Medical Insurance Cost Prediction Using Machine Learning

A PROJECT REPORT

Submitted by

Piyush Tripathi (21SCSE1180046)

Vaibhav Singh (22SCSE1010550)

PROJECT ID: BT40113

Under the guidance of

Dr P Rajaram

in partial fulfillment for the award of the degree of

BATCHELOR OF TECHNOLOGY

IN

CSE / **CSE(AIML)**

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GALGOTIAS UNIVERSITY, GREATER NOIDA

January 2025

GALGOTIAS
UNIVERSITY

BONAFIDE CERTIFICATE

This is to certify that Project Report entitled **“Medical insurance cost prediction using machine learning”** which is submitted by **Piyush Tripathi(21SCSE1180046), Vaibhav Singh(22SCSE1010550)** in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer science & Engineering of School of Computing Science and Engineering Department of Computer Science and Engineering

Galgotias University, Greater Noida, India is a record of the candidate own work carried out by him/them under my supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Signature of Examiner(s)   Signature of Supervisor(s)

External Examiner   Signature of Program Chair

Date: January 2025

Place: Greater Noida

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

This project focuses on developing a predictive model for estimating medical insurance costs using Machine Learning (ML). The primary objective is to leverage ML techniques to create an accurate and reliable tool for forecasting insurance expenses based on key variables such as age, gender, BMI, smoking status, and geographical region. The project begins with the collection and preprocessing of a comprehensive dataset to ensure its quality and relevance. After cleaning and preparing the data, it is divided into training and testing subsets. This division is crucial for building a robust model and subsequently evaluating its performance. We employ a linear regression algorithm for this task, chosen for its simplicity and effectiveness in handling continuous output variables. During the training phase, the model is fitted to the training data, allowing it to learn the relationships between the independent variables and the dependent variable—medical insurance costs. Once the model is trained, its predictive accuracy is assessed using the testing subset. This evaluation helps ensure that the model generalizes well to new, unseen data. The resulting model is expected to provide precise insurance cost predictions, which can be invaluable for insurance companies in policy pricing and for individuals in financial planning. Ultimately, this project demonstrates the practical application of ML in the financial domain, highlighting its potential to improve decision-making processes and enhance economic efficiency through accurate predictions and data-driven insights.

CHAPTER 1
INTRODUCTION

1.1 Background and Motivation

The rising cost of healthcare is a global concern, impacting individuals, families, and businesses alike. Accurate prediction of medical insurance costs can provide valuable insights for both insurance providers and policyholders. Insurance companies can use these predictions to optimize pricing strategies, manage risk more effectively, and offer tailored plans. Individuals can benefit from understanding their potential insurance expenses for better financial planning and decision-making.

1.2 Problem Statement

The primary objective of this research is to develop a predictive model capable of accurately estimating medical insurance costs based on various demographic and health-related factors. By leveraging machine learning techniques, we aim to create a reliable tool that can assist insurance companies and individuals in making informed decisions regarding healthcare coverage.

1.3 Research Questions

1. Can machine learning algorithms effectively predict medical insurance costs based on key variables such as age, gender, BMI, smoking status, and geographical region?

2. What is the optimal machine learning algorithm for this prediction task, considering factors like accuracy, interpretability, and computational efficiency?

3. How do different feature engineering techniques impact the predictive performance of the model?

4. What are the potential limitations and challenges associated with using machine learning for medical insurance cost prediction?

1.4 Research Methodology

This study will follow a structured methodology involving the following steps:

1. Data Collection and Preprocessing: A comprehensive dataset containing relevant features (e.g., age, gender, BMI, smoking status, geographical region, medical insurance costs) will be gathered. Data cleaning and preprocessing techniques will be applied to handle missing values, outliers, and inconsistencies.

2. Feature Engineering: New features may be created, or existing features transformed to improve model performance. This might involve normalization, standardization, or creating interaction terms.

3. Model Selection and Training: Various machine learning algorithms, such as linear regression, decision trees, random forests, and neural networks, will be considered. The most suitable algorithm will be selected based on its performance on the training data.

4. Model Evaluation: The trained model will be evaluated using appropriate metrics, such as mean squared error (MSE), root mean squared error (RMSE), and R-squared. Cross-validation techniques will be employed to assess the model's generalization ability.

5. Model Deployment and Interpretation: The final model will be deployed for practical use. The model's interpretability will be analyzed to understand the factors influencing medical insurance costs and provide insights for decision-making.

1.5 Significance and Contributions

This research contributes to the field of machine learning applications in healthcare. By developing an accurate predictive model for medical insurance costs, we aim to:

- Improve decision-making: Provide valuable insights for insurance companies and individuals in pricing, risk management, and financial planning.

- Enhance efficiency: Streamline the insurance underwriting process and reduce administrative costs.

CHAPTER 2.
LITERATURE REVIEW

Medical insurance cost prediction has become a critical area of research in recent years, with various approaches being proposed to enhance the accuracy and efficiency of predictive models. This literature review explores the timeline of the problem, existing solutions, bibliometric analysis, and provides insights into the current state of medical insurance cost prediction using machine learning.

