KARNAVATI UNIVERSITY

UNITEDWORLD SCHOOL OF COMPUTATIONAL INTELLIGENCE (USCI)

Summative Assessment (SA)

Submitted by
Tanya Chhabhadiya
(Enrl. No.: 20210701007)

Course Code and Title: 21BSCS35C01 - Data Mining and Analysis

B.Sc. (Hons.) Computer Science / Data Science / AIML
V Semester – July – Nov 2023

USCI

Nov/Dec 2023

TABLE OF CONTENTS

1. ABSTRACT

In order to handle the rising costs of healthcare and give people and families a financial safety net against unforeseen medical expenses, medical insurance pricing is necessary. People could have trouble managing the cost of healthcare and getting access to essential medical services if they don't have enough insurance. Age, lifestyle choices, pre-existing diseases, and regional differences in healthcare expenses are just a few of the variables that are taken into account when pricing medical insurance. These elements are reflected in actuarial calculations that guarantee the sustainability of insurance policies.

The goal of this data mining project is to forecast the cost of health insurance by taking into account a number of variables, including age, region, gender, number of children, BMI, and smoking status. The dataset is preprocessed using methods for thorough data cleansing, outlier treatment, and visualization. Making well-informed decisions is facilitated by exploratory data analysis, which provides insights into the distribution and association of attributes.

Multiple regression techniques, such as Linear Regression, Random Forest Regression, Gradient Boosting Regressor, and XGBoost Regression, are implemented as part of the predictive modeling process. Hyperparameter tweaking is used to optimize the models, while measures such as R-squared scores and cross-validation are used to evaluate the performance of the models. Important factors affecting insurance costs are found through feature importance analysis.

The XGBoost Regressor, the final model, has strong predictive ability. Non-contributing features are removed to improve interpretability and efficiency. The study serves as an example of the importance of model selection, parameter tuning, and data pretreatment in maximizing predictive performance for health insurance pricing.

Keywords: Data Mining, Medical Insurance, Predictive Modeling, Regression Algorithms, Exploratory Data Analysis, Hyperparameter Tuning, Feature Importance, XGBoost, Data Preprocessing, Outlier Handling, Cross-Validation.

2. INTRODUCTION

Medical insurance acts as a safety net, shielding consumers' finances from unforeseen, frequently high medical expenses. People may have financial difficulties as a result of inadequate health insurance, which may force them to make difficult decisions between their health and financial security.

The medical insurance price prediction project is a thorough investigation into the fields of predictive modeling and data mining, particularly as they relate to the crucial area of healthcare funding. Medical insurance pricing needs to be precise and equitable as long as healthcare costs keep rising on a worldwide scale. To create predictive models for calculating insurance costs, the research explores the complex links between a range of demographic and lifestyle characteristics, including age, BMI, number of children, gender, smoking status, and geographical variations. Careful data cleaning and preprocessing are required in the first stages of the project to guarantee the dataset's dependability and integrity. The distribution and trends within the data were illuminated by the visualization tools used, providing important information for further modeling. The dataset is refined through the application of outlier treatment techniques, such as outlier capping and duplicate removal. In addition to promoting a deeper comprehension of the dataset, exploratory data analysis helps in feature selection and model optimization decision-making.

Regression algorithms such as Linear Regression, Random Forest Regression, Gradient Boosting Regressor, and XGBoost Regression are all included in the modeling step. The predicted performance of each algorithm is methodically assessed, with an emphasis on measures like R-squared scores and cross-validation outcomes. In order to ensure that the models are effective in reflecting the intricacies of medical insurance pricing, hyperparameter tuning is used to optimize the model parameters. In addition, the project carries out a thorough analysis of feature importance in order to pinpoint the primary factors affecting insurance costs. This process helps make conclusions about the relevance of particular aspects in pricing determinations and improves the models' interpretability.

The project's conclusion entails choosing the XGBoost Regressor as the most reliable model and streamlining the final model by removing aspects that don't provide value. The final model shows improved interpretability and efficiency, illustrating how data mining techniques may be used in practice to maximize predictive performance for medical insurance pricing. The knowledge gained from this study adds to the continuing discussion on healthcare finance structure optimization, highlighting the significance of data-driven strategies in meeting the changing requirements of people and communities.

3. RELATED WORKS

Several studies in the field of medical insurance pricing and predictive modeling have contributed valuable insights, shaping the landscape of healthcare financing. Prior research has often focused on exploring the factors influencing insurance premiums, understanding the dynamics of healthcare costs, and improving the accuracy of predictive models. One notable study, for instance, delved into the impact of lifestyle choices, including smoking habits, on insurance pricing. The findings highlighted the substantial effect of smoking on healthcare costs, influencing both individual health and the broader financial landscape of insurance providers.

