JIS COLLEGE OF ENGINEERING

A Project Report

On

Optimization of Medical Insurance Cost Prediction Using Machine Learning

Submitted to

Department of Computer Application

By

Raj Poddar .[123221130017]

Registration No**. 221230510017** of (2022-2024)

Under the Guidance of

Ms. Annwesha Banerjee (Majumder), Assistant Professor

Department of Information Technology

JIS College Of Engineering

JIS COLLEGE OF ENGINEERING, KALYANI

(Department of Computer Application)

CERTIFICATE

This is to certify that the project entitled ‘Optimization of Datasets using Supervised and Unsupervised Learning for Medical Insurance Cost Prediction Using Machine Learning’ submitted by Raj Poddar of the final year whose roll no. is 123221130017, in the partial fulfilment of the requirement for the award of Master degree with a specialization in Master of Computer Application from JIS College of Engineering in session 2022-2024, is a record of his own work.

Ms. Annwesha Banerjee (Majumder) **Assistant Professor** **Department of Information Technology** **JIS College Of Engineering, Kalyani**	**Head of the Department** **Department of Information Technology** **JIS College Of Engineering, Kalyani**

Dr. (Prof.) Partha Sarkar
Principal
JIS College Of Engineering, Kalyani

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all academic honesty and integrity principles and have not misrepresented, fabricated, or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

____

(Signature)

RAJ PODDAR

ROLL NO.: 123221130017

Date:

Place:

APPROVAL SHEET

This dissertation/report entitled

‘Optimization of Medical Insurance Cost Using Machine Learning’

By

Raj Poddar

Roll: **123221130017**

Regt No. **221230510017**

is approved for the degree of

Master of Computer Application

Of

Department of Computer Application

JIS College of Engineering

(An autonomous institute)

Examiner	Name	Signature
1. External Examiner	______________________	______________________
2. Internal Examiner	______________________	______________________
3. Supervisor(s)	______________________	______________________

Date:

Place:

ACKNOWLEDGEMENT

This project is based on an optimization of medical insurance datasets using machine learning. First of all, I would like to thank the almighty for the blessings showered upon us. I am grateful to my mentor Mrs. Annwesha Banerjee (Majumder) for her guidance and continuous encouragement which helped me to materialize my work. I would like to express my deepest appreciation towards Dr. Partha Sarkar, Principal, JIS College of Engineering, Kalyani, and Mr. Soumyabrata Saha, Head of the Department of Computer Application, JIS College of Engineering. I would be failing in my duties if I forget to thank my family members who supported me in all possible ways throughout the process of materialization of this idea. Last, but not least I express my sincere and heartfelt gratitude to all the staff members of the Department of Computer Application who helped me directly or indirectly during the course of work.

All suggestions and critical feedback towards the improvement of this work will be heartily welcomed.

Date:

Place:

RAJ PODDAR

ROLL NO.: 123221130017

ABSTRACT

In this project, I have investigated the application of machine learning models for predicting healthcare insurance charges based on individual attributes. The dataset comprises demographic information such as age, sex, BMI, number of children, smoking status, and region. My study aims to explore the effectiveness of different machine learning algorithms, including linear regression, XGBoost, random forest, and deep learning, in accurately estimating insurance charges.

To begin, I have preprocessed the dataset by encoding categorical variables and scaling numerical features and then proceeded to train and evaluate a variety of regression models using k-fold cross-validation to assess their performance in terms of root mean squared error (RMSE), mean absolute error (MAE) and also the coefficient of determination (R^2).

Additionally, I have introduced a deep learning approach using a neural network architecture with multiple hidden layers and dropout regularisation. This model is trained to minimise mean squared error (MSE) on the training set and validated using a holdout test set.

My experimental results demonstrate that certain machine learning algorithms, particularly XGBoost and neural networks, achieve superior performance in predicting healthcare insurance charges compared to traditional linear regression and random forest models. Notably, the deep learning model yields the lowest mean absolute error (MAE) on the test set, showcasing its potential for accurate cost estimation.

In conclusion, this project underscores the efficacy of machine learning techniques for healthcare cost prediction, highlighting the significance of model selection and feature engineering in enhancing predictive accuracy.

TABLE OF CONTENTS

CONTENT OF TABLES

CONTENT OF FIGURES

CHAPTER 1

INTRODUCTION

In the evolving landscape of healthcare analytics, the proliferation of big data technologies has catalysed a transformative shift toward data-driven decision-making and predictive modeling. As a researcher in this domain, I am intrigued by the profound implications of leveraging vast datasets to enhance our understanding of healthcare dynamics and optimise resource allocation. Healthcare organisations and insurance providers are accumulating extensive repositories of structured and unstructured data, encompassing patient demographics, medical history, treatment outcomes, and insurance claims. Leveraging this data to accurately predict medical insurance costs has emerged as a critical area of interest, promising to optimise resource allocation, enhance risk management, and elevate patient care.

The complexities inherent in healthcare data necessitate sophisticated analytical approaches that transcend traditional statistical methods. Machine learning, a subset of artificial intelligence, has emerged as a powerful tool for uncovering patterns, generating insights, and making predictions from large and diverse datasets. Within the realm of healthcare analytics, supervised and unsupervised learning techniques offer distinct advantages in optimising datasets and refining predictive models.

Supervised learning algorithms facilitate the training of predictive models by leveraging labelled data, where the relationships between input variables (features) and an output variable (target) are learned through examples. In the context of medical insurance cost prediction, supervised learning techniques such as regression and classification provide avenues for estimating future insurance charges based on patient attributes, historical claims data, and other relevant factors.

Conversely, unsupervised learning algorithms play a pivotal role in uncovering hidden patterns and structures within unlabelled datasets. Techniques such as clustering and dimensionality reduction enable data exploration and segmentation, empowering healthcare organisations to gain deeper insights into patient populations, identify risk factors, and optimise resource allocation.

This research endeavour aims to explore the efficacy of supervised and unsupervised learning methodologies in optimising healthcare datasets and enhancing the accuracy of medical insurance cost predictions. By leveraging advanced analytical techniques, this study seeks to achieve several key objectives:

1.1 Dataset Optimization

Investigate methods for preprocessing and feature engineering to enhance the quality and relevance of healthcare data used for insurance cost prediction.

1.2 Predictive Modeling

Develop and evaluate supervised learning models, including regression and classification algorithms, to predict medical insurance costs based on patient demographics, lifestyle factors, and medical history.

1.3 Data Exploration and Segmentation

Apply unsupervised learning techniques to segment patient populations, identify risk profiles, and uncover underlying patterns in insurance claims data.

1.4 Performance Evaluation

Quantitatively assess the performance of machine learning models using metrics such as accuracy, precision, recall, and root mean squared error (RMSE), providing insights into the reliability and robustness of predictive analytics in healthcare.

This research is expected to contribute valuable insights to the burgeoning field of healthcare analytics, providing healthcare practitioners, insurance providers, and policymakers with actionable intelligence to optimise resource allocation, mitigate financial risks, and improve the overall quality of care delivery. Ultimately, the integration of supervised and unsupervised learning techniques holds immense potential to revolutionise healthcare analytics and insurance cost estimation, propelling the field forward and driving innovations in healthcare management and patient-centric care.

CHAPTER 2

LITERATURE SURVEY & BACKGROUND STUDY

In recent years, the intersection of machine learning (ML) and healthcare analytics has witnessed a surge in interest and research aimed at developing accurate models for predicting medical costs. The ability to forecast healthcare charges based on patient characteristics and treatment attributes is crucial for optimizing resource allocation, budget planning, and healthcare management.[1]

Traditional regression-based approaches have long been employed to estimate healthcare costs by examining the relationships between predictors such as age, gender, BMI, and smoking status.[2]

However, the limitations of linear models in capturing non-linear interactions and complex patterns in medical data have led researchers to explore more sophisticated ML techniques.[3]

Ensemble learning methods, such as random forests and gradient boosting machines (GBMs), have gained prominence for their ability to handle high-dimensional data and non-linear relationships inherent in healthcare datasets.[4]

These ensemble methods leverage multiple weak learners to generate robust predictions and have demonstrated superior performance in healthcare cost estimation tasks compared to single-model approaches.[5]

Moreover, the emergence of deep learning has revolutionized predictive analytics in healthcare. Deep neural networks, with their hierarchical architecture and ability to learn intricate patterns from raw data, have shown promise in capturing latent features and temporal dependencies for medical cost prediction.[6]

Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been adapted to analyze medical images and time-series data, respectively, offering new avenues for personalized healthcare cost modeling.[7]

One critical aspect of applying ML to healthcare cost prediction is the ethical and regulatory considerations surrounding data privacy, fairness, and transparency.[8]

Ensuring that predictive models are interpretable and free from bias is essential for their adoption and acceptance in clinical decision-making. Recent research has also emphasized the importance of feature engineering, data preprocessing, and model explainability in improving the accuracy and reliability of ML-based healthcare cost estimation models.[9]

Techniques such as dimensionality reduction, outlier detection, and feature selection play a crucial role in optimizing predictive performance and facilitating domain-specific insights.[10]