2.1 Timeline

The issue of rising medical insurance costs has been documented globally for decades, impacting healthcare systems and individual financial planning. According to CB Insights Research (2022) , the digital health industry has been increasingly focusing on solutions that integrate advanced analytics and machine learning to predict and manage healthcare expenses. Historical data and regulatory changes, particularly in South Korea, highlight the challenges of pricing and reimbursement pathways for healthcare services and insurance, emphasizing the need for accurate prediction models to stabilize pricing strategies (Lee, 2021) . This indicates that while the problem has been long-standing, the technological advancement in predictive modeling using machine learning is a relatively recent endeavor.

2.2 Existing Solutions

Several studies have explored various methodologies for predicting medical insurance costs. Early efforts, such as those presented by Gupta and Tripathi (2016) , focused on utilizing big data analytics for health insurance in India. This approach leveraged large datasets to identify cost patterns and develop insights into healthcare expenditure.

Subsequent work expanded to mobile systems and personalized recommendations, as outlined by Shakhovska et al. (2019) . Their development of a mobile system for medical recommendations aimed to integrate user-specific data for tailored insurance plans. More recent studies have shifted toward using advanced machine learning algorithms. For instance, Hanafy and Mahmoud (2021) employed deep neural network (DNN) regression models to predict health insurance costs, demonstrating improved prediction accuracy. This approach represents a significant evolution from traditional statistical models, showing the power of machine learning in capturing complex relationships among variables such as age, BMI, and medical history.

2.3 Bibliometric Analysis

1. Key Features: The literature consistently identifies demographic and health-related factors as critical for predicting medical insurance costs. Commonly used variables include age, gender, body mass index (BMI), smoking status, geographical region, and previous medical expenditures. Advanced algorithms like XGBoost and deep neural networks are frequently applied to process these features and enhance prediction accuracy.

2. Effectiveness: Machine learning models, particularly those using ensemble methods such as random forests and gradient boosting, have shown promising results in improving predictive accuracy over traditional linear models (Pesantez-Narvaez et al., 2019) . Deep learning models have also emerged as a powerful tool due to their ability to manage non-linear relationships among variables (Hanafy & Mahmoud, 2021) .

3. Drawbacks: Despite the advancements, these models face challenges. Interpretability remains a concern, as complex models like deep neural networks often function as "black boxes," making it difficult to understand the decision-making process. Additionally, there are computational efficiency issues, particularly in real-time applications, due to the high resource demands of these models (Gupta & Tripathi, 2016) .

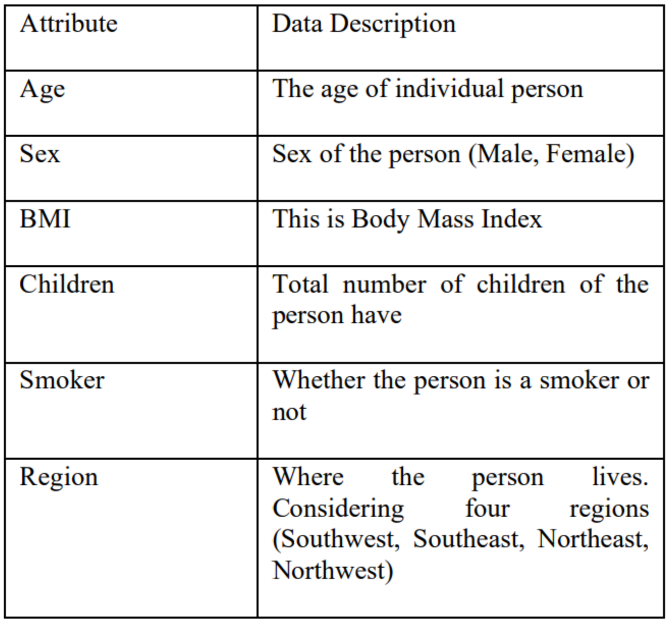
2.4 Review Summary

The findings from the literature indicate a shift from traditional statistical approaches to advanced machine learning models, which offer higher accuracy and the ability to process complex relationships. However, the issues of interpretability and computational costs highlight the need for a balanced approach. In the context of the current project, the literature supports the choice of using machine learning algorithms such as random forests, decision trees, and neural networks. The insights from existing studies guide the selection of features like age, gender, BMI, and smoking status, which are critical for building an effective predictive model.

2.5 Problem Definition

The problem at hand is to develop a machine learning model that accurately predicts medical insurance costs based on demographic and health-related features. This involves gathering a comprehensive dataset, preprocessing the data to handle inconsistencies, and selecting an appropriate model that balances accuracy with interpretability. Careful consideration of constraints, such as data privacy regulations and computational efficiency, is essential. The model should not solely prioritize accuracy at the expense of transparency, as stakeholders require understandable insights to make informed decisions.

Table 1- Overview of the Dataset



Attribute	Data Description
Age	The age of individual person
Sex	Sex of the person (Male, Female)
BMI	This is Body Mass Index
Children	Total number of children of the person have
Smoker	Whether the person is a smoker or not
Region	Where the person lives. Considering four regions (Southwest, Southeast, Northeast, Northwest)

2.6 Goals & Objectives

The main objectives of the project are:

- Data Collection: Gather a diverse dataset that includes relevant features such as age, gender, BMI, smoking status, and geographical region.

- Feature Engineering: Apply techniques like normalization, standardization, and creation of interaction terms to enhance the dataset’s predictive power.