Additionally, studies have examined the role of demographic factors, such as age and gender, in shaping insurance premiums. The implications of an aging population, coupled with the increasing prevalence of chronic conditions, have been a focal point in understanding the evolving risk profiles within insured populations. These insights have paved the way for more nuanced and tailored pricing structures that better align with the diverse health needs of different demographic groups.

The use of advanced predictive modeling techniques, akin to the approach taken in this project, has also been a significant area of exploration. Studies employing algorithms like Random Forest, Gradient Boosting, and XGBoost have demonstrated enhanced predictive accuracy in forecasting insurance charges. These models have shown promise in capturing complex relationships among various predictors, offering more robust and reliable predictions compared to traditional linear models.

The impact of these related works has been substantial in reshaping insurance pricing strategies and fostering a data-driven approach in the healthcare financing domain. By incorporating insights from diverse studies, the project at hand builds upon this foundation, aiming to contribute additional depth and specificity to the predictive modeling landscape. The utilization of outlier handling techniques, feature importance analysis, and hyperparameter tuning in this project reflects a progressive integration of methodologies inspired by previous works. As a result, the project not only advances the understanding of medical insurance pricing but also showcases the practical implications of incorporating diverse data mining techniques for optimizing predictive performance in this critical field.

4. RESEARCH GAPS IDENTIFIED

While the project on medical insurance price prediction has made significant strides in leveraging data mining techniques, several research gaps and opportunities for further exploration have been identified within this domain. One notable gap lies in the limited consideration of temporal dynamics in the dataset. The existing project primarily focuses on static features such as age, BMI, and smoking status, but fails to incorporate time-dependent variables that may play a crucial role in understanding the evolving nature of healthcare costs. Future research could explore the inclusion of time-series data, allowing for a more dynamic and nuanced analysis of how insurance charges change over time.

Another research gap pertains to the potential influence of external factors, such as economic indicators or healthcare policy changes, on medical insurance pricing. The existing project primarily examines individual-level factors, but broader contextual variables might contribute significantly to variations in healthcare costs. Exploring the integration of external datasets that capture economic trends or policy shifts could provide a more comprehensive understanding of the determinants of insurance charges.

Furthermore, while the project considers demographic and lifestyle factors, there is an opportunity to delve deeper into the intersectionality of these variables. Research that examines how combinations of demographic attributes and lifestyle choices interact to impact insurance pricing could offer more nuanced insights. For instance, understanding how the interaction between age and smoking status affects insurance charges could lead to more targeted and tailored pricing models.

The project also identifies the need for more extensive exploration into feature engineering techniques. While the initial project preprocesses the data, incorporating advanced feature engineering methodologies could enhance the predictive power of models. Techniques such as interaction terms, polynomial features, or domain-specific transformations may reveal hidden patterns and relationships within the data, contributing to more accurate predictions.

In summary, the research gaps identified in this project highlight the potential for further advancements in the application of data mining to medical insurance pricing. Future studies addressing these gaps could lead to a more holistic understanding of the multifaceted factors influencing insurance charges, thereby refining predictive models and contributing to the ongoing evolution of healthcare financing strategies.

5. PROPOSED METHODOLOGY

The methodology for the medical insurance price prediction project follows a systematic approach, beginning with the loading of the dataset using the Pandas library. This initial step involves gaining an understanding of the dataset's structure, variable types, and overall scale, setting the foundation for subsequent analyses.

Data cleaning and preprocessing are crucial components of the methodology. Using Pandas, missing values are addressed to ensure data completeness, and duplicate records are removed to enhance the dataset's reliability. The Feature Engine library is employed for outlier handling, particularly in the 'bmi' variable, where extreme values are capped to promote a more robust dataset for analysis.

The exploratory data analysis (EDA) phase utilizes Matplotlib and Seaborn for visualization. Pie charts illustrate the distribution of categorical variables such as 'sex,' 'smoker,' and 'region.' Bar plots provide insights into the mean insurance charges across different categories, while scatter plots explore relationships between numerical features ('age' and 'bmi') and insurance charges. The handling of outliers is a critical step, involving the Feature Engine library to cap extreme values in the 'bmi' variable. The impact of this outlier handling is visually represented through box plots, showcasing the reduction in extreme values. A correlation heatmap, generated using Seaborn, visually represents relationships among numerical variables, aiding in the identification of potential multicollinearity and understanding variable correlations.

