 Lebanese International University-Beirut Campus

Department of Computer Science

CSCI441-Machine Learning

**Team**: Houssam Alrifaii & Bilal Alhalabi

**ID**: 12030024 & 12133799

**Instructor**: Ali Sabra

**Topic**: Predicting Medical Costs and Classifying Smokers: A Machine Learning Approach

Contents

1. Acknowledgements
2. History
3. Objective
4. Theoretical Part
   1. Logistic Regression
   2. Linear Regression
   3. Ridge Regression
   4. Decision Trees
   5. Random Forest
5. Description of the Database
6. Our Study Approach
7. Data Loading and Importing Libraries
8. Exploratory Analysis
9. Model Validation
10. Data Transformation and Partitioning
11. Regression Modeling
    1. Evaluation Metrics (Table Format)
    2. Predictions and Insights
12. Classification Modeling
    1. Evaluation Metrics (Table Format)
    2. Predictions and Insights
13. Feature Importance for Regression and Classification
14. Model Comparison for Regression and Classification
15. Conclusion
16. References
17. Acknowledgements

We extend our gratitude to Kaggle for providing the dataset and to the contributors of open-source libraries such as pandas, sklearn, seaborn, and matplotlib, which were instrumental in completing this project

.

1. History

Machine learning models have increasingly been used in the insurance industry to predict costs and assess risk factors. Our study builds on this foundation to explore and improve predictive capabilities.

1. Objective

The objective of this study is to analyze and model insurance data to:

1. Predict medical costs using regression methods.
2. Classify individuals as smokers or non-smokers using classification methods.
3. Provide insights for cost management and policy optimization.
4. Theoretical Part
5. Logistic Regression: A statistical method used for binary classification. Logistic regression models the probability of a dependent variable as a function of independent variables using the sigmoid function. It predicts probabilities that help classify data points into two categories.
6. Linear Regression: A fundamental regression algorithm that predicts continuous outputs by finding the line of best fit. The model establishes a linear relationship between independent variables and the target variable by minimizing the sum of squared errors.
7. Ridge Regression: A regularized version of linear regression that adds a penalty to large coefficients, reducing overfitting. Ridge regression minimizes a loss function with an added penalty term proportional to the square of the coefficients.
8. Decision Trees: A non-parametric supervised learning algorithm used for both regression and classification tasks. Decision trees split data into subsets based on feature thresholds, forming a tree structure. The algorithm aims to minimize entropy or maximize information gain at each split.
9. Random Forest: An ensemble learning method that builds multiple decision trees during training and merges their predictions to improve accuracy and control overfitting. It uses techniques like bootstrapping and feature randomness to enhance performance.
10. Description of the Database

The dataset consists of 1338 entries, with attributes such as age, sex, BMI, number of children, smoking status, region, and charges. Details of each feature are provided below:

age: age of primary beneficiary

sex: insurance contractor gender, female, male

bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

children: Number of children covered by health insurance / Number of dependents

smoker: Smoking

region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

charges: Individual medical costs billed by health insurance

- Dataset Size: 1338 rows and 7 columns.

- BMI Categories:

- 705 obese

- 385 overweight

- 221 healthy

- 24 underweight

1. Our Study Approach

➢ Step 1: Data Import and Preparation: Loading the dataset and necessary libraries.

➢ Step 2: Descriptive Analysis: Exploring the characteristics of the data and detecting missing values.

➢ Step 3: Preprocessing: Cleaning the data, converting qualitative variables into quantitative ones, and verifying the conditions for regression.

➢ Step 4: Data Partitioning: Splitting the data into training and testing sets.

➢ Step 5: Model Application: Training and evaluating the regression and classification models using metrics.

➢ Step 6: Results Comparison: Analyzing the models' performance and execution time.

1. Data Loading and Importing Libraries

We used the following libraries:

- pandas: For data manipulation and preprocessing.

- numpy: For numerical computations.

- seaborn and matplotlib: For visualization of data patterns.

- scikit-learn: For implementing machine learning models and evaluation metrics.

The dataset was loaded, cleaned, and prepared for analysis using these libraries.