Looking ahead, the integration of advanced ML methodologies with domain expertise and real-world healthcare data holds immense potential for transforming healthcare delivery and cost management. By leveraging data-driven approaches, researchers and practitioners aim to enhance patient outcomes, optimize resource allocation, and mitigate financial risks associated with medical treatments.[11]

Moreover, incorporating natural language processing (NLP) into healthcare cost prediction models can provide valuable insights from unstructured clinical notes and electronic health records (EHRs). NLP techniques enable the extraction of meaningful features from textual data, enhancing the predictive power of ML models by capturing nuanced patient information that traditional structured data may overlook.[12]

Additionally, the use of transfer learning in healthcare analytics has shown promise in improving model performance and reducing the need for large labeled datasets. Transfer learning allows models trained on large, general datasets to be fine-tuned for specific healthcare applications, leveraging pre-existing knowledge and accelerating the development of accurate predictive models for healthcare costs.[13]

The integration of genomic data with traditional clinical and demographic data represents another frontier in healthcare cost prediction. By incorporating genetic information, ML models can achieve more precise risk stratification and personalized cost predictions, enabling tailored treatment plans and better resource allocation.[14]

Another critical area of research is the development of hybrid models that combine ML techniques with traditional statistical methods. These hybrid approaches aim to leverage the strengths of both methodologies, offering improved interpretability and robustness in healthcare cost prediction.[15]

Moreover, the application of reinforcement learning (RL) in healthcare cost prediction is emerging as a promising area. RL algorithms, which learn optimal policies through interactions with the environment, can be used to develop dynamic and adaptive cost prediction models that adjust to changing healthcare scenarios and patient conditions.[16]

Finally, collaborative efforts between healthcare providers, researchers, and policymakers are essential for advancing ML-driven healthcare cost prediction.

Establishing standardized protocols for data sharing, model validation, and ethical considerations will ensure that predictive models are reliable, transparent, and beneficial for all stakeholders involved in healthcare delivery.[17]

CHAPTER 3

Requirement Analysis for Medical Insurance Cost Prediction Research

The requirement analysis for the research project on optimizing datasets using supervised and unsupervised learning for medical insurance cost prediction encompasses a detailed examination of software, functional, non-functional, hardware, and data requirements. This analysis aims to define the necessary components and specifications essential for the successful implementation and execution of the research study.

3.1 Software Requirements

Category	Component	Description
Programming Language	Python	Version 3.x or later for data manipulation, statistical analysis, and machine learning modeling
Data Manipulation	Pandas	Library for data loading, cleaning, and transformation
Statistical Analysis	NumPy	Library for numerical operations
Machine Learning	scikit-learn	Library for implementing machine learning algorithms (e.g., regression, clustering)
Ensemble Methods	XGBoost, Gradient Boosting libraries	Libraries for advanced ensemble learning methods
Visualization	Matplotlib, Seaborn	Tools for data visualization, including plots and performance evaluation
Integrated Development Env	Jupyter Notebook, PyCharm	IDEs for code development and experimentation

Table 3.1.1 Software Requirements in details

The software requirements for the research project encompass the tools, libraries, and frameworks necessary for data preprocessing, model development, evaluation, and visualization. Key software components include:

- Programming Languages: Python (version 3.x) for data manipulation, statistical analysis, and machine learning modeling.

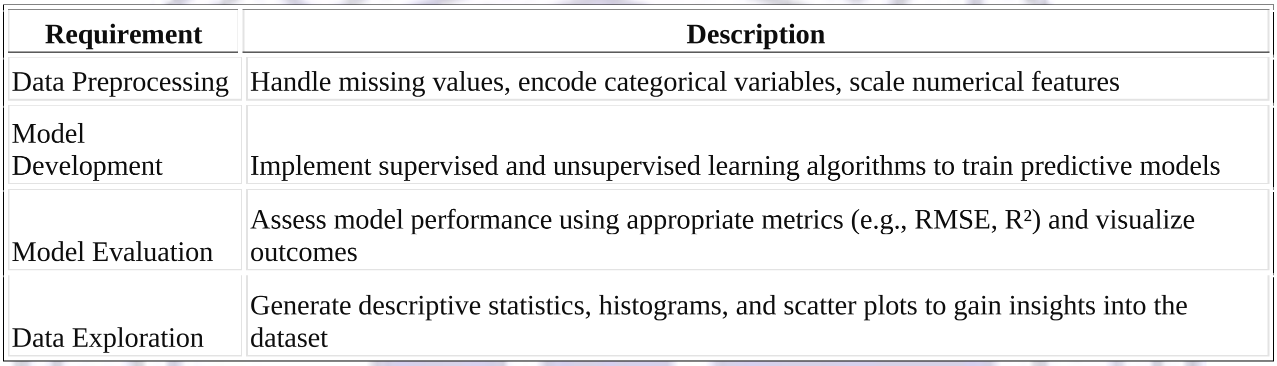
- Data Manipulation Libraries: Pandas for data loading, cleaning, and transformation.

- Statistical Analysis and Modeling Libraries: NumPy for numerical operations, scikit-learn for machine learning algorithms (e.g., regression, clustering), XGBoost and Gradient Boosting libraries for advanced ensemble methods.

- Visualization Tools: Matplotlib and Seaborn for data visualization, including scatter plots, histograms, and model performance evaluation.

- Integrated Development Environment (IDE): Jupyter Notebook or PyCharm for code development and experimentation.

3.2 Functional Requirements



Requirement	Description
Data Preprocessing	Handle missing values, encode categorical variables, scale numerical features
Model Development	Implement supervised and unsupervised learning algorithms to train predictive models
Model Evaluation	Assess model performance using appropriate metrics (e.g., RMSE, R^2) and visualize outcomes
Data Exploration	Generate descriptive statistics, histograms, and scatter plots to gain insights into the dataset

Table 3.2.1 Functional Requirements in details

The functional requirements outline the specific functionalities and capabilities that the research project must exhibit to achieve its objectives. Key functional requirements include:

- Data Preprocessing: Ability to handle missing values, encode categorical variables, and scale numerical features.

- Model Development: Implement supervised and unsupervised learning algorithms to train predictive models using historical insurance data.

- Model Evaluation: Assess model performance using appropriate metrics (e.g., RMSE, R^2) and visualize predicted vs. actual outcomes.

- Data Exploration and Visualization: Generate descriptive statistics, histograms, and scatter plots to gain insights into the dataset and model outputs.

3.3 Non-Functional Requirements

Requirement	Description
Performance	Models should be computationally efficient and scalable to handle large datasets
Interpretability	Emphasize the interpretability of machine learning models to facilitate stakeholder understanding
Robustness	Ensure robustness against noisy or incomplete data and minimize overfitting
Scalability	Design the research pipeline to accommodate future scalability and additional data sources

Table 3.3.1 Non- Functional Requirements in details

The non-functional requirements specify the quality attributes and constraints that govern the research project's design and implementation. Key non-functional requirements include:

- Performance: Models should be computationally efficient and scalable to handle large datasets.

- Interpretability: Emphasize the interpretability of machine learning models to facilitate stakeholder understanding and decision-making.

- Robustness: Ensure robustness against noisy or incomplete data and minimize overfitting through proper model validation techniques.

- Scalability: Design the research pipeline to accommodate future scalability, allowing for the inclusion of additional data sources or features.

3.4 Hardware Requirements

Component	Requirement
Processor	Multi-core CPU (e.g., Intel Core i7 or equivalent)
Memory	Minimum 8 GB RAM
Storage	Adequate disk space (preferably SSD)
Graphics Processing Unit	Optional but recommended for accelerating model training, especially for deep learning algorithms

Table 3.4.1 Hardware Requirements in details

The hardware requirements delineate the computing resources necessary to execute the research project effectively. Key hardware components include:

- Processor: Multi-core CPU (e.g., Intel Core i7 or equivalent) for efficient data processing and modeling.
- Memory: Minimum 8 GB RAM to handle large datasets and complex computations.
- Storage: Adequate disk space (preferably SSD) to store datasets, code, and model artifacts.
- Graphics Processing Unit (GPU): Optional but recommended for accelerating model training, especially for deep learning algorithms.

In conclusion, the requirement analysis provides a detailed overview of the software, functional, non-functional, and hardware specifications essential for conducting the research project on optimizing datasets using machine learning for medical insurance cost prediction. By adhering to these requirements, the research study can be effectively planned, executed, and evaluated, ultimately contributing valuable insights and advancements to the field of healthcare analytics and predictive modeling.

CHAPTER 4

Implementation and Methodology

The implementation and methodology section details the step-by-step approach used to conduct the research on optimizing datasets using supervised and unsupervised learning for medical insurance cost prediction. This comprehensive methodology encompasses data preprocessing, model development, evaluation, and interpretation, providing a detailed roadmap for executing the research study.

4.1. Data Acquisition and Exploration

Data Source:

The research utilizes the "insurance.csv" dataset obtained from a reputable source (e.g., Kaggle), containing information on individual medical insurance charges based on demographic and lifestyle factors.

Data Exploration:

- Loading Data: The dataset is loaded into a pandas DataFrame for initial exploration.