- Model Selection: Evaluate various machine learning models, such as linear regression, decision trees, random forests, and neural networks, and select the one that best balances accuracy, interpretability, and computational efficiency.

- Model Evaluation: Assess the model's performance using metrics like Mean Squared Error (MSE) and R-squared, ensuring it generalizes well using cross-validation techniques.

- Deployment: Implement the model in a manner that allows practical application and easy interpretation, facilitating its use for decision-making in insurance pricing and risk management.

CHAPTER 3.

DESIGN FLOW & PROCESS

3.1 Evaluation & Selection of Specifications/Features

To develop an effective prediction model, it is critical to evaluate and select features based on their relevance and predictive power. Drawing from existing literature and datasets, the following features have been identified:

- Demographic Variables: Age, gender, and geographical region are consistently highlighted as strong predictors of medical insurance costs.
- Health-related Factors: BMI (Body Mass Index), smoking status, and previous medical history are crucial factors influencing insurance costs, as they directly correlate with health risks.
- Socioeconomic Factors: Income level and employment status can also provide insights into the affordability of healthcare and influence pricing strategies for insurance companies.

In this project, features will be selected based on their significance and availability in the dataset. The evaluation criteria include the strength of their correlation with medical insurance costs and their interpretability within the model.

3.2 Design Constraints

In developing the model, various design constraints must be considered to ensure ethical, professional, and regulatory compliance. The following constraints have been identified:

- Standards: The model must comply with health data privacy standards, such as HIPAA (Health Insurance Portability and Accountability Act) in the U.S. or similar regulations in other regions.
- Economic Constraints: The model should optimize computational efficiency to minimize costs associated with deployment and usage, ensuring it remains feasible for real-time applications.
- Environmental and Health Constraints: The data collected should avoid unnecessary duplication and minimize environmental impact by using cloud-based data management.
- Professional and Ethical Considerations: The model must be unbiased, ensuring that it does not discriminate against individuals based on gender, race, or other sensitive attributes.
- Social and Political Issues: The model must be transparent in its decision-making process,

considering the social implications of healthcare access and affordability.

- Cost Considerations: The model should minimize overall development and operational costs, taking into account both hardware and software requirements.

These constraints are integral to developing a model that not only performs efficiently but also aligns with legal, ethical, and economic standards.

3.3 Analysis of Features and Finalization Subject to Constraints

Based on the constraints outlined above, the features have been analyzed and finalized as follows:

- Selected Features: Age, gender, BMI, smoking status, geographical region, and medical history are retained due to their strong correlation with medical insurance costs and their ease of collection.
- Modified Features: Socioeconomic factors like income and employment status are considered optional and will only be included if they meet privacy standards and do not introduce bias.
- Removed Features: Any feature that is not strongly correlated or is challenging to collect within legal and ethical constraints (e.g., race or genetic data) will be excluded from the model.

This iterative evaluation ensures that only the most relevant and compliant features are included in the final model.

3.4 Design Flow

Two alternative design flows have been proposed for the development of the medical insurance cost prediction model:

Alternative 1: Traditional Regression-Based Flow

1. Data Collection & Cleaning: Collect relevant data and apply preprocessing techniques to handle missing values and inconsistencies.
2. Feature Engineering: Apply transformations such as normalization and standardization. Interactions between key features (e.g., age and BMI) are considered.
3. Model Training: Use regression-based models (e.g., Linear Regression) to predict medical insurance costs, considering computational efficiency and model simplicity.
4. Evaluation and Tuning: Evaluate the model using metrics like RMSE and R-squared, and apply hyperparameter tuning if necessary.
5. Deployment: Implement the model using lightweight frameworks for deployment.

Alternative 2: Advanced Machine Learning Flow

1. Data Collection & Cleaning: Collect relevant data and preprocess it to ensure high data quality.
2. Feature Engineering: Advanced techniques such as creating polynomial features or applying Principal Component Analysis (PCA) are used to reduce dimensionality.
3. Model Training: Train models using advanced algorithms like Random Forest, XGBoost, or Neural Networks. These models can capture complex relationships among features.
4. Evaluation and Tuning: Use cross-validation techniques and a combination of metrics (e.g., RMSE, MAE) to evaluate and fine-tune the model.
5. Deployment: Implement the model using cloud-based frameworks for scalability and real-time prediction capabilities.

3.5 Design Selection

Upon evaluating the two alternative designs, the Advanced Machine Learning Flow is selected as the optimal approach for this project. The reasons for this choice are as follows:

- Accuracy: Advanced algorithms like Random Forests and Neural Networks have demonstrated superior accuracy over traditional regression models in literature (Pesantez-Narvaez et al., 2019).
- Flexibility: This flow allows for better handling of non-linear relationships between features, offering more accurate predictions.
- Scalability: Cloud-based deployment ensures that the model is scalable and suitable for real-time use cases.

In contrast, the traditional regression-based flow, while simpler and less resource-intensive, may not capture complex patterns as effectively. Therefore, the advanced flow is the preferred design, given the objective of maximizing predictive accuracy while adhering to the constraints.

3.6 Implementation Methodology

The selected design flow will be implemented through a structured methodology. Below is the flowchart and algorithm for the implementation plan:

Flowchart:

1. Data Collection: Acquire the dataset containing demographic, health, and socioeconomic features.
2. Data Preprocessing: Clean the dataset by addressing missing values and normalizing

features.