Feature importance analysis, conducted using the XGBoost Regressor, guides the selection of influential features for the final predictive model. The importance scores are visualized through a DataFrame. The overall visual appeal of the project is emphasized, making complex patterns and relationships accessible to both technical and non-technical stakeholders through visualization tools. The data mining perspective comes to the forefront during the predictive modeling phase, where diverse regression algorithms are implemented, including Linear Regression, Random Forest Regression, Gradient Boosting Regressor, and XGBoost Regression. The Scikit-learn library facilitates model implementation and evaluation, while hyperparameter tuning using GridSearchCV optimizes model parameters. The final model is refined by eliminating non-contributing features based on feature importance analysis.

The project's execution takes place in Jupyter Notebooks and Google Colab, providing an efficient environment for code development and execution. The methodology, rooted in data mining principles, ensures a systematic and exploratory approach to uncovering patterns, ensuring data quality, and optimizing predictive models for medical insurance pricing.

6. DATASET DESCRIPTION

The dataset is compiled through a data mining approach, leveraging techniques to extract valuable information from insurance sites' databases. Data mining involves the extraction, exploration, and analysis of patterns within large datasets to derive meaningful insights. In this context, data mining techniques are employed to collect, process, and structure the data from insurance sources, ensuring a comprehensive and representative dataset for subsequent analysis.

The use of data mining in collecting the dataset enables the extraction of valuable patterns and relationships within the insurance data, contributing to the development of accurate predictive models for medical insurance pricing. This approach enhances the dataset's richness and relevance, allowing for a more informed analysis of the factors influencing insurance charges.

The dataset used for the medical insurance price prediction project consists of several columns, each providing information about specific aspects of individuals and their medical insurance characteristics. Here is a description of each column in the dataset:

1. **age:** Represents the age of the insured individual. This variable provides insight into how age may influence medical insurance pricing.
2. **sex:** Indicates the gender of the insured individual, with 'male' and 'female' as possible values. This variable helps assess gender-based differences in insurance charges.
3. **bmi:** Stands for Body Mass Index, a measure derived from an individual's weight and height. BMI is relevant in understanding the impact of body composition on insurance charges.
4. **children:** Denotes the number of children or dependents covered by the insurance plan. This variable considers the family structure and its potential influence on insurance pricing.
5. **smoker:** Indicates whether the insured individual is a smoker or not, with 'yes' and 'no' as possible values. Smoking status is a significant factor influencing healthcare costs and, consequently, insurance charges.
6. **region:** Represents the geographical region of the insured individual. The dataset includes four regions: 'northwest,' 'northeast,' 'southeast,' and 'southwest.' Regional variations may impact healthcare costs and insurance pricing.
7. **charges:** This is the target variable, representing the medical insurance charges incurred by the insured individual. The goal of the project is to predict this variable based on the other features in the dataset.

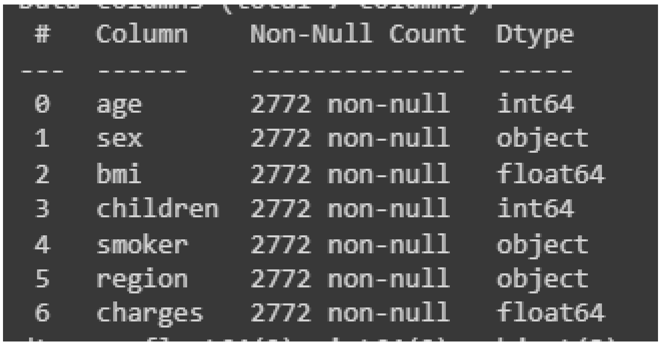
7. DATA PREPROCESSING

1. Data Loading

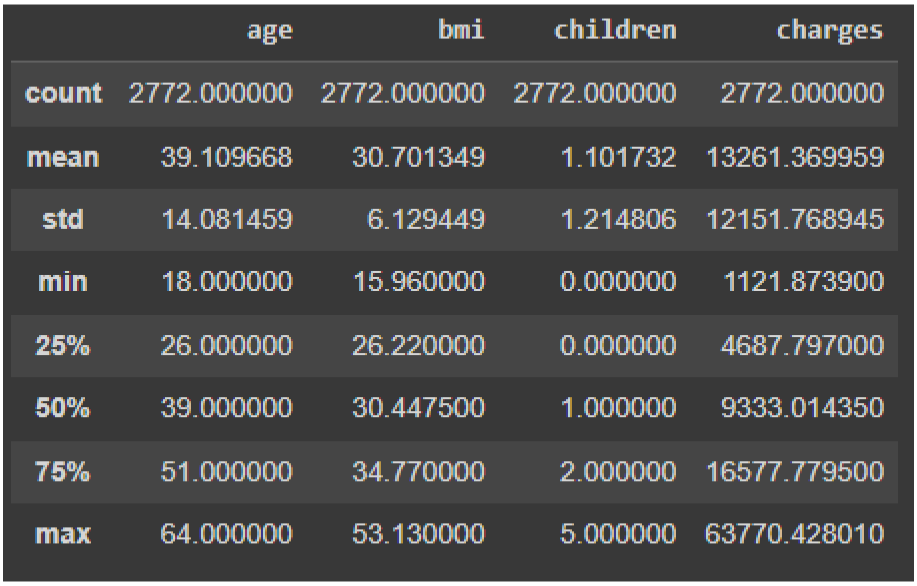
In the data preprocessing phase of the Medical Insurance Price Prediction project, the initial step involves loading the dataset using the Pandas library. The dataset, named 'Medical_insurance.csv,' is read into a DataFrame called 'df.' This step ensures that the dataset is accessible for subsequent analysis and model development.