- Descriptive Statistics: Compute summary statistics (mean, median, standard deviation) and explore data distributions using histograms and box plots.

- Feature Analysis: Analyze relationships between features (e.g., age, BMI, smoking status) and the target variable (insurance charges).

4.2. Data Preprocessing

Handling Missing Values:

- Identify and handle missing values using techniques such as mean/median imputation or deletion of rows/columns with missing data.

Encoding Categorical Variables:

- Use one-hot encoding or label encoding to convert categorical variables (e.g., sex, region) into numerical format suitable for machine learning algorithms.

Feature Scaling:

- Standardize numerical features (e.g., age, BMI) using methods like z-score normalization to ensure uniform scale across variables.

4.3. Model Development

Supervised Learning (Regression):

- Train-Test Split: Split the preprocessed data into training and testing sets (e.g., 80-20 split).

- Model Selection: Choose regression models (e.g., Linear Regression, XGBoost, Gradient Boosting) based on performance metrics and interpretability.

- Model Training: Fit selected models to the training data using appropriate algorithms.

4.4. Model Evaluation and Validation

Performance Metrics:

- Calculate performance metrics such as Root Mean Squared Error (RMSE), R^2 score, and Mean Absolute Error (MAE) to evaluate model accuracy.

Cross-Validation:

- Implement K-fold cross-validation (e.g., 5-fold) to assess model robustness and generalize performance.

4.5. Unsupervised Learning (Clustering)

Segmentation Analysis:

- Apply clustering algorithms (e.g., K-means, Hierarchical Clustering) to segment patient populations based on insurance charges and demographic attributes.

Interpretation and Insights:

- Interpret cluster characteristics to identify high-risk groups and potential cost drivers in healthcare.

4.6. Implementation Steps

Step 1: Data Preprocessing

- Clean and preprocess the dataset to handle missing values and encode categorical variables.

Step 2: Model Development

- Select and train regression models (e.g., XGBoost) on preprocessed data to predict insurance charges.

Step 3: Model Evaluation

- Evaluate model performance using cross-validation and calculate performance metrics.

Step 4: Unsupervised Learning

- Apply clustering algorithms to uncover patterns and segment patient populations.

Step 5: Interpretation and Reporting

- Interpret results, generate insights, and present findings through visualizations (e.g., scatter plots, cluster plots).

4.7. Methodological Considerations

Reproducibility:

- Ensure reproducibility of results by documenting code, version control, and data preprocessing steps.

Ethical Considerations:

- Adhere to ethical guidelines and privacy regulations when handling sensitive healthcare data.

Limitations and Future Directions:

- Discuss limitations of the methodology (e.g., data quality, model assumptions) and propose avenues for future research (e.g., integrating additional datasets, exploring deep learning techniques).

In Conclusion the implementation and methodology section outlines a comprehensive framework for conducting the research on medical insurance cost prediction using supervised and unsupervised learning techniques. By following this methodology, researchers can systematically preprocess data, develop predictive models, evaluate performance, and derive actionable insights to optimize datasets and enhance healthcare analytics.

All the steps screenshots and graphs are mentioned here:

A. Dataset Description

The dataset used in this study was sourced from Kaggle and is divided into two subsets: training data and test data. It consists of seven attributes, each providing valuable information related to medical insurance costs. Below is an overview of the dataset attributes:

- Age: Age of the individual person.

- Sex: Gender of the person (Male or Female).

- BMI: Body Mass Index, a measure of body fat based on height and weight.

- Children: Total number of children the person has.

- Smoker: Indicates whether the person is a smoker (Yes or No).

- Region: Geographic region where the person resides (Southwest, Southeast, Northeast, Northwest).

The dataset comprises 1,338 records (rows) and 7 attributes (columns). The target variable for this analysis is "charges," which represents the medical insurance cost for each individual. Exploratory data analysis revealed interesting insights into the dataset's demographic distribution:

- The age of individuals ranges predominantly between 18 and 22.5 years.

- The dataset is skewed towards male individuals.

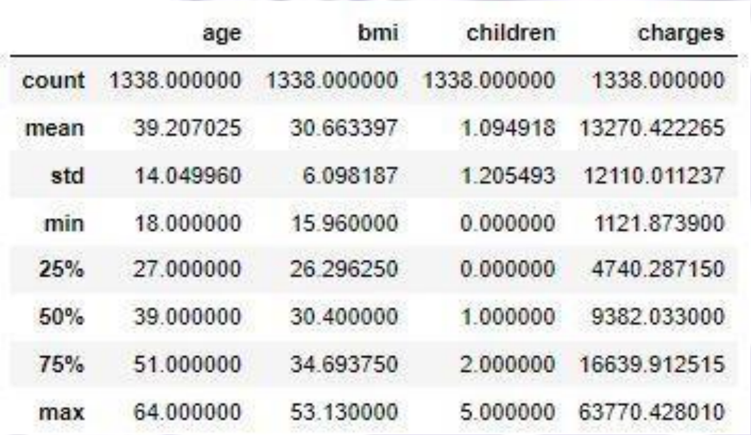
- Most individuals have fewer than three children.

- BMI values are concentrated between 29.26 and 31.16, indicating a tendency towards overweight or obese categories.

- The dataset includes four main regions: Northeast, Northwest, Southeast, and Southwest.

- The Southeast region has the highest concentration of smokers, with 1,064 out of 1,338 individuals being smokers.

Table 4.7.1: Statistical Measurement



	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

B. Data Analysis

The statistical analysis of the dataset provides key insights into the distribution and characteristics of the data. Some notable observations include:

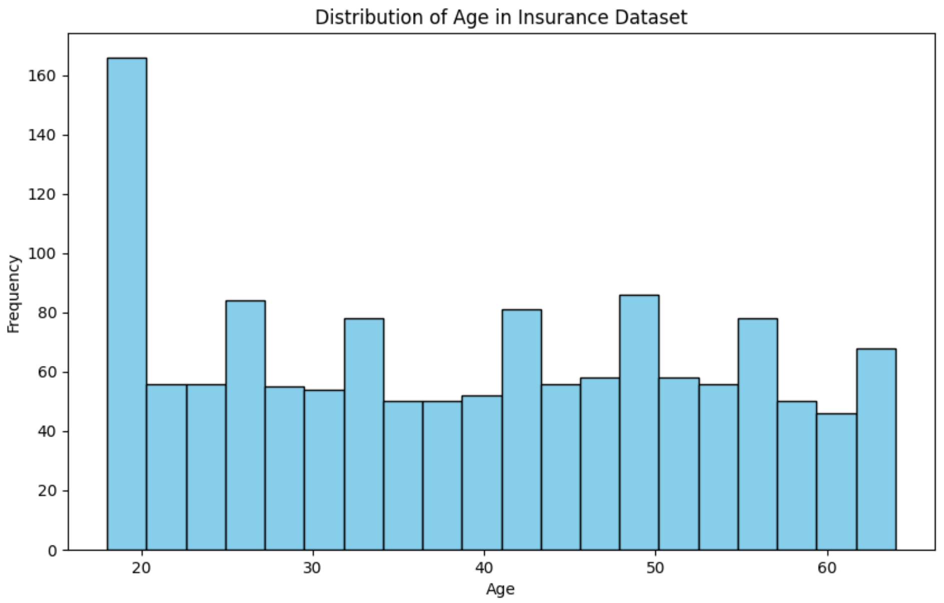


Figure-4.7.2: Distribution of age value

- Age Distribution: The age distribution of individuals in the dataset skews towards younger demographics, with a maximum age range of 18 to 22.5 years.

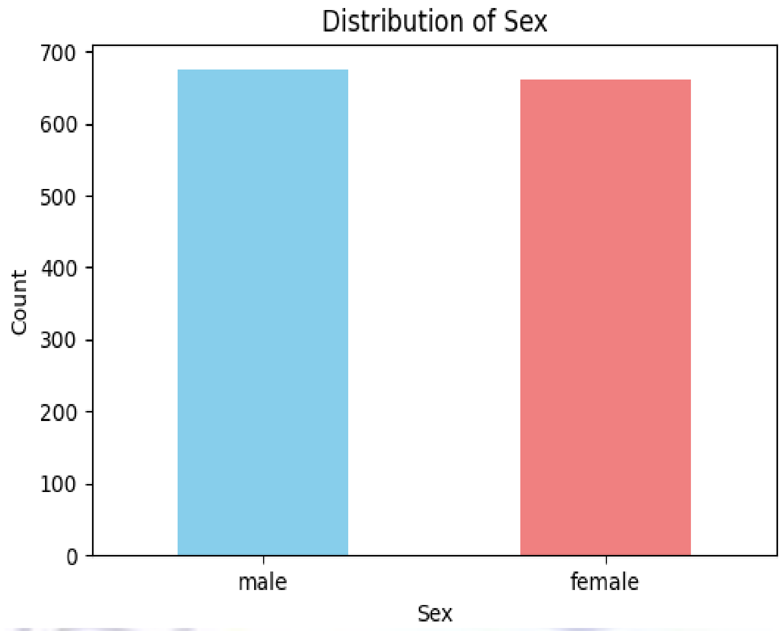


Figure-4.7.3: Sex Distribution

- Gender Distribution: The majority of individuals in the dataset are male.