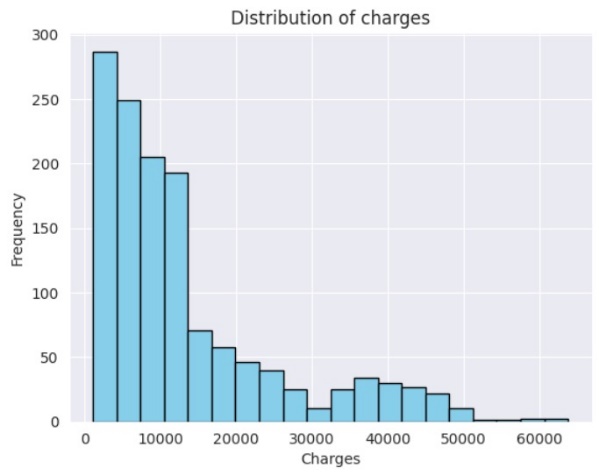
1. Exploratory Analysis

The dataset contains information about individuals' age, sex, BMI, number of children, smoking status, and insurance charges. Before applying any machine learning models, we first explored the data to understand its characteristics and patterns.

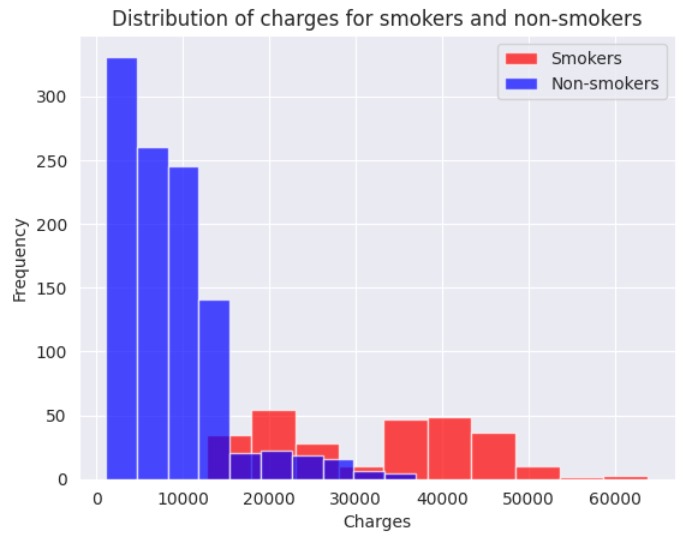
We performed statistical and graphical analyses to identify patterns and relationships:

Distribution of Charges:

Charges are right-skewed, with most individuals incurring costs below $10,000. Smokers typically incur higher charges compared to non-smokers.

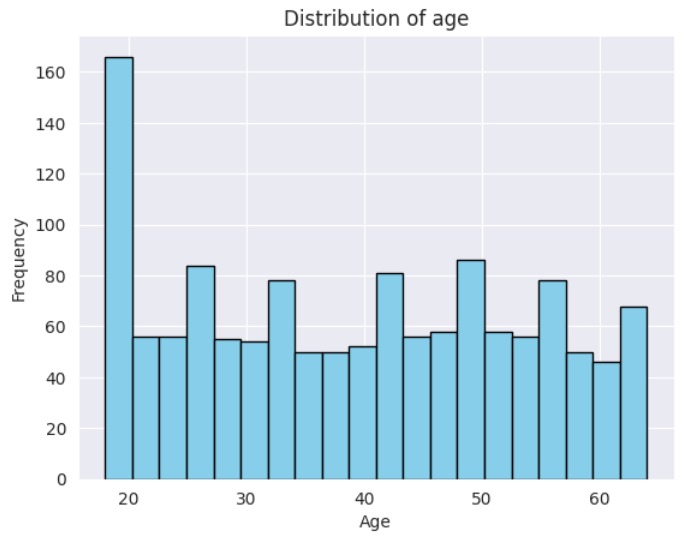


Distribution of Charges for Smokers and Non-Smokers:

Smokers tend to have significantly higher insurance charges, with their costs predominantly shifted towards the higher range. Non-smokers, on the other hand, have a much lower frequency of charges exceeding $30,000.