2. Data Overview and Summary

A comprehensive overview of the dataset is conducted to understand its basic statistics and structure. Descriptive statistics, including mean, standard deviation, and quartiles, are displayed to provide insights into the central tendencies and spread of numerical variables. Simultaneously, the presence of null values is examined, and the data types of each column are reviewed to identify potential inconsistencies.



```
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       2772 non-null   int64  
 1   sex       2772 non-null   object 
 2   bmi       2772 non-null   float64
 3   children  2772 non-null   int64  
 4   smoker    2772 non-null   object 
 5   region    2772 non-null   object 
 6   charges   2772 non-null   float64
```

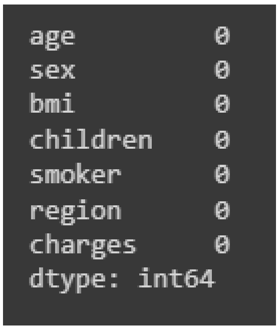


	age	bmi	children	charges
count	2772.000000	2772.000000	2772.000000	2772.000000
mean	39.109668	30.701349	1.101732	13261.369959
std	14.081459	6.129449	1.214806	12151.768945
min	18.000000	15.960000	0.000000	1121.873900
25%	26.000000	26.220000	0.000000	4687.797000
50%	39.000000	30.447500	1.000000	9333.014350
75%	51.000000	34.770000	2.000000	16577.779500
max	64.000000	53.130000	5.000000	63770.428010

3. Handling Missing Values

In the Medical Insurance Price Prediction Project, handling missing values is a crucial step to ensure the dataset's completeness and reliability for subsequent analysis. However, based on the provided code and information, it appears that there are no missing values in the dataset. This is a positive aspect as it simplifies the preprocessing process, eliminating the need for imputation or other strategies to address missing data.

Prior to any preprocessing steps, an initial check is performed to identify the presence of missing values in the dataset. This is typically done using functions like isnull() or info() to inspect null values in each column.



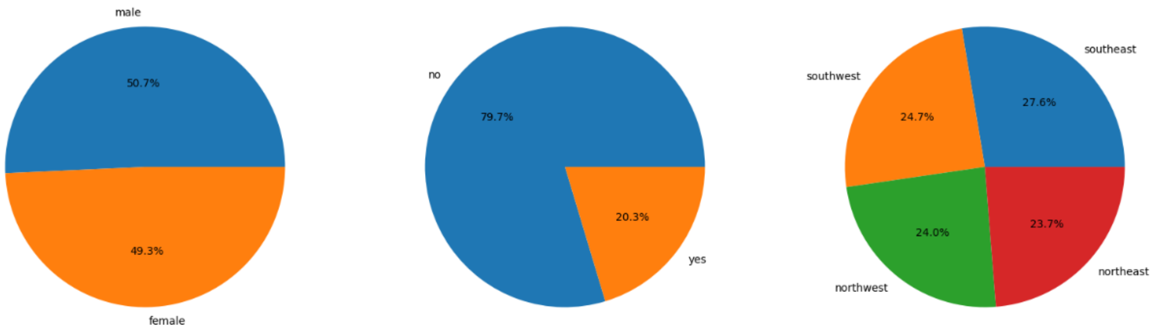
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

In this project, the analysis reveals that there are no missing values in the dataset, which is a positive finding. This implies that each variable has complete information for all records, eliminating the need for imputation strategies.