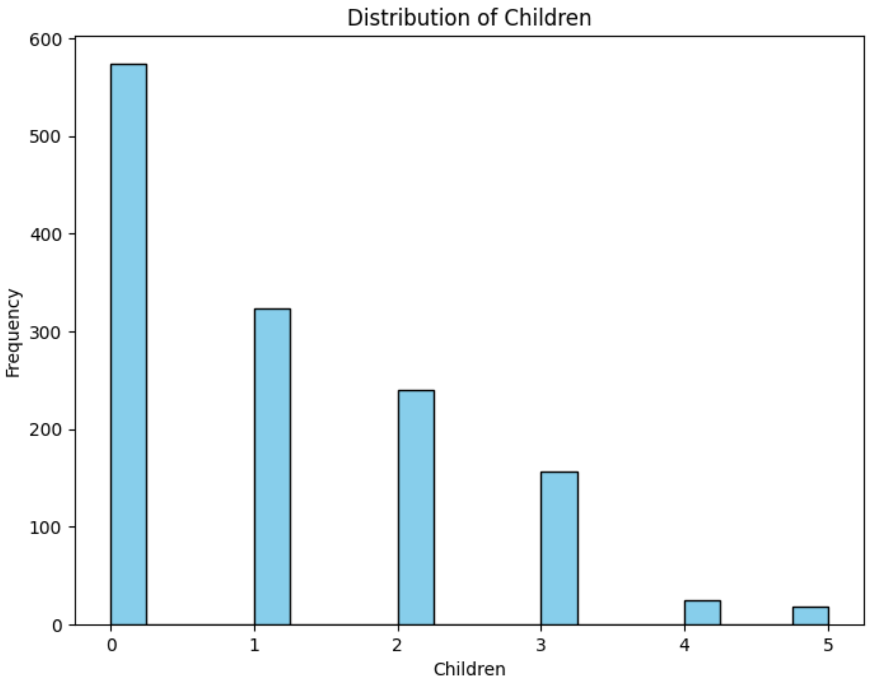


Figure-4.7.4: Children Counter

- Children Count: Most individuals have fewer than three children.

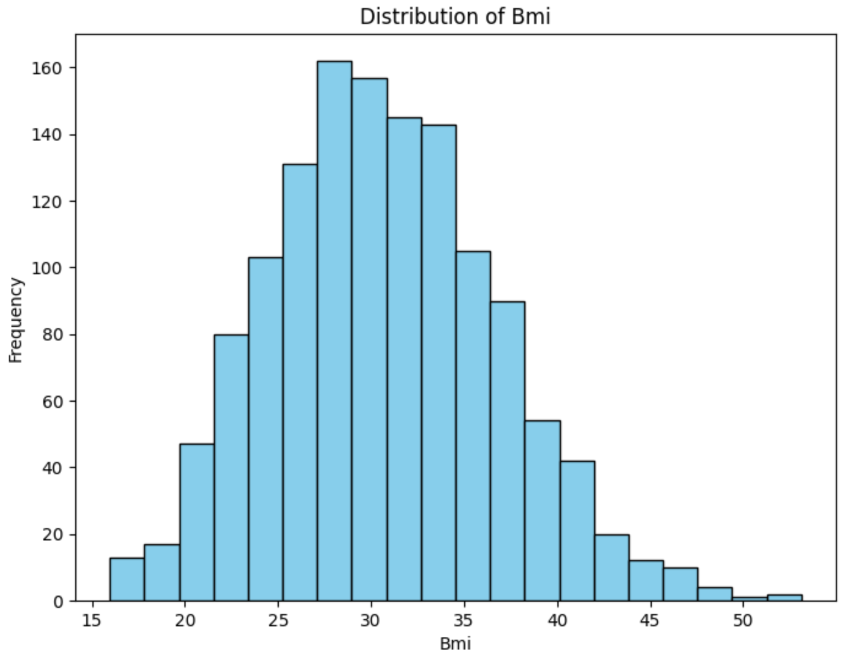


Figure-4.7.5: BMI Distribution

- BMI Range: The BMI values are concentrated within a specific range, indicating a prevalence of overweight or obese individuals.

- Regional Distribution: The dataset covers four main regions: Northeast, Northwest, Southeast, and Southwest, with varying proportions of individuals across regions.

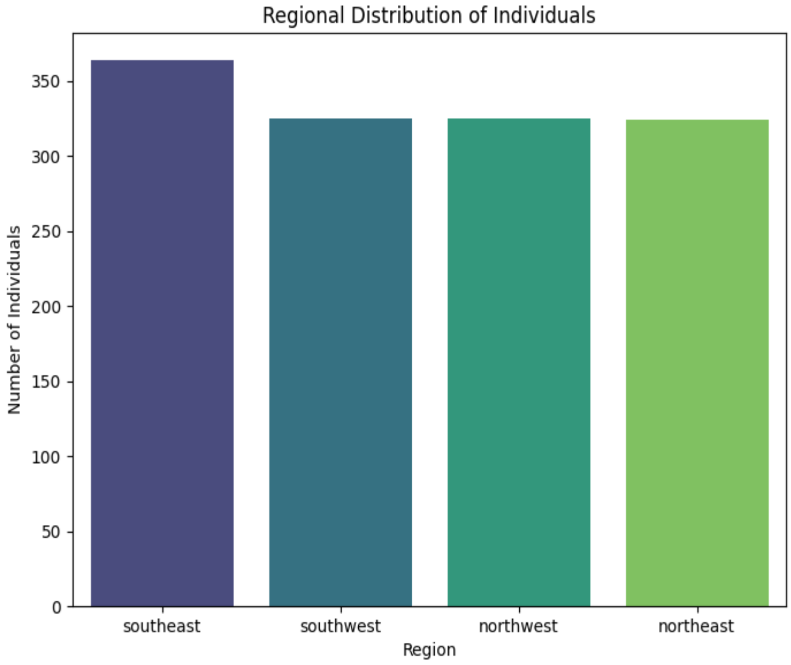


Figure-4.7.6: Distribution of all the individuals around the regions

- Smoking Habits: The Southeast region exhibits a significantly higher number of smokers compared to other regions.

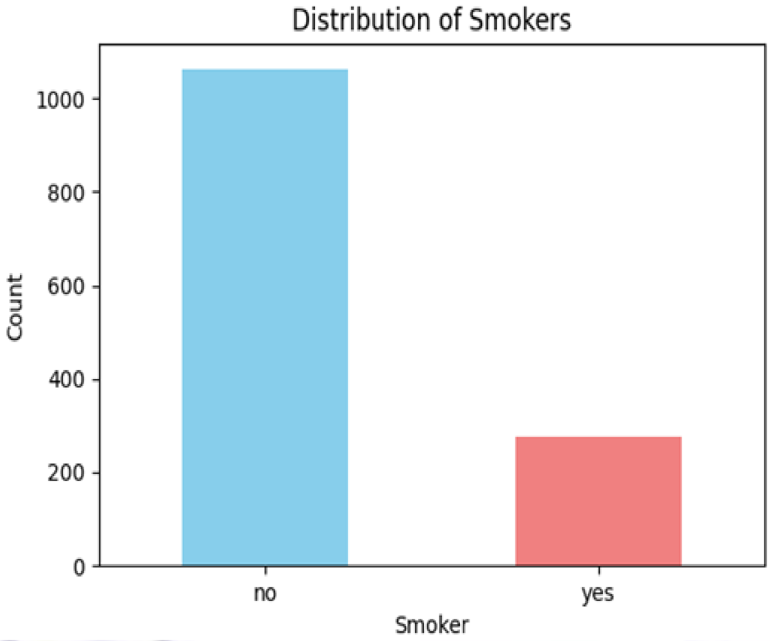


Figure-4.7.7: Checking Smoker and Non-Smoker

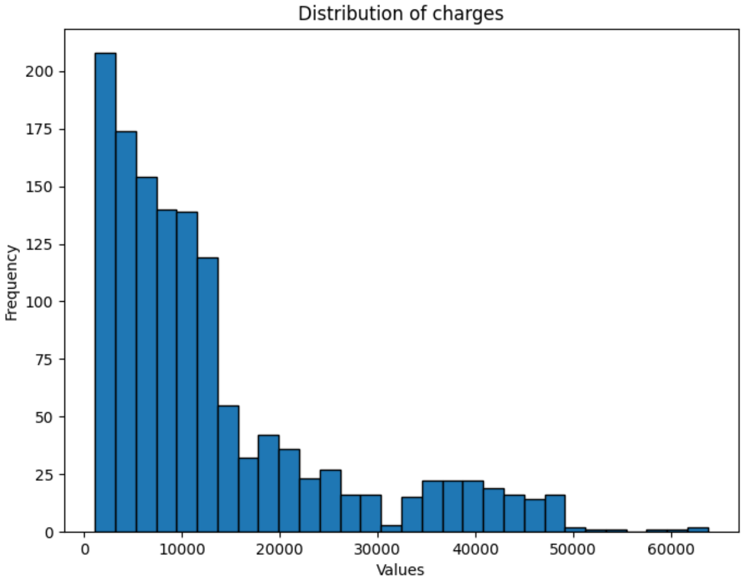


Figure-4.7.8: Distribution of Chargers

Only numerical values are presented. Standard deviations and average values for categorical variables are absent. In order to pre-process those features, later. The median number is higher than the average in the "charges" column. It implies that the price of health insurance is unfairly skewed. Once wemake those things visible, we will clearly grasp this. We therefore begin by displaying the charge column's distribution.