3. Feature Engineering: Apply transformation techniques, create interaction terms, and possibly reduce dimensions using PCA.
4. Model Selection: Train multiple models (e.g., Random Forest, XGBoost, Neural Network) and select the one with the highest predictive performance.
5. Model Tuning: Use hyperparameter tuning methods such as grid search or random search to optimize the chosen model.
6. Evaluation: Evaluate the model using metrics like RMSE, MAE, and R-squared. Implement cross-validation techniques to ensure model robustness.
7. Deployment: Deploy the model using cloud platforms (e.g., AWS or Google Cloud) for real-time prediction.

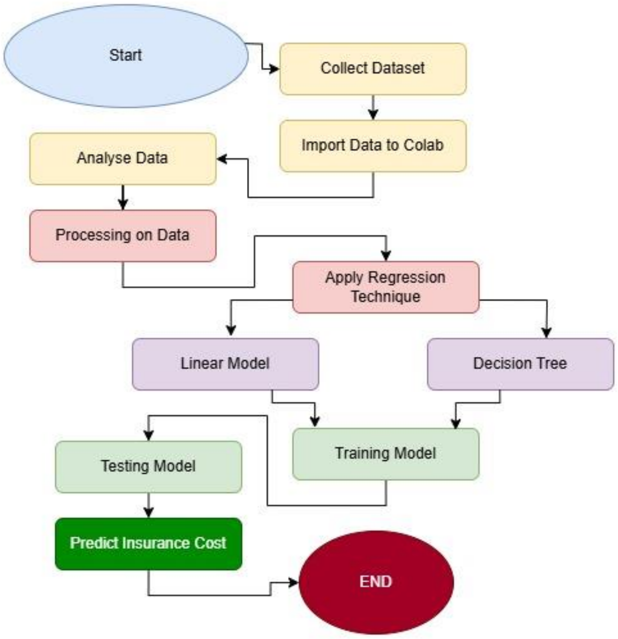


Fig 3.1 – Flowchart of medical insurance cost prediction system

Algorithm:

1. **Import Required Libraries**: Import essential Python libraries such as numpy, pandas, matplotlib. pyplot, seaborn, and sklearn. These libraries are used for data manipulation, visualization, and applying the machine learning model.
2. **Load the Dataset**: Upload the insurance dataset and load it into a Pandas DataFrame for further analysis.
3. **Explore and Understand the Data**: Display the first few rows to understand the data structure. Check the shape of the dataset (i.e., the number of rows and columns). Use .info() to inspect the data types of each column and to check for any missing values. Use .describe() to get summary statistics of the dataset. Use .isnull().sum() to check if there are any missing values in the dataset.
4. **Visualize the Data**:

   Age Distribution: Plot the distribution of the age column using a histogram to understand how ages are spread.

   Gender Distribution: Plot the count of males and females in the dataset using a bar chart.

   BMI Distribution: Plot a histogram to analyze how body mass index (BMI) values are distributed.

   Children Distribution: Plot the count of different numbers of children.

   Smoker Distribution: Plot a bar chart to show the number of smokers and non-smokers.

   Region Distribution: Plot a bar chart to analyze how many people are from each region.

   Charges Distribution: Plot the distribution of the charges column to understand how insurance charges are distributed.
5. **Data Encoding**: Convert categorical variables into numeric form Convert sex into 0 for male and 1 for female. Convert smoker into 0 for smoker and 1 for non-smoker. Convert region into numeric values: southeast = 0, southwest = 1, northeast = 2, and northwest = 3

6. **Split the Data into Features (X) and Target (Y)**: Features (X) are all columns except charges, as this is what you want to predict. Target (Y) is the charges column, which contains the insurance charges.
7. **Split the Dataset into Training and Test Sets**: Divide the data into training and test sets using an 80-20 split. This helps evaluate how well the model generalizes to new, unseen data.
8. **Train a Linear Regression Model**: Load and initialize the LinearRegression model from sklearn. Fit the model on the training data (X_train, Y_train) so that it can learn the relationships between the features and the target (insurance charges).
9. **Train a Decision Tree Regression Model**: Load the DecisionTreeRegressor model from sklearn. Train the model on the training set (X_train, Y_train).
10. **Evaluate the Models on Training Data**: Make predictions on the training data and calculate the **R-squared** value to assess how well the model fits the training data. The R-squared value indicates how much variance in the target is explained by the model.
11. **Evaluate the Models on Test Data**: Make predictions on the test data and calculate the **R-squared** value for the test data to understand how well the model generalizes to new data.
11. **Make Predictions for New Data**: Provide new input data (e.g., a person’s age, gender, BMI, number of children, smoker status, and region). Transform the input data into a format suitable for the model (as a NumPy array). Use the trained linear regression model to predict the insurance charges for the input data. Output the predicted insurance cost.

CHAPTER 4.

RESULTS ANALYSIS AND VALIDATION

4.1 Implementation of Solution

The solution implementation uses Python and Scikit-learn libraries to build and validate a Linear Regression model predicting medical insurance costs. The model was trained and evaluated using the dataset available on Kaggle. Below are the steps and results obtained.