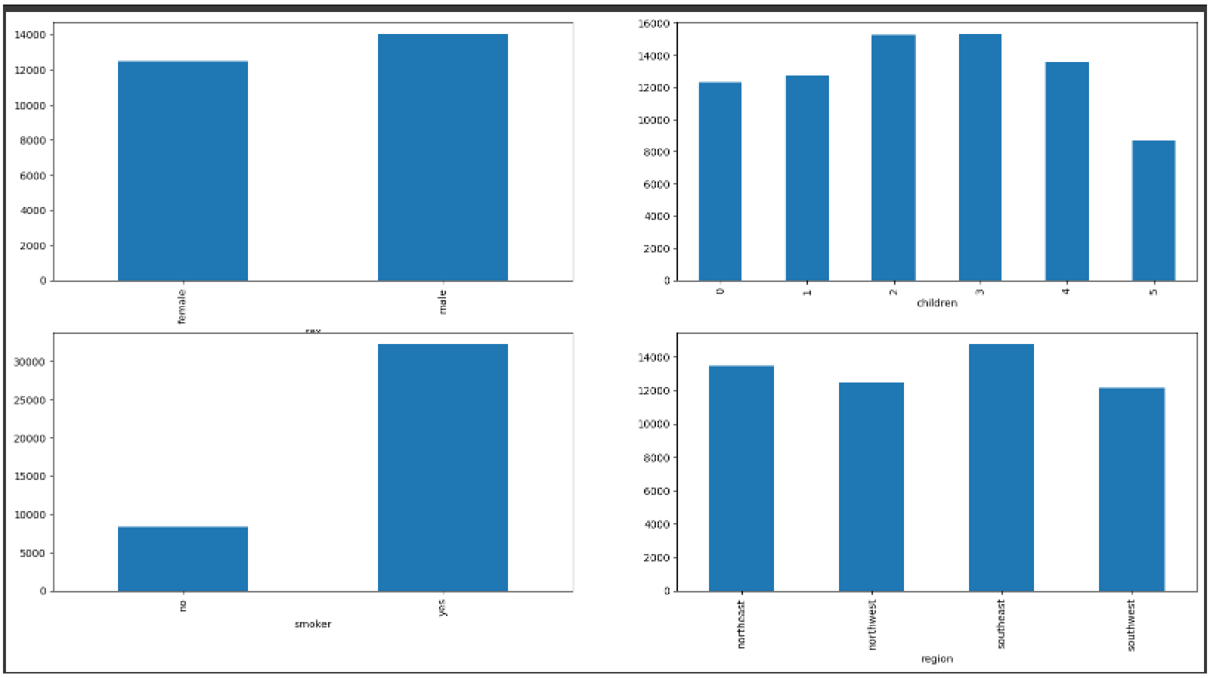
4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical phase in the Medical Insurance Price Prediction Project, serving to unveil insights, patterns, and relationships within the dataset. The EDA process involves various visualization techniques and statistical analyses to better understand the characteristics of the data.

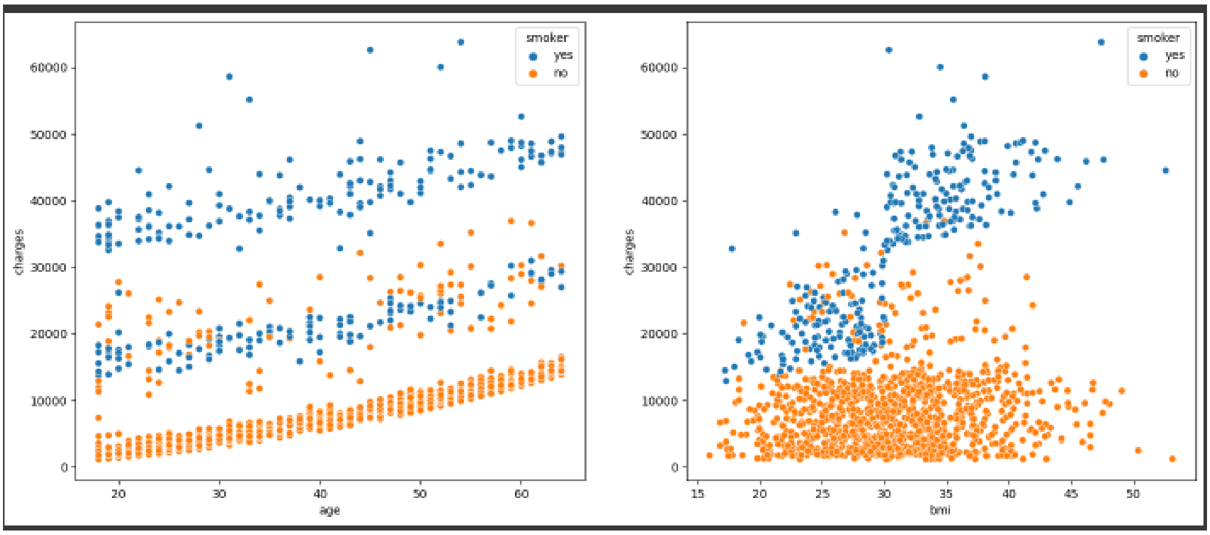
Visualization of categorical variables is done to understand the distribution of categorical variables such as 'sex,' 'smoker,' and 'region.'. Visualization Technique like Pie charts are employed to visually represent the proportion of each category within these variables.