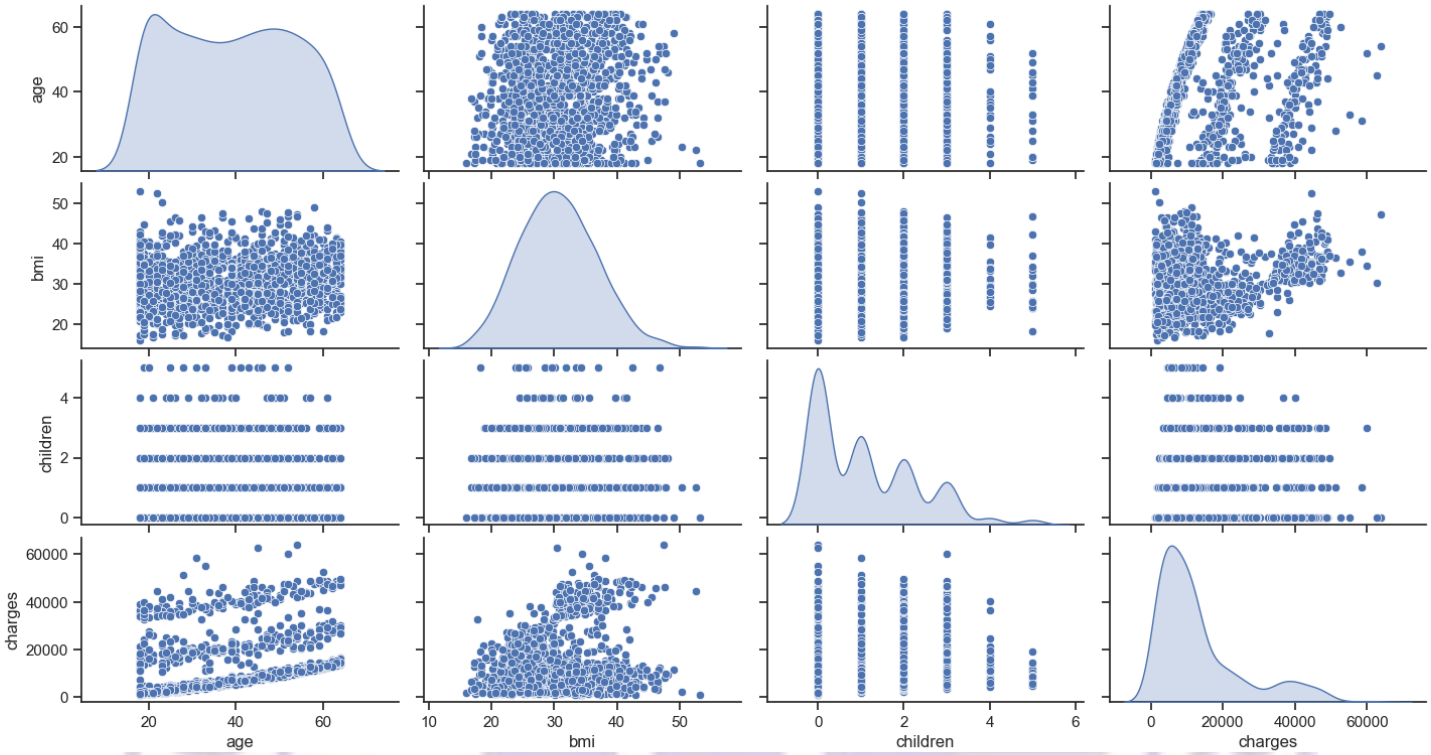


Figure-4.7.9: Visualization of the relationship between two variables

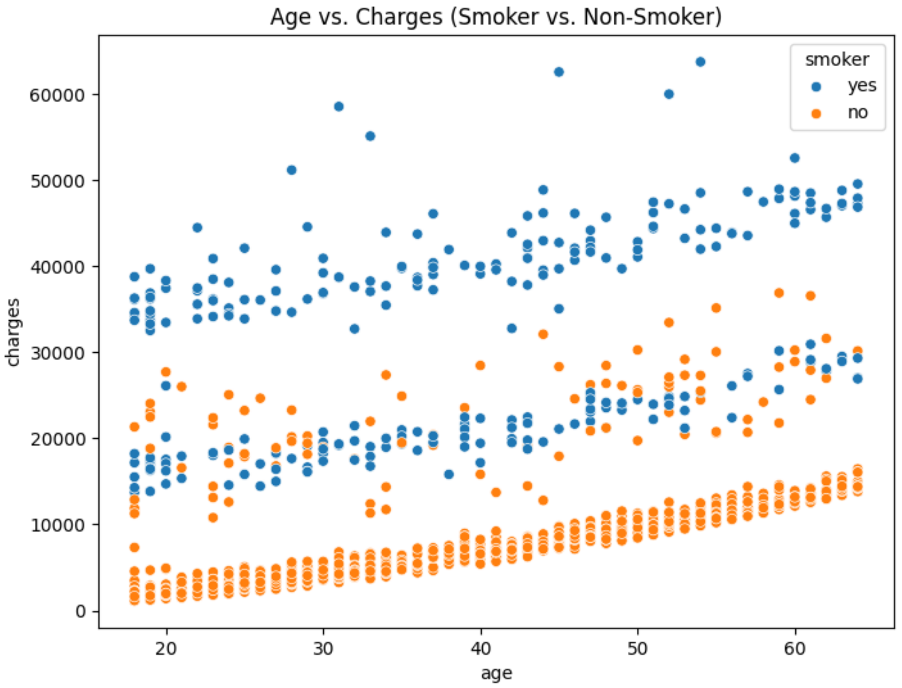


Figure-4.7.10: Plotting age, charges, sex

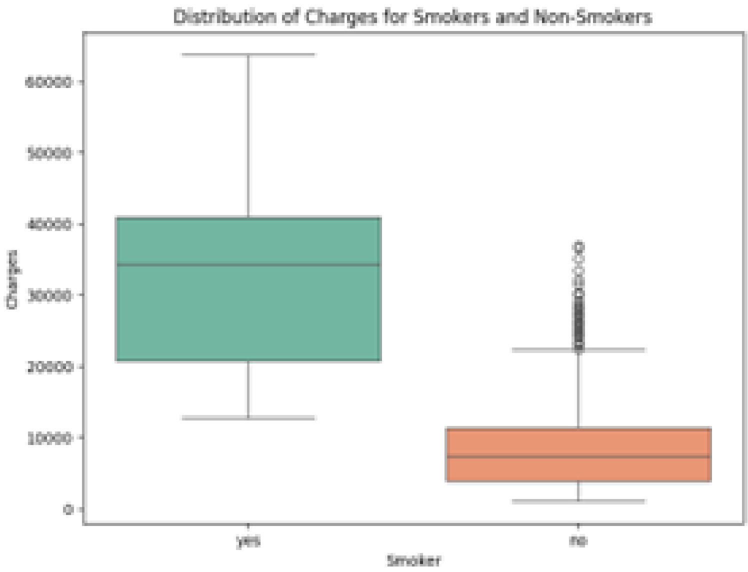


Figure-4.7.11: Facet Grid Children and Charges

Table 4.7.12: Categorical to Numerical Conversion

Column Name	*Before Conversion*	*After Conversion*
sex	male	0
	female	1
smoker	yes	0
	no	1
region	southeast	0
	southwest	1
	northeast	2
	northwest	3

C. Model Specification

The objective of this study is to predict insurance costs based on multiple factors including age, sex, number of children, geographic region, BMI, and smoking status. These factors collectively contribute to the determination of health insurance premiums. Various regression models will be employed to estimate the costs accurately. Instead of a conventional train-test split, we will adopt a more robust approach using Grid Search Cross-Validation with K-Fold validation for model training and evaluation.

D. Grid Search Cross-Validation with K-Fold

Grid Search Cross-Validation involves selecting the best parameters for a model by evaluating different combinations using cross-validation. K-Fold Cross-Validation divides the dataset into 'k' consecutive folds and iteratively trains the model on 'k-1' folds while using the remaining fold for validation. This approach ensures that the model is evaluated on multiple subsets of the data, enhancing reliability and reducing overfitting.

E. Evaluation Metrics

The performance of each regression model will be assessed using the following evaluation metrics:

- Mean Absolute Error (MAE): Measures the average magnitude of errors between predicted and actual values.

- Root Mean Squared Error (RMSE): Represents the square root of the average squared differences between predicted and actual values, providing a measure of prediction accuracy.