Distribution of Age:

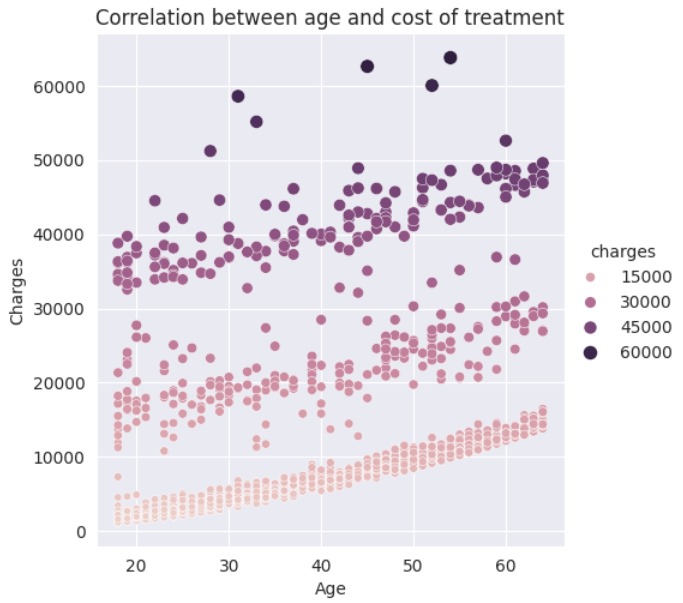
This histogram shows the distribution of age in the dataset. It appears to be relatively uniform, indicating that the dataset includes individuals from a wide range of ages.

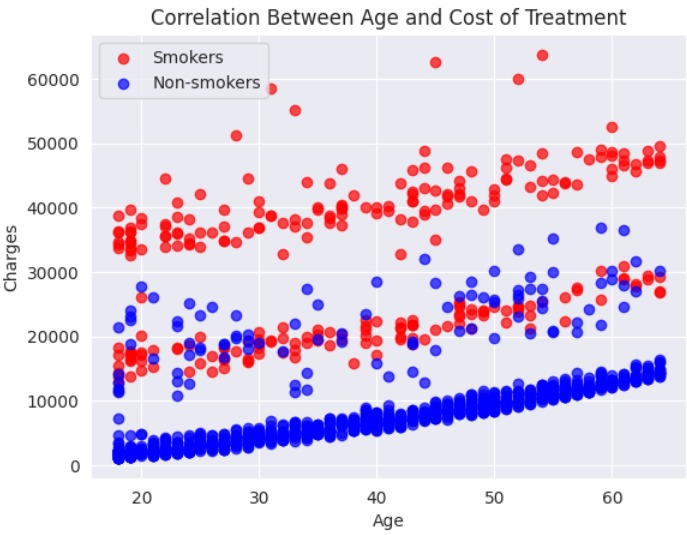


Correlation Between age and cost of treatment:

There is a positive correlation between age and charges, indicating that older individuals tend to have higher insurance costs.

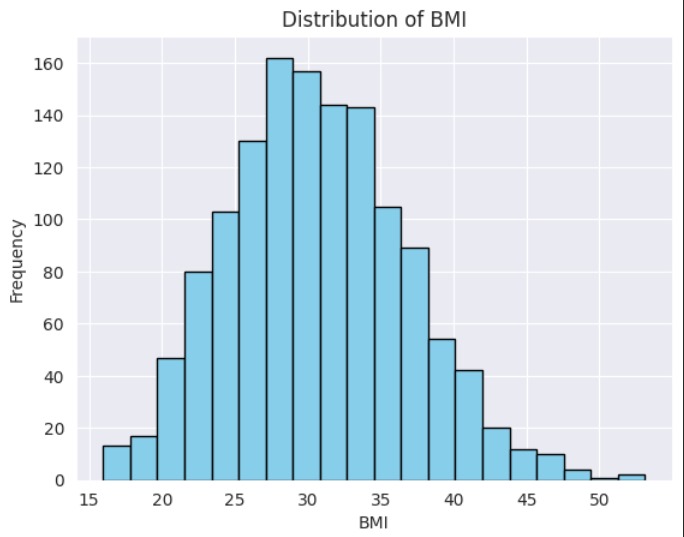
Younger individuals (below 30 years) generally have lower charges, while older individuals (above 50 years) tend to have higher charges, with some outliers.





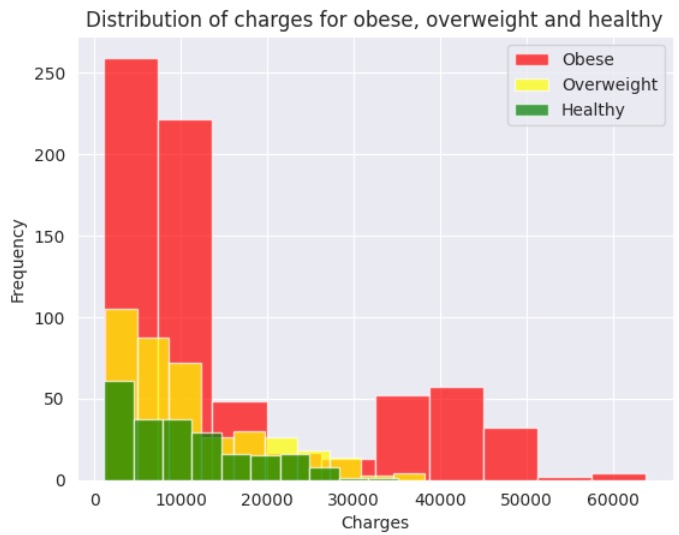
Distribution of BMI:

This histogram displays the distribution of BMI (Body Mass Index) values in the dataset. It appears to be approximately normally distributed, centered around 30.



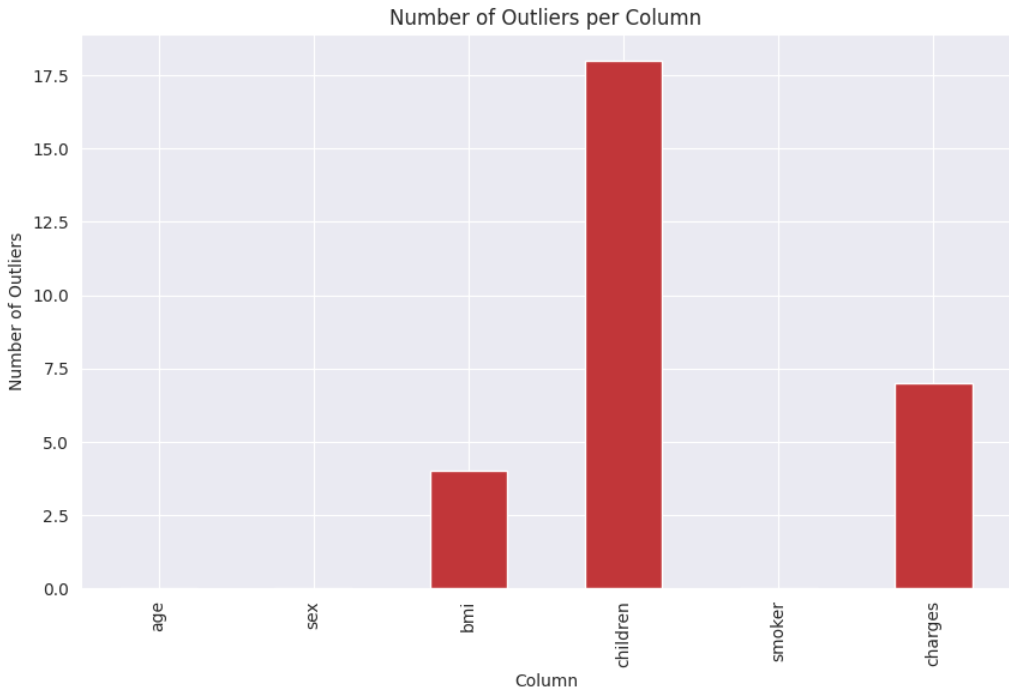
Distribution of Charges for Different BMI Categories:

This histogram compares the distribution of charges for different BMI categories: obese (red bars), overweight (yellow bars), and healthy (green bars). Obese individuals tend to have the highest frequency of high charges, followed by overweight individuals. Healthy individuals have a lower frequency of high charges. Obese individuals have a higher frequency of charges exceeding $30,000 compared to overweight and healthy individuals.



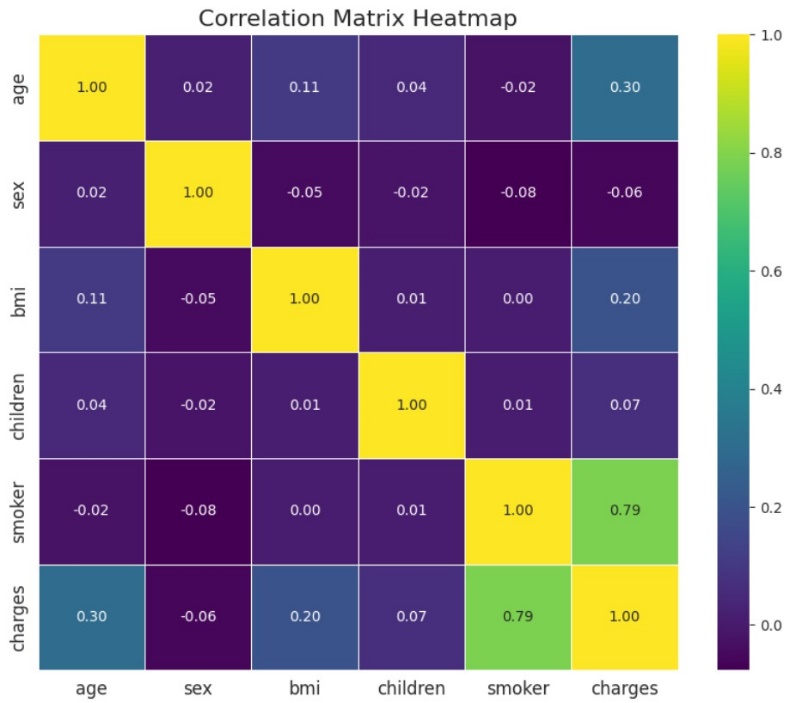
Number of outliers:

This bar plot shows the number of outliers detected in each column of the dataset. Outliers are data points that significantly deviate from the typical pattern. This information helps us understand the data quality and potential need for handling outliers.



Correlation Matrix Heatmap:

This heatmap visualizes the correlation between different features in the dataset. Smoking (smoker) has a strong positive correlation with charges, confirming our previous observations.



1. Model Validation

We ensured model robustness by applying the following:

Splitting the data into training (80%) and testing (20%) sets.

Cross-validation to ensure generalizability.

Evaluating metrics such as MAE, MSE, RMSE, R², accuracy, precision, recall, and F1 score.

1. Data Transformation and Partitioning

Label Encoding: Applied to `sex` and `smoker` to convert categorical variables into numerical format for machine learning models.

1. Regression Modeling

Evaluation Metrics:

* **Mean Absolute Error (MAE):** Simple average error.
* **Mean Squared Error (MSE):** Average squared error, penalizing larger errors more
* **Root Mean Squared Error (RMSE):** MSE but in the same units as the target variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MAE | MSE | \*RMSE | \*R² Score\* |
| Linear Regression | 0.067 | 0.0096 | 0.0981 | 0.788 |
| Ridge Regression | 0.066 | 0.0097 | 0.0986 | 0.888 |
| Decision Tree | 0.045 | 0.005 | 0.071 | 0.888 |

Predictions and Insights:

The regression models demonstrated that smoking status, BMI, and age are the most significant predictors of medical costs. Decision Tree performed the best in terms of minimizing error metrics.

1. Classification Modeling

Evaluation Metrics:

 **Accuracy**: Overall correctness.

 **Precision**: Focuses on avoiding false positives.

 **Recall**: Focuses on capturing all actual positives.

 **F1 Score**: Balances precision and recall in a single score.

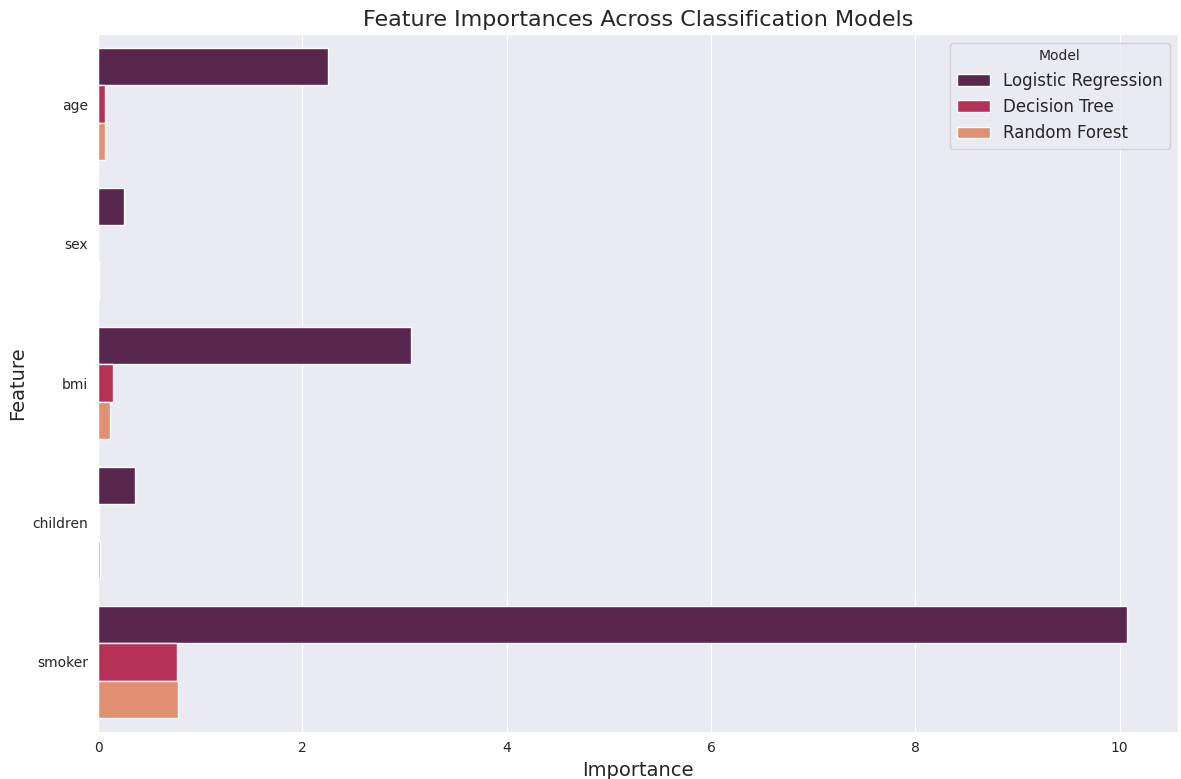
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| Logistic Regression | 92.13% | 0.94 | 0.88 | 0.9 |
| Decision Tree | 97.75% | 0.96 | 0.97 | 0.98 |
| Random Forest | 97.00% | 0.95 | 0.97 | 0.96 |