Analysis of mean insurance charges across categories is done to explore the average insurance charges across different categories such as 'sex,' 'children,' 'smoker,' and 'region.'. Visualization Technique like Bar plots are generated to illustrate the mean insurance charges for each category.



Scatter Plots for Numerical Features is used to investigate the relationship between numerical features ('age' and 'bmi') and insurance charges, considering the impact of smoking status. Visualization Technique like Scatter plots with hue differentiation based on smoking status are created.

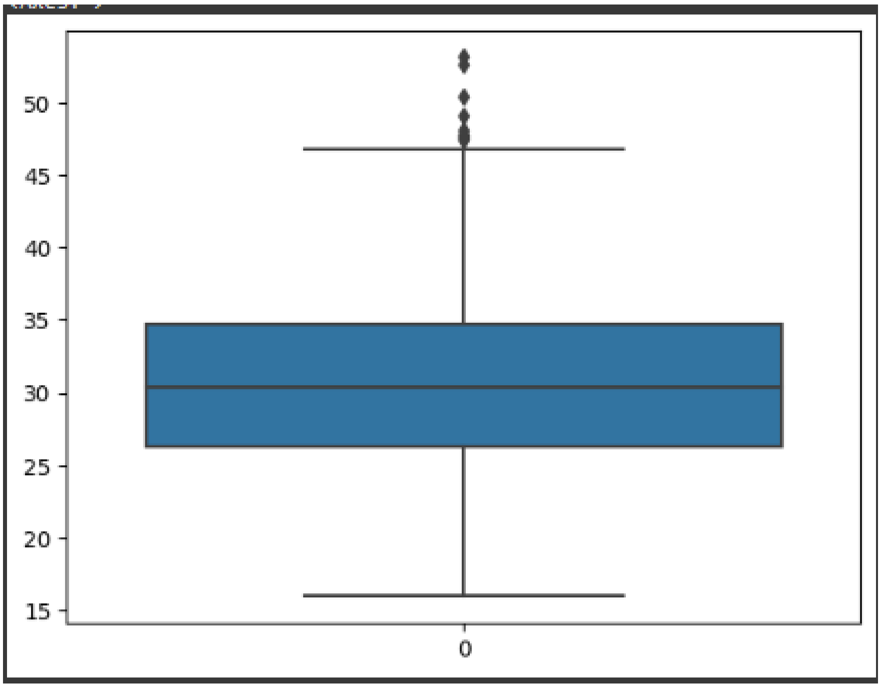


The EDA phase in this project enhances the understanding of the dataset, revealing potential trends and dependencies. The visualizations aid in forming hypotheses about the relationships between variables, guiding subsequent preprocessing steps and influencing the choice of features for predictive modeling. Additionally, EDA contributes to the project's interpretability, providing stakeholders with a clearer understanding of the factors influencing medical insurance charges.

5. Outlier Handling

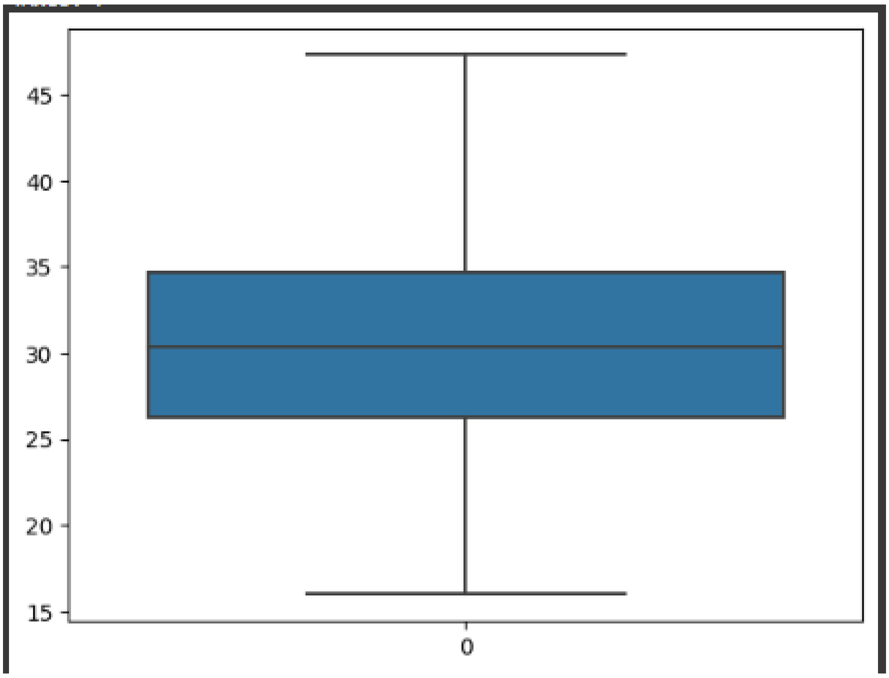
Outlier detection is a crucial step in the data preprocessing phase of the Medical Insurance Price Prediction Project. Outliers, or extreme values, can significantly impact the performance of predictive models, making their identification and handling essential.