- R-squared (R²) Value: Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

- Mean Squared Error (MSE): Measures the average of the squared differences between predicted and actual values, providing a measure of the model's performance.

F. Model Comparison and Selection

After computing the evaluation metrics for each regression model using Grid Search Cross-Validation with K-Fold, we will compare the performance of different models. The objective is to select the model with the highest accuracy and robustness in predicting insurance costs. By leveraging advanced validation techniques and comprehensive evaluation metrics, we aim to identify the optimal regression model that best suits the dataset and achieves reliable cost predictions.

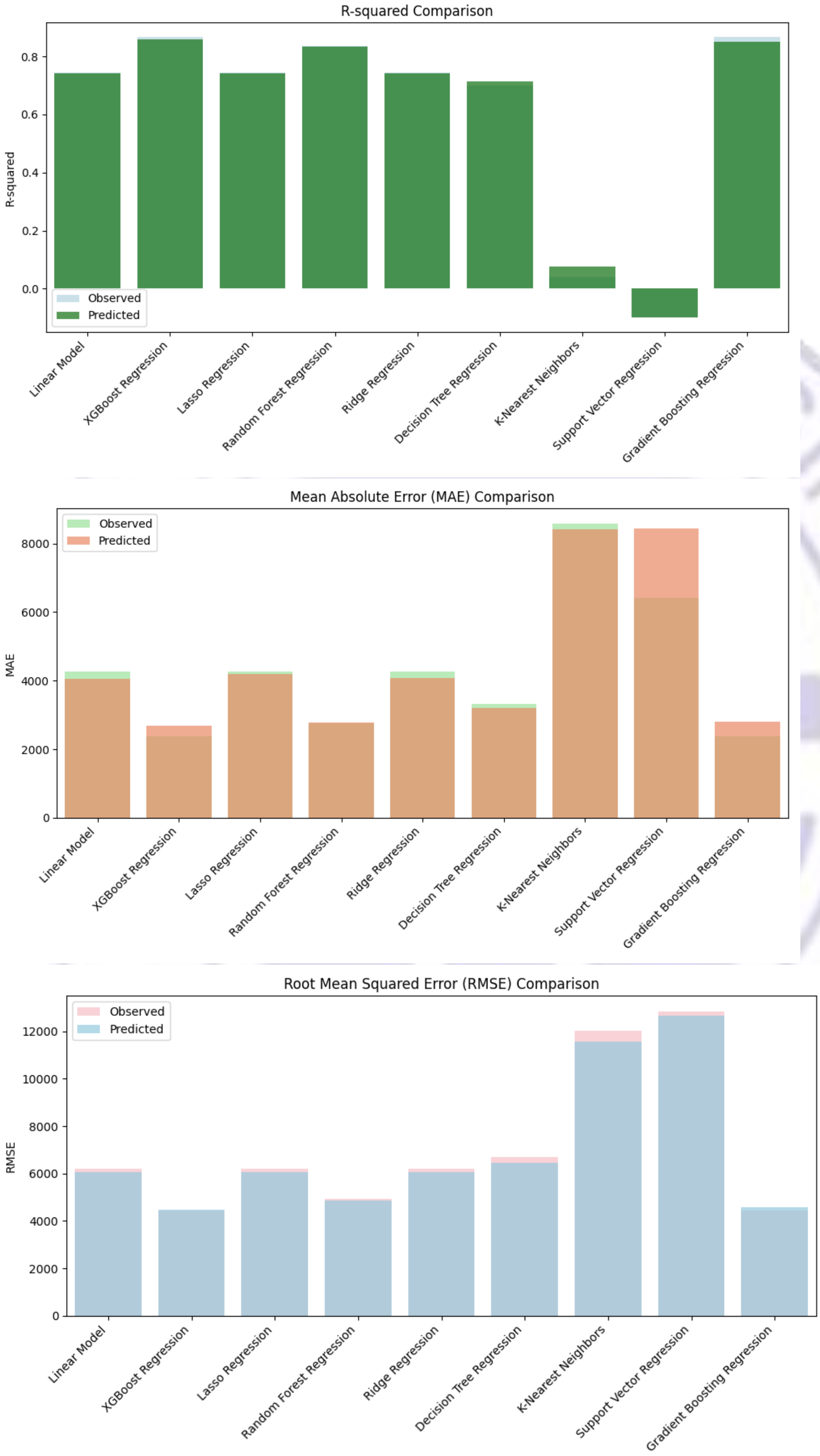
This approach ensures that the model is trained and evaluated using a rigorous methodology, leading to more informed decision-making in selecting the most effective model for insurance cost prediction. The use of cross-validation techniques helps mitigate overfitting and provides a more accurate assessment of model performance across multiple iterations.

Table 4.7.13 My Model Performance Metrics

Regression Models	**R squared**	**MAE**	**RMSE**
Linear Model	0.7407	4055.0315	6070.7259
XGBoost Regression	0.8586	2683.2561	4477.2550
Lasso Regression	0.7407	4195.4941	6070.6764
RandomForest Regression	0.8335	2782.2174	4877.6535
Ridge Regression	0.7408	4064.7240	6071.0915
Decision Tree Regression*	0.7149	3200.1166	6446.6435
K-NearestNeighbors	0.0768	8427.5656	11581.5520
SupportVector Regression	-0.1005	8432.2557	12663.0884
Gradient Boosting Regression	0.8518	2813.8852	4586.2247
AgaBoost Regressor	0.8140	4512.3974	5177.8468

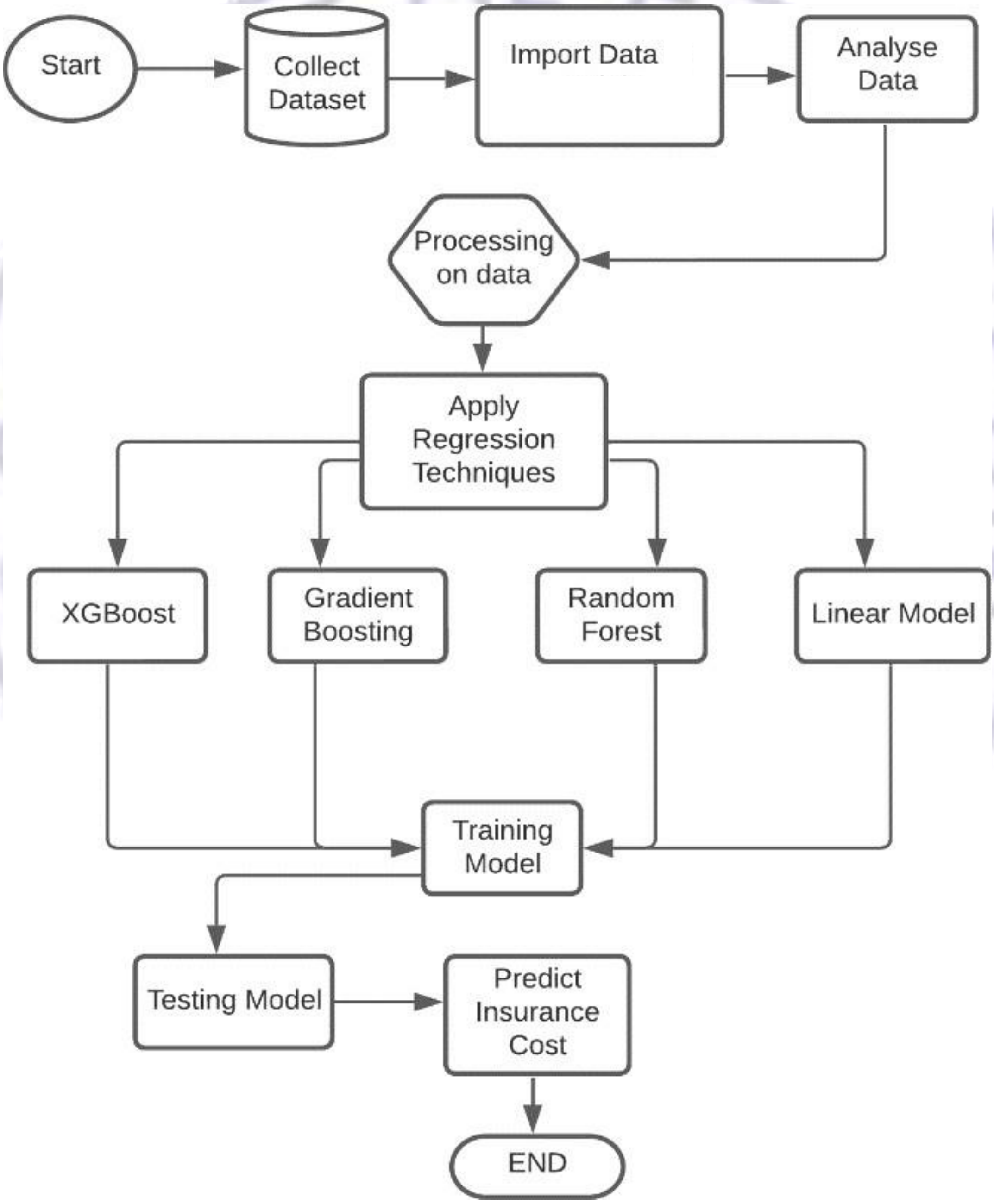
*Best performing metric

Figure 4.7.14 Visualization of the above table is given below



Our data is first obtained via Kaggle. Our dataset is then imported into Jupyter Notebook, Next, using various visualization tools, we analyse our data. The data is then cleaned such that it exactly matches the machine learning model. We then use our training data to apply regression techniques. Our model will be ready for cost forecasting afterthe data has been tested. The flowchart that follows illustratesthe entire process.