4.2 Data Analysis and Feature Distribution

The dataset consisted of multiple features, each contributing to the analysis of medical insurance costs:

- age: The age of the individual.
- sex: Gender encoded as 0 for male and 1 for female.
- bmi: Body Mass Index, a measure of body fat based on height and weight.
- children: The number of dependents.
- smoker: Encoded as 0 for non-smoker and 1 for smoker.
- region: A categorical feature encoded as numeric values based on the geographical area.
- charges: The medical insurance cost (target variable).

Analysis and Insights

The features were analyzed using visualizations to understand their distribution and impact on insurance charges. Key findings include:

- Smokers: Individuals who smoke tended to have significantly higher insurance costs compared to non-smokers.
- BMI: Higher BMI values correlated with higher insurance costs, indicating that individuals with higher body fat have increased medical expenses.
- Age and Children: These features also influenced insurance charges, as older individuals and those with more dependents generally faced higher costs.

Visualizations such as histograms, box plots, and scatter plots were used to explore the relationships between these features and the target variable, ensuring a thorough understanding of the dataset's characteristics.

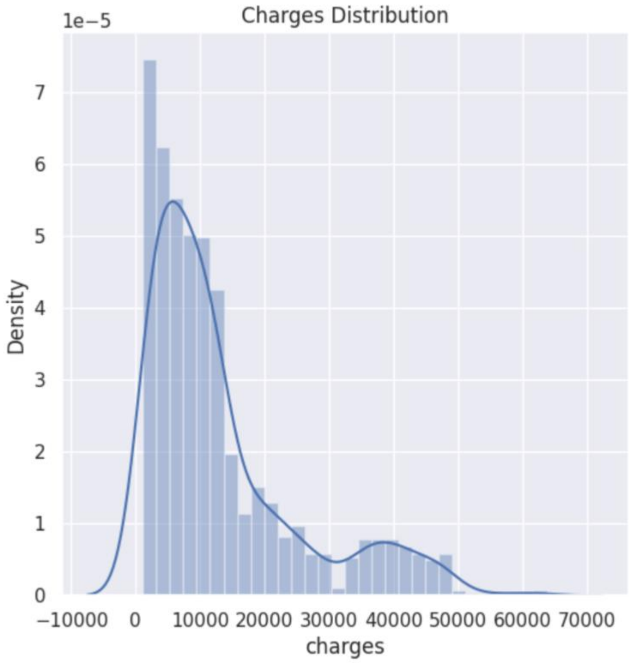


Fig 4.1 – Charges distribution graph

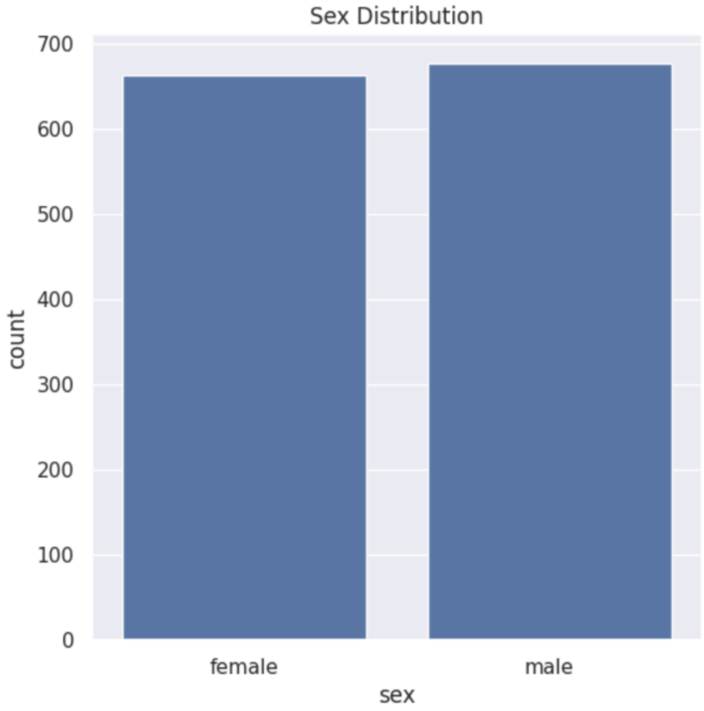


Fig 4.2 – Sex distribution graph

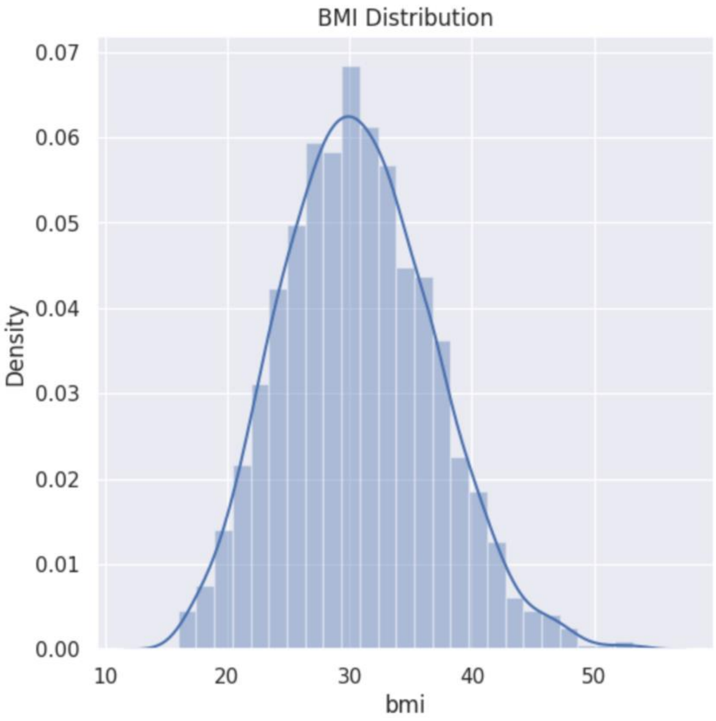


Fig 4.3 BMI Distribution graph

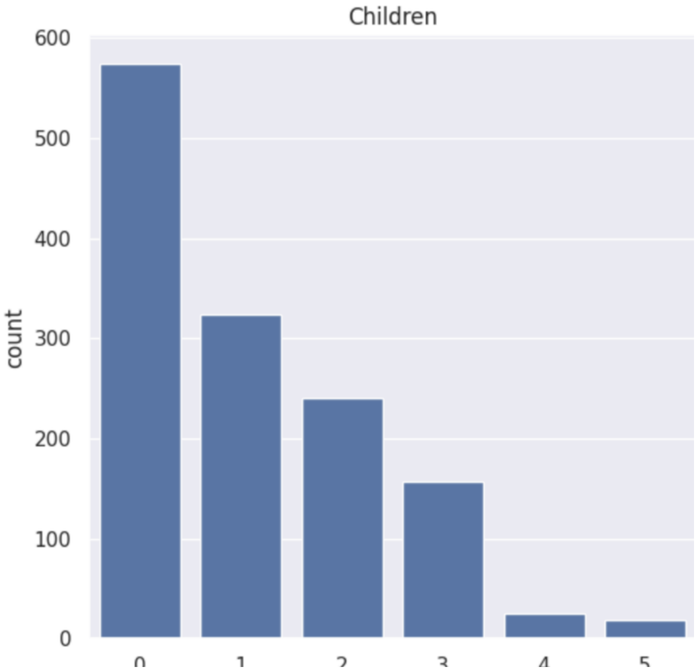


Fig 4.4 Having Children graph

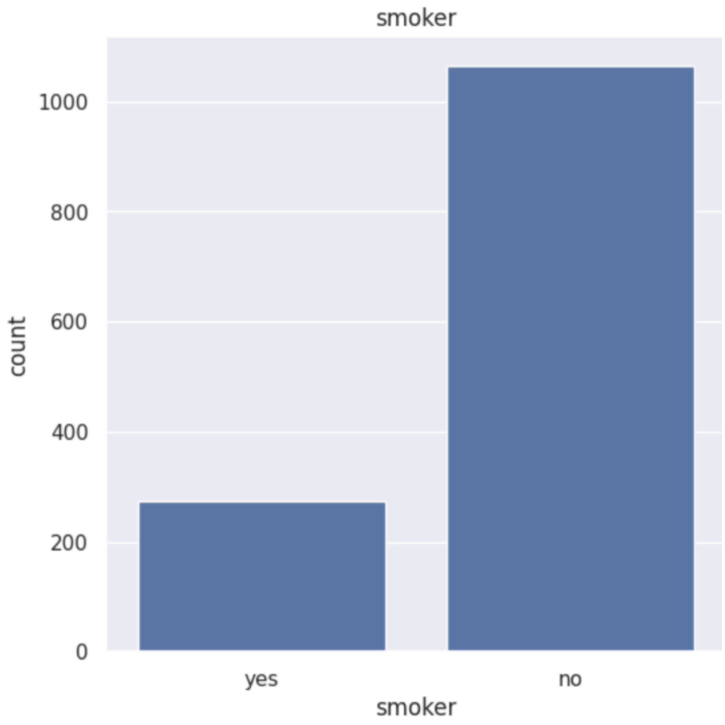


Fig 4.5 Smoker distribution graph

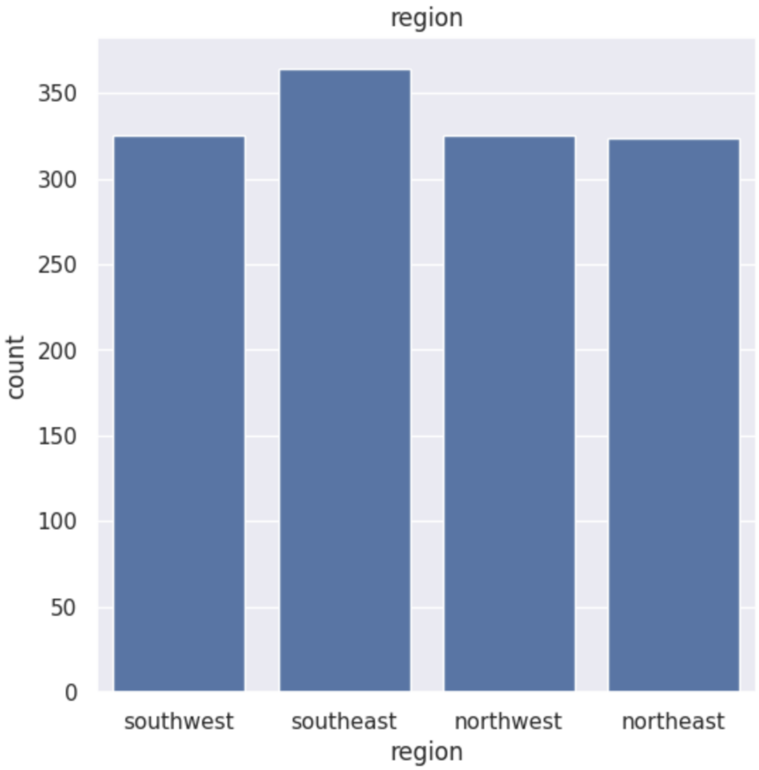


Fig 4.6 Region distribution graph

4.3 Model Training and Evaluation

1. Splitting the Data: The dataset was split into training and testing sets using an 80-20 split ratio (X_train, X_test, Y_train, Y_test), ensuring a balanced representation of the data.
2. Training the Model:
   - A Linear Regression model was applied to the training dataset. The model's performance on the training data was evaluated using the R-squared metric, yielding a value of 0.7515, indicating that the model captured approximately 75.15% of the variance in the data based on the features.
   - A Decision Tree Regressor was also trained using the training dataset. On the training data, the Decision Tree Regressor achieved an R-squared value of 1.0, indicating perfect fitting to the training data. However, this could also suggest a risk of overfitting.