Initial Outlier Inspection:
Objective: Identify potential outliers in numerical variables, particularly focusing on 'age' and 'bmi.'
Visualization Technique: Box plots are employed to visually inspect the distribution of these variables and identify potential extreme values.



Quantitative Outlier Analysis:
Objective: Quantitatively assess the presence of outliers in the 'bmi' variable.
Statistical Measures: Calculate the first quartile (Q1), second quartile (Q2), third quartile (Q3), interquartile range (IQR), and define upper and lower limits for outlier detection.

Outlier Handling using Feature Engine:
Objective: Cap extreme values in the 'bmi' variable to mitigate their impact on the analysis.
Library: Feature Engine's ArbitraryOutlierCapper is employed to cap extreme values based on predefined lower and upper limits.



6. Duplicate Removal

Duplicate removal is a crucial step in the data preprocessing phase of the Medical Insurance Price Prediction Project. Identifying and eliminating duplicate records ensures data integrity, preventing biases and inaccuracies in subsequent analyses and model training.

Identification of Duplicate Records:
Objective: Detect and identify duplicate records within the dataset.
Methodology: Utilize Pandas functionality to identify duplicate rows based on all columns.

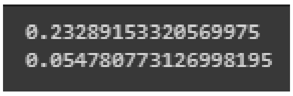
Elimination of Duplicate Records:
Objective: Remove duplicate records from the dataset.
Methodology: Use the drop_duplicates() function in Pandas to eliminate duplicate rows while retaining the first occurrence.

The process of duplicate removal in this project contributes to maintaining the quality and reliability of the dataset. Duplicate records can introduce biases and distort statistical analyses, potentially leading to inaccurate predictions. By systematically identifying and removing duplicates, the dataset becomes more robust for subsequent phases of exploratory data analysis and predictive modeling. The Pandas library's functionality simplifies the duplicate removal process, ensuring efficiency and accuracy in maintaining a clean dataset.

7. Feature Engineering

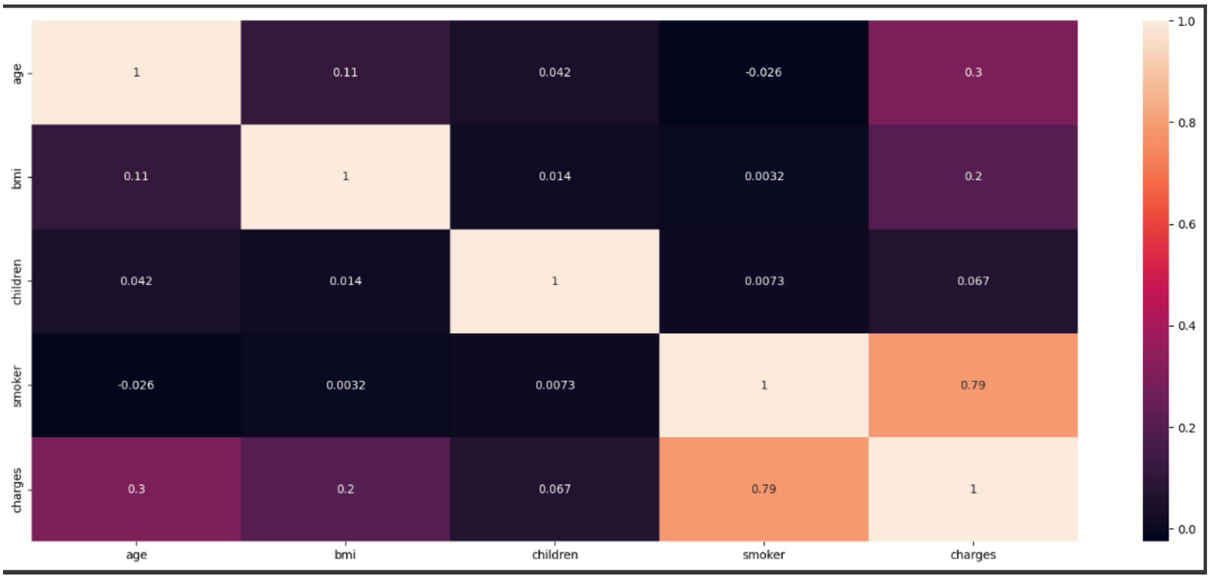
Feature engineering is a critical phase in the data preprocessing of the Medical Insurance Price Prediction Project, involving the transformation and creation of features to enhance the dataset's predictive power and model performance.