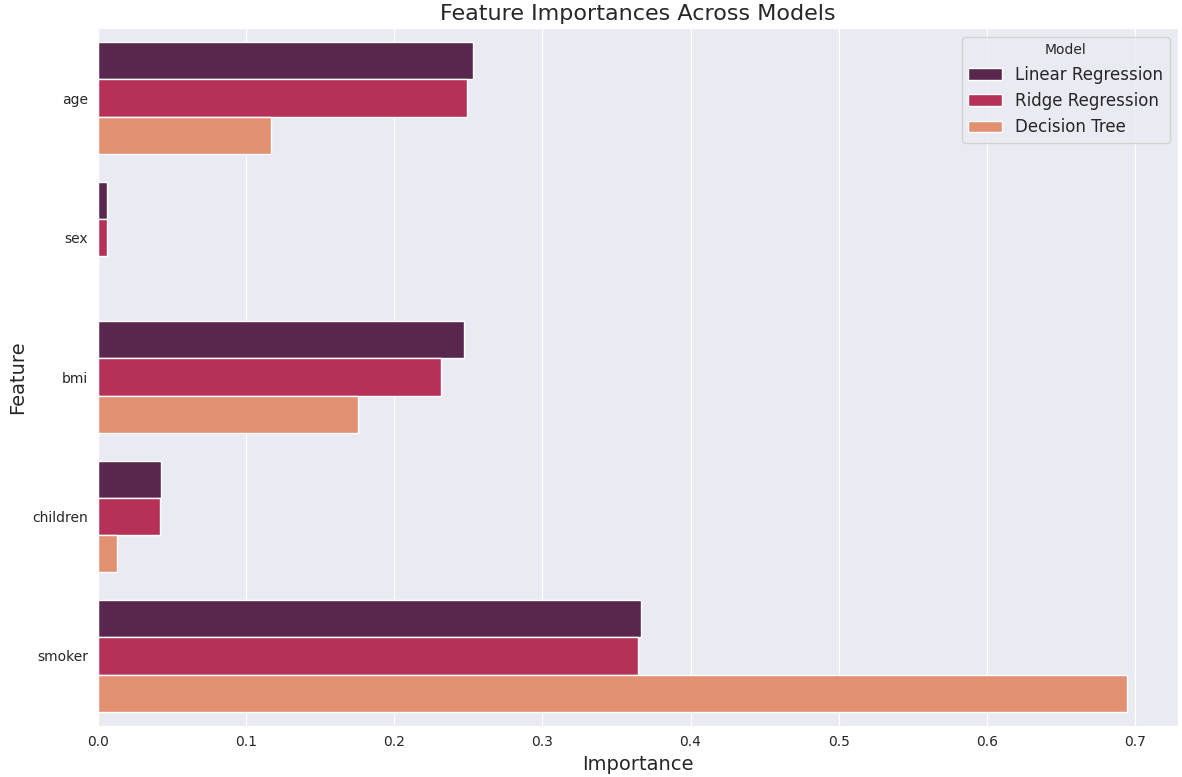
Predictions and Insights:

Decision Tree achieved the highest accuracy in classifying smokers, highlighting the robustness of ensemble methods.