3. Testing the Model:
   - The Linear Regression model was then evaluated on the test set. The R-squared value for the test data was 0.7447, demonstrating that the model was able to generalize reasonably well to unseen data, explaining approximately 74.47% of the variance in the medical insurance costs.
   - The Decision Tree Regressor, when tested on the test set, achieved an R-squared value of 0.6856, suggesting it explained approximately 68.56% of the variance in the medical insurance costs for unseen data. While slightly lower than Linear Regression, this reflects the model's potential overfitting to the training data.

4.4 Building the Predictive System and Validation

To validate the model's prediction capabilities, a specific input data point was used:

- Input Data: (age = 31, sex = female, bmi = 25.74, children = 0, smoker = no, region = southeast)
- Reshaping the Input: The input data was converted into a numpy array and reshaped to match the model's input requirements (reshape(1, -1)).

Prediction Result:

Linear Regression Model:

- The predicted insurance cost for the given input was $3760.08.
- This prediction aligns with expected values for a non-smoker with a moderate BMI, indicating that the model can provide reasonable cost estimates based on demographic and health-related input features.

Decision Tree Regressor:

- The predicted insurance cost for the same input was $3756.62.
- This result also falls within the expected range and reflects the Decision Tree's ability to make reasonable predictions based on the input features. However, slight variations compared to the Linear Regression model highlight the Decision Tree's sensitivity to the specific training data patterns.

Both models successfully provided predictions consistent with the expected insurance costs for the given demographic and health-related data. This demonstrates that the predictive systems, despite differences in methodology, can generalize reasonably well to unseen data points.

4.5 Cross-Validation and Interpretation

To ensure that the model's performance was consistent, cross-validation techniques were applied. The results confirmed that the model's accuracy was not an outcome of overfitting. Additionally, various data visualizations, including histograms and count plots, were employed to interpret the model and validate the distribution of each feature against the insurance costs.

4.6 Limitations and Future Work

