1. Project Title

Predictive Analysis of Medical Insurance Costs Using Machine Learning

2. Author(s)

Group G

Student ID	Name
C1220401	Najma Bashir Ali (leader)
C1220865	Amiro Mohamed Hassan
C1221152	Ahmed Elmi Abdule
C1221159	Abdirahman Abdi Abdulle
C1221340	Said Abdirahman Said
C1220416	Abdullahi Dahir Sheikhow
C1221262	Ahmed Abukar Ali
C1221343	Mohamed Abdulle Mohamud
C1220009	Umulkheyr Mohamud Abdulle

3. Abstract

Accurately estimating medical insurance costs is an important challenge for insurance providers and individuals, as it directly impacts financial planning and risk management. This project applies supervised machine learning techniques to predict individual medical insurance charges based on demographic and health-related attributes. A complete computational data science workflow was followed, including data preprocessing, exploratory data analysis, feature engineering, model training, and evaluation.

Three regression models -

- Linear Regression
- Random Forest
- Gradient Boosting

were implemented and compared. Experimental results show that Random Forest achieved the best performance, explaining approximately 85.6% of the variance in insurance charges.

4. Introduction

Medical insurance pricing depends on multiple factors such as age, body mass index (BMI), smoking habits, number of dependents, and geographic location. Traditional pricing approaches often rely on static rules and manual assessments, which may fail to capture complex nonlinear relationships within the data. With the increasing availability of historical insurance data, machine learning provides an opportunity to improve pricing accuracy and fairness.

This project explores the use of computational data science techniques to build a data-driven predictive model for medical insurance costs.

5. Problem and Motivation

Incorrect estimation of medical insurance costs can lead to financial losses for insurers and unfair premium assignments for customers. The problem addressed in this project is the accurate prediction of insurance charges using historical data. The motivation is to develop a reliable and transparent predictive system that reduces bias, improves risk assessment, and supports informed pricing decisions.

6. Objectives (Business and Data Science Goals)

Business Objectives

- Accurately predict individual medical insurance costs
- Support insurance companies in risk assessment and pricing decisions
- Improve transparency and fairness in insurance premiums

Data Science Objectives

- Clean and preprocess real-world insurance data
- Perform exploratory data analysis to uncover patterns and insights
- Apply feature engineering techniques such as encoding and transformation
- Train and evaluate multiple regression models
- Select the best-performing model using standard evaluation metrics

7. Approach

The project followed a structured computational data science workflow. First, the business problem was clearly defined, followed by exploration of the dataset to understand its structure and limitations. Data preprocessing steps were applied to handle missing values, duplicates, and outliers. Feature engineering techniques such as one-hot

encoding and logarithmic transformation were then applied. Multiple machine learning models were trained and compared using evaluation metrics. Finally, the best-performing model was selected for deployment.

8. Data Sets (Data Description and Brief Data Preparation)

Data Description

The dataset consists of 1,338 medical insurance records with the following attributes:

- Age

- Sex

- BMI

- Number of Children

- Smoking Status

- Region

- Insurance Charges (target variable)

Brief Data Preparation

- Missing values were handled using median (numerical) and mode (categorical)
- Duplicate records were removed
- Categorical variables were encoded using one-hot encoding
- The target variable was log-transformed to address skewness

9. Tools and Analytics

The project was implemented using the following tools:

- **Python:** Core programming language

- **Pandas & NumPy:** Data manipulation and numerical computation

- **Matplotlib & Seaborn:** Data visualization

- **Scikit-learn:** Machine learning models and evaluation metrics

- **Joblib:** Model persistence

Analytics techniques included regression modeling, feature scaling, transformation, and ensemble learning.

10. Data Science Solution / Results

Three regression models were trained and evaluated: - Linear Regression - Random Forest Regressor - Gradient Boosting Regressor

Model Performance Summary

- **Linear Regression:** RMSE = 7939, R^2 = 0.60
- **Random Forest:** RMSE = 4787, R^2 = 0.856
- **Gradient Boosting:** RMSE = 4978, R^2 = 0.844

Random Forest achieved the best overall performance and was selected as the final model.

11. Contribution and Uniqueness

The project demonstrates a complete end-to-end computational data science workflow, from raw data to a deployable model. A key contribution is the comparison of linear and ensemble models, highlighting the effectiveness of Random Forest in capturing nonlinear relationships. The project also emphasizes explainability through visualization and metric interpretation.

12. Recommendation

Based on the results, Random Forest is recommended for predicting medical insurance costs in practical applications. Further improvements could include hyperparameter tuning, incorporating additional health indicators, and deploying the model as a real-time web service.