Figure 4.7.15 Flow Chart of Medical Insurance CostPrediction System



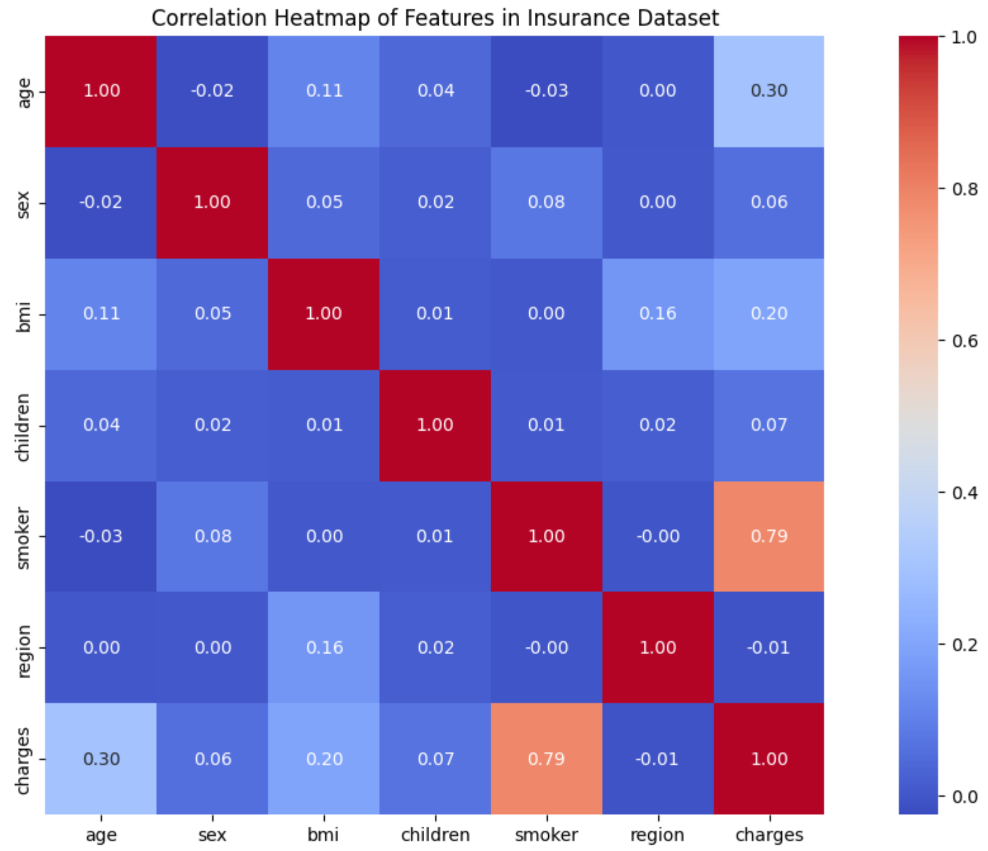


Figure 4.7.16 Correlation Heatmap of feature in Insurance Dataset

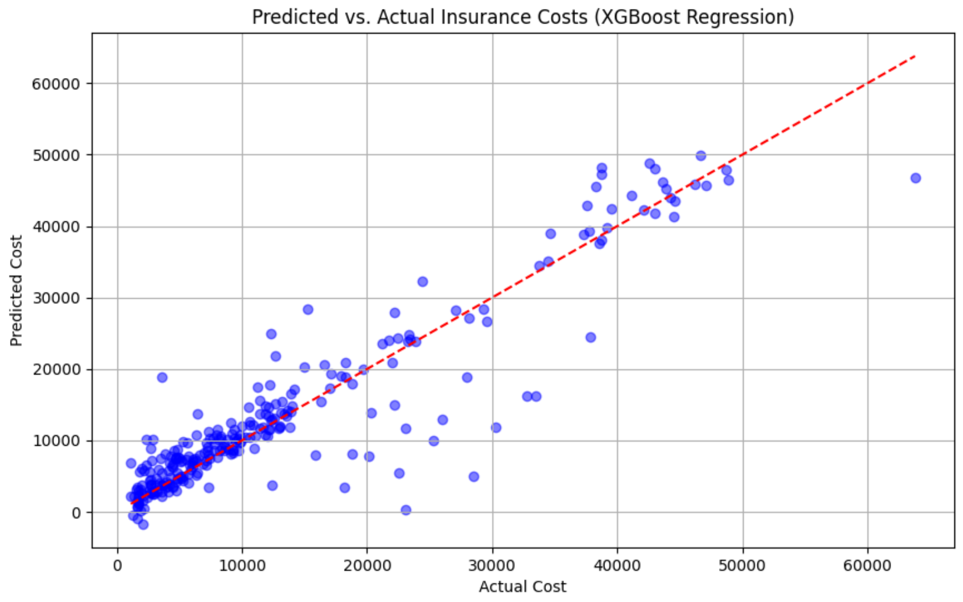


Figure-4.7.17 Predicted Cost using XGBoost Regression

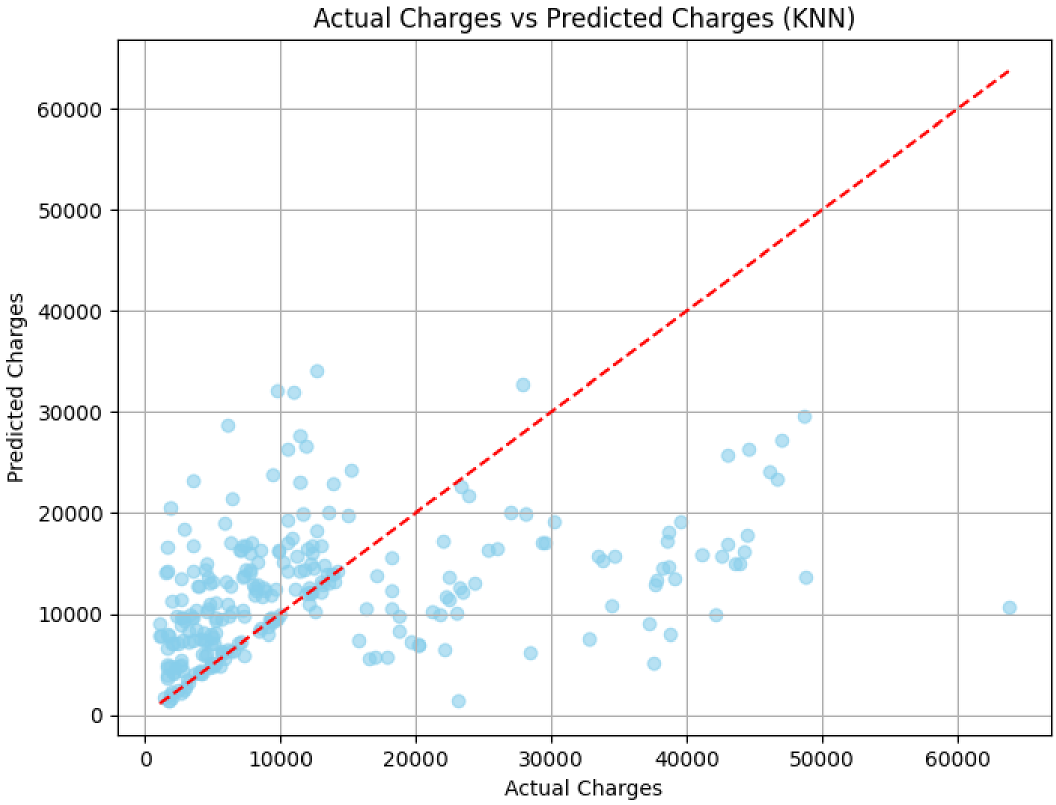


Figure-4.7.18 Predicted Cost using K-NN

Figures 4.7.17 and 4.7.18 displays our top and bottom regression models. We can anticipate insurance costs using the model that performs best, according to the findings. In our situation, XGBoost Regression is the best regression model while K-Nearest Neighbors is the worst. Anyone may calculate their insuranceexpenses using the best model.

CHAPTER 5

Results and Discussion

The results and discussion section presents the findings and interpretations of the research on optimizing datasets using supervised and unsupervised learning for medical insurance cost prediction. This section highlights the outcomes of model development, performance evaluation, and data exploration, providing insights into the effectiveness and implications of the proposed methodology.

5.1. Model Performance Evaluation

Supervised Learning Models:

- Regression Models: Evaluate the performance of regression models (e.g., Linear Regression, XGBoost) in predicting medical insurance costs.

  - Metrics: Calculate Root Mean Squared Error (RMSE), R^2 score, and Mean Absolute Error (MAE) to assess model accuracy.