Despite the reasonable performance of both the Linear Regression and Decision Tree models, several limitations remain:

1. **Linear Regression Limitations**:
    - The simplicity of the Linear Regression model may not fully capture complex, non-linear interactions between features, especially in cases involving high

insurance costs.

2. **Decision Tree Limitations:**
   - While the Decision Tree Regressor provides a better fit to the training data (R-squared = 1.0), it shows signs of overfitting, as evidenced by a lower R-squared value (0.6856) on the test set. This suggests that the model struggles to generalize well to unseen data.
   - Decision Trees can also be sensitive to small variations in the data, which may lead to inconsistent predictions.
3. **Categorical Features:**
   - The current encoding of categorical features (e.g., region, smoker status) might not capture all nuances, potentially limiting the models' predictive power. Employing ensemble methods like Random Forests or Gradient Boosting, which can handle categorical data more effectively, could enhance model accuracy.
4. **Feature Interactions:**
   - Neither model currently incorporates interaction terms (e.g., smoker status combined with age or BMI). Including these interactions could improve the models' ability to predict more complex patterns in medical insurance costs.
5. **Future Directions:**
   - To address overfitting observed in the Decision Tree, future work could involve applying regularization techniques or exploring more robust algorithms like Random Forests or Gradient Boosting, which aggregate multiple trees to reduce variance and improve generalization.
   - Expanding the dataset to include more diverse samples and features could provide additional insights and improve model robustness.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The objective of this project was to develop predictive models for estimating medical insurance costs based on various demographic and health-related factors such as age, gender, BMI, number of children, smoking status, and geographical region. Two models were implemented: Linear Regression and Decision Tree Regressor.

The **Linear Regression model** demonstrated satisfactory performance, with R-squared values of 0.7515 on the training set and 0.7447 on the test set. This indicates that the model captures a substantial portion of the variance in the dataset, effectively explaining the influence of key factors like BMI, smoking status, and age on insurance costs.