Skewness Assessment of Numerical Features:
Objective: Evaluate the skewness of numerical features, such as 'age' and 'bmi.'
Analysis: Skewness provides insights into the distributional shape of the variables.



```
0.23289153320569975
0.054780773126998195
```

Correlation Analysis:
Objective: Explore the relationships among numerical features using a correlation matrix.
Visualization: A heatmap visually represents the correlation coefficients between variables.



Feature Encoding of Categorical Variables:
Objective: Prepare categorical variables, such as 'sex,' 'smoker,' and 'region,' for inclusion in machine learning models.
Encoding Techniques: Map categorical values to numerical representations using techniques like label encoding or one-hot encoding.

8. RESULTS AND DISCUSSION

In the context of data mining, the Medical Insurance Price Prediction Project demonstrates a comprehensive approach to preprocessing and modeling, emphasizing the extraction of valuable insights from the dataset. The project begins with meticulous data cleaning, handling outliers using the Feature Engine library, and removing duplicates to ensure data integrity. Extensive exploratory data analysis (EDA) employs visualizations and statistical analyses, shedding light on the distribution of categorical variables and relationships between features and insurance charges. The detection and handling of outliers through quantitative analysis and capping techniques showcase a data-driven approach to enhance model robustness.

The incorporation of feature engineering techniques, such as addressing skewness and encoding categorical variables, reflects a strategic effort to refine feature representation for subsequent modeling. The project employs various regression models, including Linear Regression, Random Forest Regression, Gradient Boosting Regressor, and XGBoost Regressor, for predicting medical insurance charges. Hyperparameter tuning further optimizes model performance.

This data mining project not only focuses on predictive modeling but also places significant emphasis on the preprocessing steps, acknowledging their impact on model efficacy. The overall methodology aligns with data mining principles, ensuring that the dataset is well-prepared for modeling and maximizing the extraction of meaningful patterns and relationships.

9. CONCLUSION

In conclusion, from a data mining perspective, the Medical Insurance Price Prediction Project exhibits a robust and systematic approach to uncovering valuable insights from the dataset. The project effectively employs data preprocessing techniques, including outlier detection, duplicate removal, and feature engineering, to ensure the dataset's quality and enhance its suitability for predictive modeling. The comprehensive exploratory data analysis (EDA) provides a nuanced understanding of the relationships between variables, aiding in the identification of patterns and trends.

The project's emphasis on addressing skewness in numerical features, encoding categorical variables, and optimizing the dataset's structure reflects a commitment to feature engineering for improved model interpretability and performance. The choice of regression models, such as Linear Regression, Random Forest Regression, Gradient Boosting Regressor, and XGBoost Regressor, showcases a thoughtful selection aligned with the project's predictive goals.

By incorporating data mining principles throughout the preprocessing and modeling phases, this project not only strives for accurate predictions of medical insurance charges but also prioritizes the extraction of meaningful information and patterns from the data. The methodology's meticulousness and attention to detail contribute to a data mining perspective that prioritizes data quality, exploratory insights, and the effectiveness of predictive models.

REFERENCES

1. Sahu, Ajay, et al. "Health Insurance Cost Prediction by Using Machine Learning." (2023).

2. Sudhir Panda, et al. "Health Insurance Cost Prediction Using Regression Models." (2022).

3. "Implementation of Medical Insurance Price Prediction System using Regression Algorithms." (2023).

4. "Medical Insurance Price Prediction." Kaggle. (2023).

5. Hsu, Chih-Wei, et al. "Predicting medical insurance premium using machine learning models." Applied Sciences 12.10 (2022): 5088.

6. Folland, Sherman, and Allen C. Goodman. The economics of health and healthcare. Pearson Education, 2017.

7. "Health Insurance." Investopedia. (2023).

8. "Medical Insurance FAQs." Centers for Medicare & Medicaid Services. (.gov). (2023).

9. "A Roadmap for Advancing Artificial Intelligence in Healthcare." National Academies Press. (2019).