1. Feature Importance for Regression and Classification

Bar plots were generated to compare the significance of features in predicting charges (regression) and classifying smoking status (classification). Smoking emerged as the most impactful feature in both tasks.

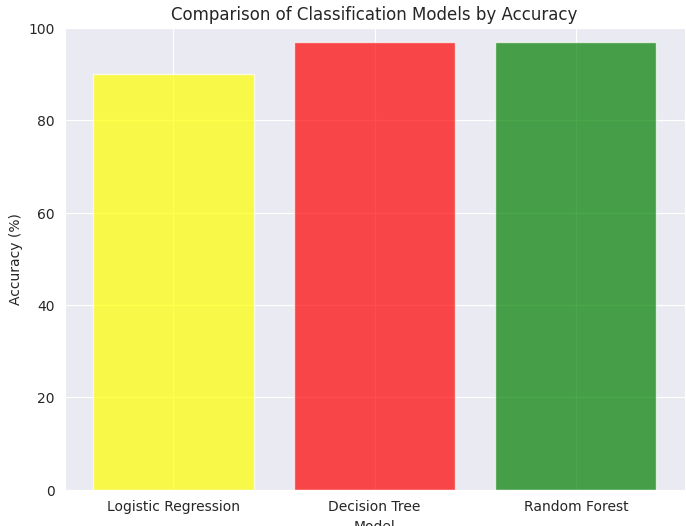


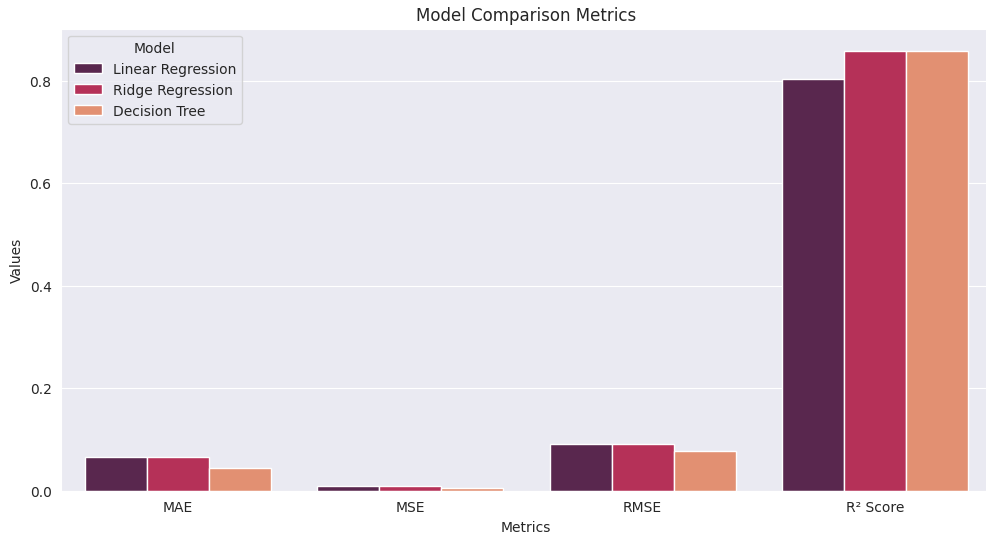


1. Model Comparison for Regression and Classification

**Regression Models**: Decision Tree outperformed Linear and Ridge Regression with the lowest RMSE and highest R² score.

**Classification Models**: Random Forest and Decision Tree showed comparable results, with Random Forest slightly outperforming.





1. Conclusion

This study demonstrates the feasibility of using machine learning to predict medical costs and classify smokers. Smoking, BMI, and age emerged as key factors influencing predictions. We recommend using Decision Tree to classify if the person is smoker or not, and use Decision Tree to calculate medical costs.

1. References

1. Kaggle Dataset: [https://www.kaggle.com/](<https://www.kaggle.com/>)

2. Scikit-learn Documentation: [https://scikit-learn.org/](https://scikit-learn.org/)

3. Matplotlib Documentation: [https://matplotlib.org/](https://matplotlib.org/)