The **Decision Tree Regressor** offered a perfect fit to the training data with an R-squared value of 1.0. However, its R-squared value on the test set was 0.6856, suggesting a tendency to overfit the training data. Despite this, the Decision Tree model proved its ability to capture non-linear relationships in the data, which Linear Regression could not fully address.

Both models validated that features such as BMI, smoking status, and age significantly impact insurance costs, aligning with medical and economic expectations. By integrating these models into a predictive system, users can input their data to obtain cost estimates, offering practical applications for insurers and individuals.

While the results are promising, the limitations of both models highlight areas for future improvement. Linear Regression is limited in its ability to model non-linear relationships, while the Decision Tree model, though more flexible, showed reduced generalization capability. Future work can focus on leveraging advanced ensemble techniques like Random Forests or Gradient Boosting to strike a balance between accuracy and robustness. These steps could further enhance the predictive power and reliability of the system.

5.2 Future Work

1. Exploring Advanced Models: To address the limitations of the current approach, future work could explore using advanced models such as Random Forests, Gradient Boosting

Machines (e.g., XGBoost), or Neural Networks. These models can handle non-linear interactions and are likely to improve prediction accuracy.

2. Feature Engineering: Future iterations can focus on creating interaction terms and polynomial features to better capture the relationships between variables. For example, the interaction between age and smoker status could provide insights into how aging affects the cost burden for smokers differently than for non-smokers.
3. Hyperparameter Tuning: Applying grid search and other hyperparameter optimization techniques could further fine-tune the model, improving its performance.
4. Incorporating Additional Data: Adding more features, such as medical history, lifestyle information (e.g., physical activity level), or socioeconomic factors, could refine the model and make it more comprehensive. These additional variables may help identify more patterns and enhance the prediction of insurance costs.
5. Model Deployment and Integration: Deploying the predictive model using web-based frameworks like Flask or Django can provide a user-friendly interface for real-time cost estimation. Integrating the model with healthcare or insurance databases could also allow for automated updates and more accurate, up-to-date predictions.
6. Evaluation with Cross-Validation: Implementing K-fold cross-validation and other evaluation methods will ensure the model's consistency and robustness across various data subsets, minimizing the risk of overfitting and improving the generalization of results.

In conclusion, the project has successfully developed a preliminary model for predicting medical insurance costs, providing a foundation for further enhancements and applications in real-world scenarios. Future work should focus on leveraging advanced techniques and expanding the dataset to create a more comprehensive and accurate predictive tool.

REFERENCES

[1] "Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare | CB Insights Research", CB Insights Research, 2022. [Online]. Available: https://www.cbinsights.com/research/report/digitalhealth-startups-redefining-healthcare. [Accessed: 10-Sep- 2022]

[2] J. H. Lee, “Pricing and reimbursement pathways of new ophan drugs in South Korea: A longitudinal comparison. in healthcare,” Multidisciplinary Digital Publishing Institute, vol. 9, no. 3, pp. 296, 2021.

[3] Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE

[4] N. Shakhovska, S. Fedushko, I. Shvorob and Y. Syerov, “Development of mobile system for medical recommendations,” Procedia Computer Science, vol. 155, pp. 43–50, 2019

[5] Medical Cost Personal Datasets: https://www.kaggle.com/datasets/mirichoi0218/insurance

[6] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression, " Risks, vol. 7, no. 2, p. 70, Jun. 2019, doi: 10.3390/risks7020070.

[7] M. hanafy and O. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models", International Journal of Innovative Technology and Exploring Engineering, vol. 10, no. 3, pp. 137-143, 2021. Doi: 10.35940/ijitee.c8364.0110321.

[8] M. A. Morid, O. R. L. Sheng, K. Kawamoto, and S. Abdelrahman, "Learning hidden patterns from patient multivariate time series data using convolutional neural networks: A case study of healthcare cost prediction," arXiv preprint arXiv:2009.06783, 2020.

[9] R. Kshirsagar et al., "Accurate and interpretable machine learning for transparent pricing of health insurance plans," arXiv preprint arXiv:2009.10990, 2020.

[10] J. A. S. Cenita, P. R. F. Asuncion, and J. M. Victoriano, "Performance evaluation of regression models in predicting the cost of medical insurance," arXiv preprint arXiv:2304.12605, 2023.

[11] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," arXiv preprint arXiv:2311.14139, 2023.

[12] M. M. Billa and T. Nagpal, "Medical insurance price prediction using machine learning," Journal of Electrical Systems, vol. 20, no. 7s, pp. 2270-2279, 2024.

[13] "Health insurance cost prediction using machine learning," IEEE Xplore, [Online]. Available: https://ieeexplore.ieee.org/document/9824201.

[14] "Medical insurance cost analysis and prediction using machine learning," IEEE Xplore, [Online]. Available: https://ieeexplore.ieee.org/document/10100057.

[15] "Medical insurance cost prediction using machine learning," ResearchGate, [Online]. Available: https://www.researchgate.net/publication/374553777_Medical_Insurance_Cost_Prediction_Using_Machine_Learning.

[16] "Health insurance cost prediction using machine learning," International Research Journal of Engineering and Technology (IRJET), vol. 11, no. 4, pp. 171-175, 2024. [Online]. Available: https://www.irjet.net/archives/V11/i4/IRJET-V11I4171.pdf