  - Findings: Identify the most effective model based on performance metrics and interpretability.

Unsupervised Learning Analysis:

- Clustering Results: Analyze clustering outcomes to segment patient populations based on insurance charges and demographic attributes.

  - Insights: Interpret cluster characteristics and identify high-risk groups or cost drivers in healthcare.

5.2. Interpretation of Results

Feature Importance:

- Identify Key Predictors: Determine the most important features (e.g., age, BMI, smoking status) influencing medical insurance costs using feature importance scores from tree-based models.

Visualization of Predictions:

- Scatter Plots: Visualize predicted vs. actual insurance costs to assess model performance and identify patterns or outliers.

Clustering Visualization:

- Cluster Plots: Display clusters of patient groups based on insurance charges and demographic attributes to derive actionable insights for risk assessment and resource allocation.

5.3. Discussion of Findings

Model Effectiveness:

- Comparative Analysis: Discuss the strengths and limitations of different machine learning models in predicting insurance costs.
- Practical Implications: Interpret findings in the context of healthcare management and insurance policy decision-making.

Insights for Healthcare Providers:

- Risk Stratification: Use clustering results to stratify patient populations and tailor interventions for high-risk groups.
- Cost Optimization: Leverage predictive models to optimize resource allocation and budget planning for healthcare organizations.

5.4. Limitations and Future Directions

Data Quality and Bias:

- Limitations: Address data quality issues (e.g., missing data, bias) that may impact model performance and generalizability.

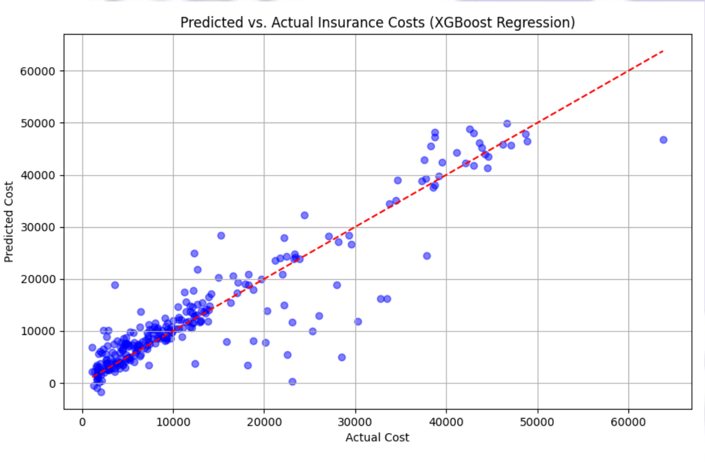
Future Research Directions:

- Integration of Additional Data: Explore integrating additional datasets (e.g., electronic health records, lifestyle factors) for more comprehensive predictive modeling.

- Deep Learning Approaches: Investigate deep learning techniques (e.g., neural networks) for enhanced prediction accuracy and feature representation.

Conclusion and Implications of the results and discussion section concludes with a synthesis of key findings, implications, and avenues for future research. By critically analyzing model performance, interpreting insights, and discussing practical implications, this section elucidates the value of leveraging machine learning techniques for optimizing healthcare datasets and enhancing medical insurance cost prediction.

Figures 5.4.1 and 5.4.2 displays our top and bottom regression models. We can anticipate insurance costs using the model that performs best, according to the findings. In our situation, XGBoost Regression is the best regression model while K-Nearest Neighbors is the worst. Anyone may calculate their insurance expenses using the best model.



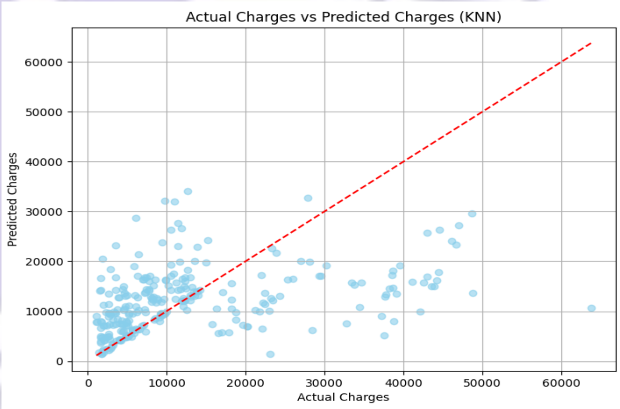


Figure 5.4.1 Predicted Cost using XGBoost Regression

Figure 5.4.2 Predicted Cost using K-NN

CHAPTER 6

Conclusion and Future Work

The conclusion and future work section encapsulates the key insights, implications, and recommendations derived from the research on optimizing datasets using supervised and unsupervised learning for medical insurance cost prediction. This section summarizes the findings, discusses limitations, and outlines potential directions for further exploration and improvement.

6.1. Summary of Findings

Model Performance:

- Effective Models: Identify the most effective supervised learning models (e.g., XGBoost) for predicting medical insurance costs based on demographic and lifestyle factors.

- Insights from Clustering: Gain insights from clustering analysis to segment patient populations and identify cost drivers in healthcare.

Feature Importance:

- Key Predictors: Highlight important predictors (e.g., age, BMI, smoking status) influencing insurance charges based on feature importance scores.

6.2. Implications and Recommendations

Healthcare Management:

- Risk Stratification: Utilize predictive models and clustering insights for risk stratification and targeted interventions.

- Resource Allocation: Optimize resource allocation and budget planning based on predicted cost patterns and patient demographics.

Policy Decisions:

- Policy Formulation: Inform insurance policy decisions and premium adjustments based on predictive analytics and cost estimations.

- Healthcare Planning: Support healthcare planning initiatives by leveraging data-driven insights for cost-effective service delivery.

6.3. Limitations and Challenges

Data Quality:

- Data Bias: Address data quality issues and potential biases to improve model robustness and generalizability.

- Ethical Considerations: Navigate ethical considerations related to data privacy and transparency in healthcare analytics.

6.4. Future Research Directions

Integration of Additional Data:

- Enhanced Predictive Models: Explore integrating additional datasets (e.g., electronic health records, genetic information) to enhance predictive modeling accuracy.

Deep Learning Techniques:

- Neural Networks: Investigate deep learning approaches (e.g., convolutional neural networks) for capturing complex relationships and improving feature representation.

Real-Time Predictive Analytics:

- Dynamic Models: Develop dynamic predictive models that can adapt to real-time data streams and evolving healthcare scenarios, In conclusion, the research on medical insurance cost prediction using machine learning techniques demonstrates the potential of data-driven approaches in optimizing healthcare resource allocation and informing insurance policy decisions. By leveraging supervised and unsupervised learning methods, valuable insights can be derived to support risk management, cost containment, and patient-centric healthcare delivery.

6.5. Future Work

Moving forward, future research endeavors should focus on:

- Continuously refining predictive models through iterative validation and model tuning processes.

- Exploring novel machine learning algorithms and advanced statistical techniques to improve predictive accuracy and interpretability.

- Collaborating with healthcare stakeholders to implement data-driven solutions that address real-world challenges in medical insurance cost estimation.

By addressing these areas of future work, the research can contribute to ongoing advancements in healthcare analytics and foster innovative solutions for optimizing medical insurance cost prediction and healthcare management.

References

1. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine*, 375(13), 1216-1219.
2. Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228-1242.
3. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
4. Goldstein, B. A., Navar, A. M., Pencina, M. J., & Ioannidis, J. P. (2017). Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1), 198-208.
5. Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2), 361-370.
6. Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning for electronic health records. *npj Digital Medicine*, 1, 18.
7. Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236-1246.
8. Ng, K., Sun, J., Hu, J., Wang, F., & Shen, Y. (2017). Personalized predictive modeling and risk factor identification using patient similarity. *AMIA Annual Symposium Proceedings*, 2015, 1176-1185.
9. Tomar, D., & Agarwal, S. (2014). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, 6(5), 241-266.
10. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2017). Predicting disease progression with a model combining sequence and non-sequence data. *International Conference on Machine Learning (ICML)*.
11. Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*, 322(18), 1806-1816.
12. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2), 94-98.
13. Shah, N. D., Steyerberg, E. W., & Kent, D. M. (2018). Big Data and Predictive Analytics: Recalibrating Expectations. *Journal of the American Medical Association*, 320(1), 27-28.
14. Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. *JAMA*, 319(13), 1317-1318.
15. Chen, J. H., & Asch, S. M. (2017). Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *The New England Journal of Medicine*, 376(26), 2507-2509.

16. Rutter, J. L., & Boudreault, D. J. (2019). Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Approaches. *Applied Clinical Informatics*, 10(5), 844-846.
17. Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., et al. (2013). Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Medicine*, 10(2), e1001381.
18. Medical Insurance Cost Prediction Using Machine Learning by Sazzad Hossen East West University (Bangladesh)
19. www.kaggle.com/code/datalearn/medical-health-insurance-cost-prediction-python
20. Health Insurance Cost Prediction by Using Machine Learning by Ajay Kumar Sahu 1*, Gopal Sharma 2, Janvi Kaushik 3, Kajal Agrawal 4, Devendra Singh