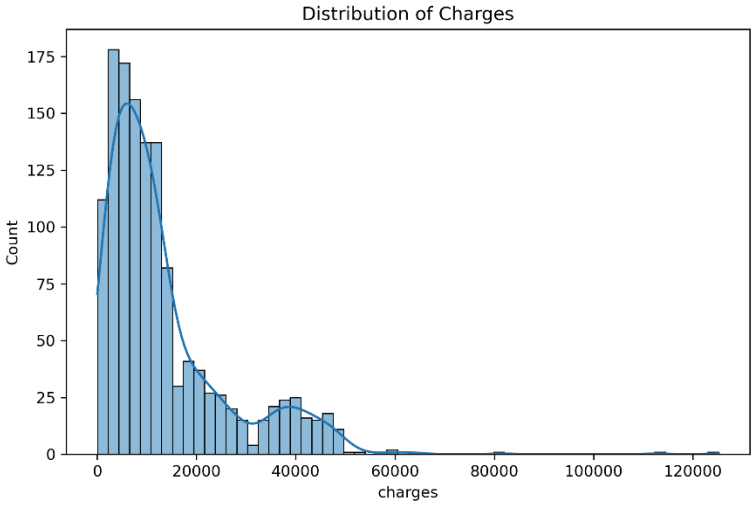
13. Conclusion

This project successfully applied machine learning techniques to predict medical insurance costs with high accuracy. Through systematic data preprocessing, feature engineering, and model evaluation, Random Forest emerged as the most effective model. The study highlights the importance of ensemble methods for complex tabular data in real-world applications.

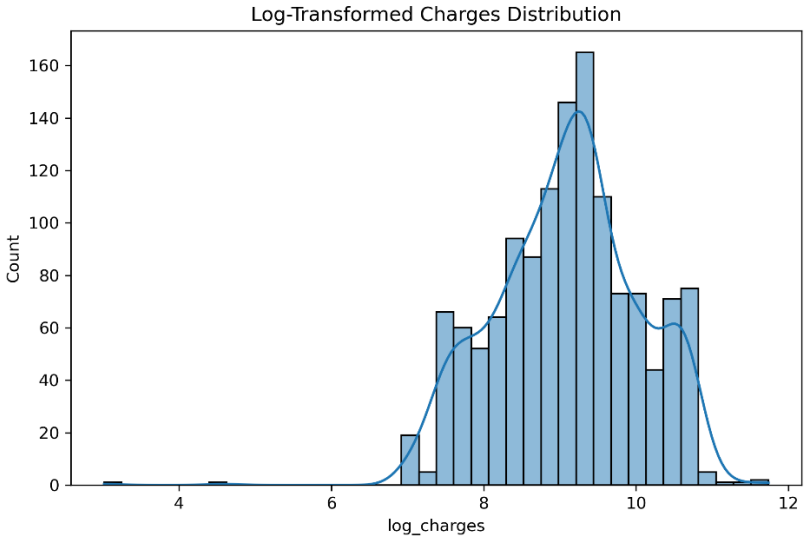
14. Appendix: Visualization / Analytics Summary

The appendix includes key visualizations used in the project, such as:

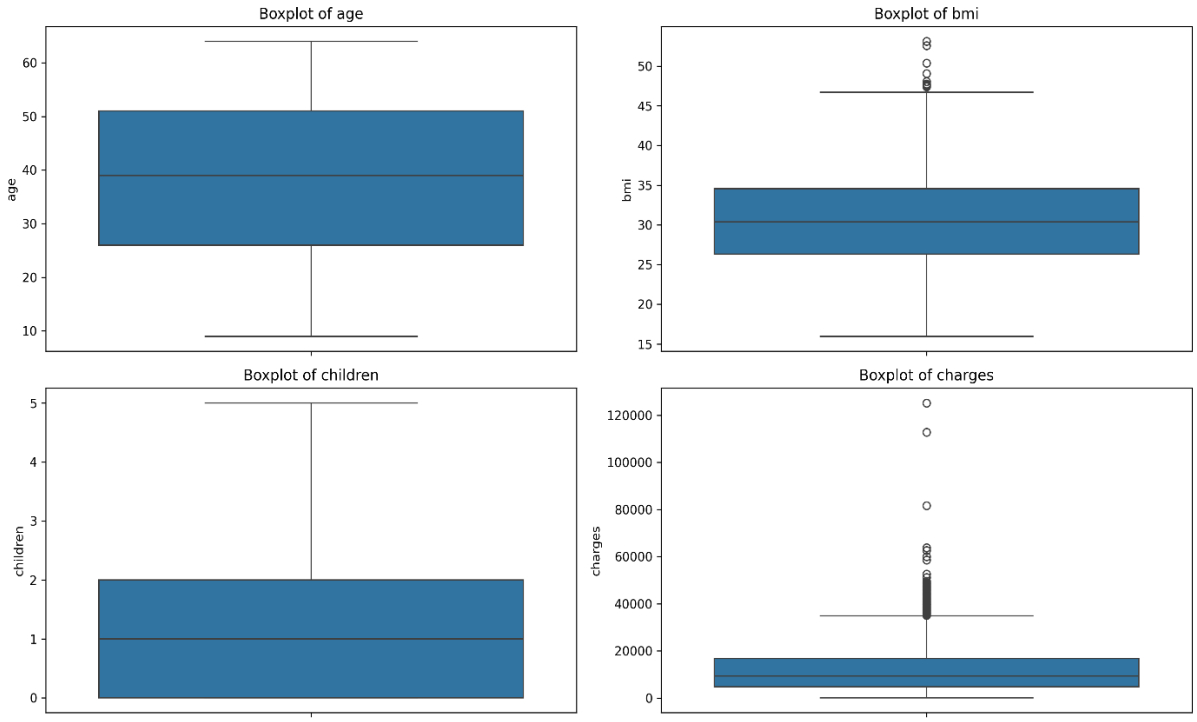
- Distribution of insurance charges



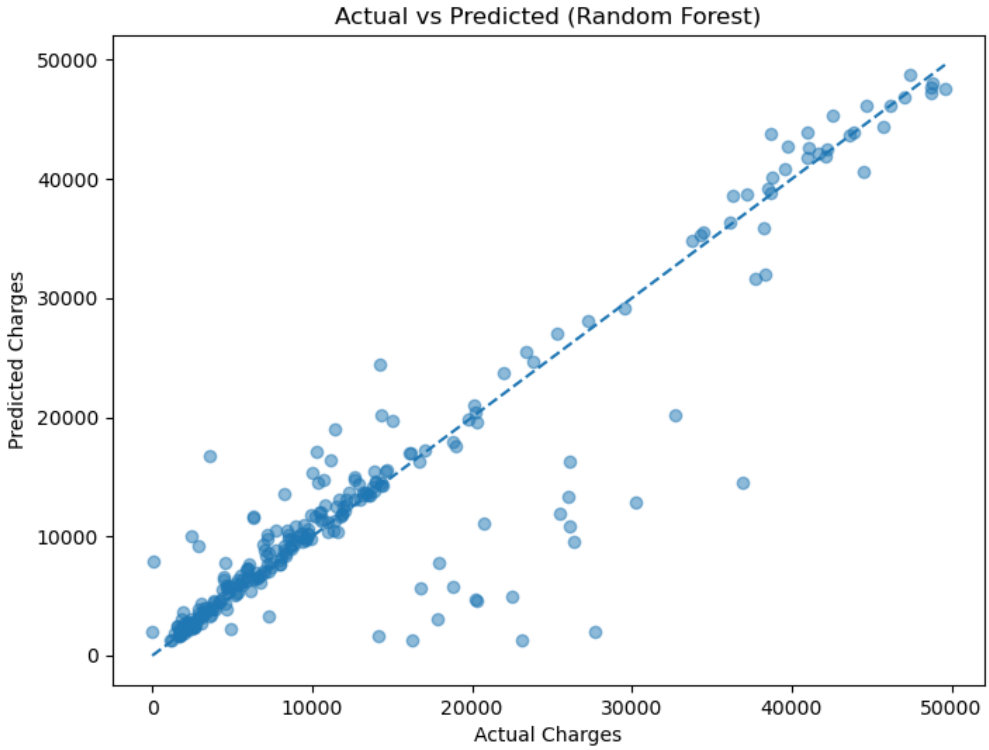
- Log-transformed charge distribution



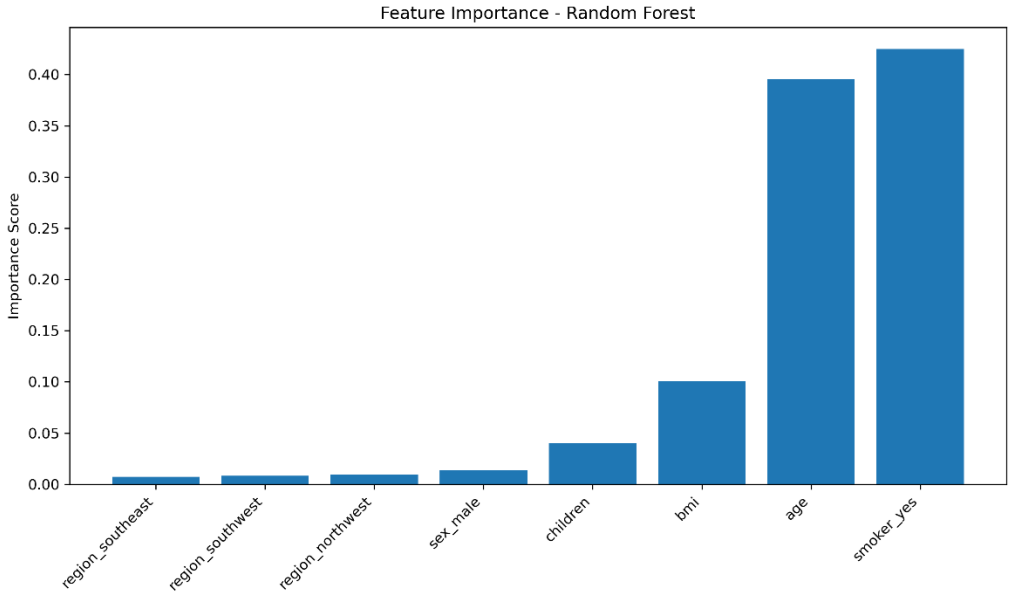
- Boxplots for numerical features



- Actual vs. predicted values



- Feature importance plots



15. References

- Scikit-learn Documentation
- Kaggle Medical Insurance Cost Dataset
- (Hastie et al., 2001)
- (Samant, 2025)