medical Charges" button initiates the regression model, processing user inputs to compute a precise prediction of healthcare expenses. The backend model applies pre-trained coefficients for each feature to deliver an accurate estimate.
    c. Output Display

        The predicted medical charges are dynamically rendered beneath the button, providing an immediate and clear result derived from the OLS regression model's computations.

(Please note that this interface utilizes an earlier version of our regression model, as the newer model is currently experiencing technical difficulties.)

III. 4 Business Implications

The ability to accurately predict medical costs has significant business implications, especially in industries like healthcare and insurance, where cost management and resource optimization are critical.

1. Precision in Pricing and Risk Assessment

   With a clear understanding of the key factors influencing medical costs—such as age, BMI, smoking status, and region—insurers can offer personalized and fair premium pricing. This ensures that:

   a. High-risk individuals (e.g., smokers or those with high BMI) are charged appropriate premiums based on their predicted healthcare needs.
   b. Low-risk individuals benefit from more competitive pricing, enhancing customer satisfaction and loyalty. This precision helps mitigate financial risks while maintaining profitability.

2. Strategic Customer Segmentation

   By clustering individuals with similar medical cost patterns, businesses can:

   a. Develop targeted wellness programs and preventive care strategies aimed at specific groups.
   b. Personalize healthcare plans to better meet customer needs, ensuring relevance and increasing customer retention. For example, a group with low predicted costs might receive incentives for maintaining healthy lifestyles, while higher-cost groups could be offered proactive care packages.

3. Data-Driven Cost Management

Predicting medical expenses transforms raw healthcare data into actionable insights, allowing businesses to:

a. Identify the most significant cost drivers and address them with targeted policies or initiatives.
b. Reduce overall healthcare expenses by investing in preventive measures for high-cost groups, ultimately lowering claims and improving financial efficiency.

CHAPTER IV: CONCLUSION AND RECOMMENDATIONS

IV. 1 Conclusion

This study demonstrates the effective use of data mining techniques, particularly regression analysis and clustering, to predict and analyze medical costs. By identifying key cost predictors such as age, BMI, smoking status, and region, the study provides actionable insights into the factors driving healthcare expenses. Clustering techniques, on the other hand, enable segmentation of individuals into meaningful groups based on cost patterns, paving the way for targeted interventions and personalized strategies.

The integration of predictive modeling with segmentation into a user-friendly deployment platform showcases the practical utility of these techniques, offering a scalable solution for real-world applications. The CRISP-DM methodology proved invaluable in systematically guiding the project from business understanding to deployment, ensuring alignment with the overarching goal of reducing healthcare costs through data-driven insights.

IV. 2 Recommendations

1. Broaden the Scope of Analysis
    a. Extending the dataset to include variables like socioeconomic status, lifestyle behaviors, and genetic predispositions, which could improve the accuracy and generalizability of the models.
    b. Consider advanced machine learning techniques (e.g., XGBoost) to capture complex relationships that linear regression may miss.
2. Enhance the Deployment Platform
    a. Develop dashboards to visualize cost drivers and clustering results for stakeholders, improving decision-making and communication.

b. Incorporate live data inputs for real-time cost predictions and scenario simulations (e.g., "How will quitting smoking affect premiums?").

By addressing these recommendations, the project can evolve into an ethical solution for tackling the rising costs of healthcare while enhancing the decision-making capabilities of insurers, healthcare providers, and policymakers.

REFERENCES

Anderson, R., Booth, A. orcid.org/0000-0003-4808-3880, Eastwood, A. et al. (11 more authors) (2021) Synthesis for health services and policy : case studies in the scoping of reviews. Health Services and Delivery Research, 9 (15). pp. 1-84.

Guo, C., & Chen, J. (2019). Big data analytics in healthcare: Data-driven methods for typical treatment pattern mining. *Journal of Systems Science and System Engineering*, 28(6), 694–714.

Elisabeth Thomas, S.N. Kumar, Divya Midhunchakkaravarthy. Effect of Segmentation of White Matter, Grey Matter and CSF in the Prediction of Neurological Disorders. International Journal of Neurological Nursing. 2024; 10(2): 9–14p.

Puka K, Buckley C, Mulia N, Lasserre AM, Rehm J, Probst C. Educational Attainment and Lifestyle Risk Factors Associated With All-Cause Mortality in the US. *JAMA Health Forum*. 2022;3(4):e220401. doi:10.1001/jamahealthforum.2022.0401

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *DaimlerChrysler AG*. Retrieved from https://www.daimler.com