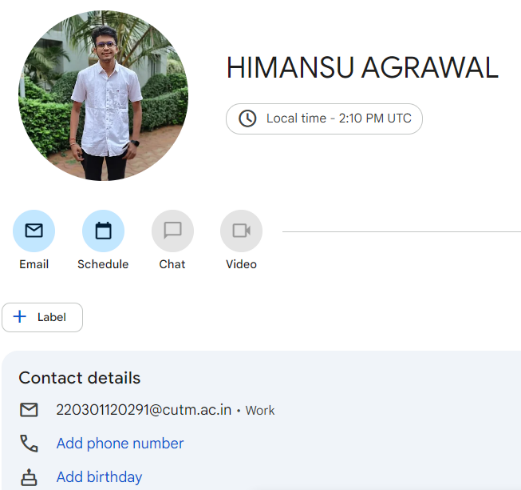
**Health Insurance Cost Prediction:**   
A Comparative Analysis of Machine Learning Algorithms for Predicting Health Insurance Costs



**Himansu Agrawal**

Regd. No. – 220301120291

Mail – [himansubansal1701@gmail.com](mailto:himansubansal1701@gmail.com)

Content

Abstract

1. Introduction
2. Data Overview
3. Data Preprocessing
4. Data Cleaning and Encoding
5. Data Visualization
6. Feature Scaling
7. Model Selection and Training
8. Linear Regression
9. Support Vector Regression
10. Ridge Regressor
11. Random Forest Regressor
12. Model Evolution
13. Comparison of Models
14. Visualization of Model Performance
15. Prediction Example
16. Conclusion
17. Appendix

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Abstract**   
The primary goal of this project is to develop a machine learning model that can predict health insurance costs based on various factors such as age, gender, BMI, number of children, smoking status, and region. The project employs several machine learning algorithms to find the most accurate and efficient model for this task. The models tested include Linear Regression, Support Vector Regression (SVR), Ridge Regression, and Random Forest Regressor. This report outlines the methodology, data preprocessing steps, model training, evaluation metrics, and a comparison of the performance of different models.

1. **Introduction**

This section introduces the broader context and motivation behind the project. Health insurance costs vary significantly based on individual factors such as age, gender, and lifestyle habits. By accurately predicting these costs, insurance companies can better price their policies and offer more personalized packages to their customers. The introduction emphasizes the practical applications of the project, such as helping insurers make data-driven decisions.

1. **Data Overview**

This section details the dataset and its features, which are crucial for building the predictive models. It lists and explains each feature in the dataset:

* ***Age:*** A continuous variable representing the individual's age, a critical factor influencing health costs.
* ***Sex:*** A binary variable representing gender, where 0 is male and 1 is female.
* ***BMI (Body Mass Index):*** A measure of body fat based on weight and height, which can impact health and, consequently, insurance costs.
* ***Children:*** A count of the number of children covered by the insurance plan, which can influence total charges.
* ***Smoker:*** A binary indicator of smoking status (0 for non-smokers, 1 for smokers), a major determinant of health costs.
* ***Region:*** A categorical variable indicating the individual’s geographical region (encoded numerically as 0 for northwest, 1 for northeast, etc.).
* ***Charges:*** The target variable represents the health insurance costs.

**Formula:** β0​+β1​×age+β2​×bmi+β3​×children+β4​×smoker

1. **Data Preprocessing**

Data preprocessing is a critical step to prepare the raw data for machine learning models. It includes:

1. *Data Cleaning and Encoding*

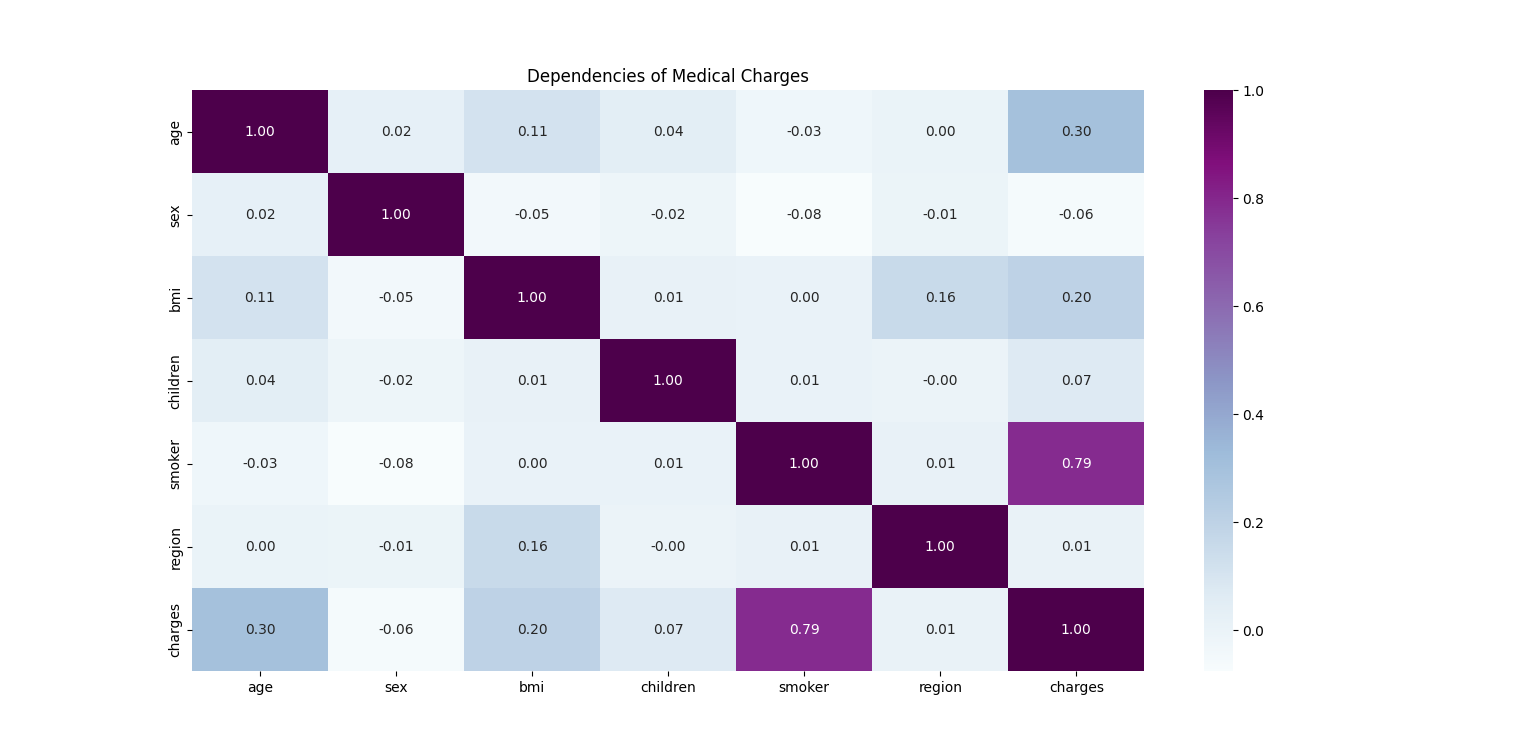
This subsection describes converting categorical variables into numerical forms, such as encoding sex, smoker, and region variables. Handling missing values ensures that the dataset is complete and ready for analysis.

1. *Data Visualization*

Data visualization helps explore relationships between different features and insurance charges, providing insights into which factors might strongly influence costs. Examples of visualizations used include:

* ***Age vs. Charges:*** Highlights how charges increase with age, reflecting higher health risks in older individuals.
* ***Region vs. Charges:*** This shows regional variations, potentially due to differing healthcare costs.
* ***BMI vs. Charges:*** Reveals the correlation between BMI and charges, indicating higher costs for overweight individuals.
* ***Smoker vs. Charges:*** Depicts that smokers incur significantly higher charges than non-smokers.
* ***Sex vs. Charges:*** Explores the impact of gender on insurance costs.

1. *Feature Scaling*

**Feature scaling standardizes the data, which is especially important for models sensitive to the scale of input data, like SVR. This ensures that each feature contributes equally to the model’s learning process.

1. **Model Selection and Training**

Different machine learning models were trained and evaluated to identify the best-performing model.

1. *Linear Regression*

Linear Regression serves as the baseline model. It assumes a linear relationship between the input features and the target variable (charges). Metrics like R² and RMSE are used to evaluate its performance. Cross-validation helps ensure that the results are not just due to overfitting but are generalizable to unseen data.

1. *Support Vector Machine*

SVR is used to capture nonlinear relationships. It uses hyperplanes in a multidimensional space to separate data points. GridSearchCV, a hyperparameter tuning technique, helps find the optimal settings for the model to improve accuracy.

1. *Ridge Regression*

Ridge Regression adds a regularization term to the Linear Regression model, helping prevent overfitting by penalizing large coefficients. This is particularly useful when dealing with multicollinearity or when the model complexity needs to be controlled.

1. *Random Forest Regression*

Random Forest is an ensemble method that uses multiple decision trees to improve prediction accuracy. It handles non-linearity and interactions between features well. Hyperparameters are tuned using GridSearchCV to find the optimal configuration.

1. **Model Evaluation**

The models are evaluated using various performance metrics:

* ***R² Score:*** Indicates the proportion of variance in the dependent variable explained by the model. A higher score suggests a better fit.
* ***RMSE (Root Mean Square Error):*** Measures the average magnitude of errors between predicted and actual values. Lower values indicate better predictive accuracy.
* ***Cross-Validation Score:*** Ensures that the model's performance is consistent across different data subsets, providing a robust measure of its generalizability.

1. *Comparison of Models*

The performance of the models is summarized below:

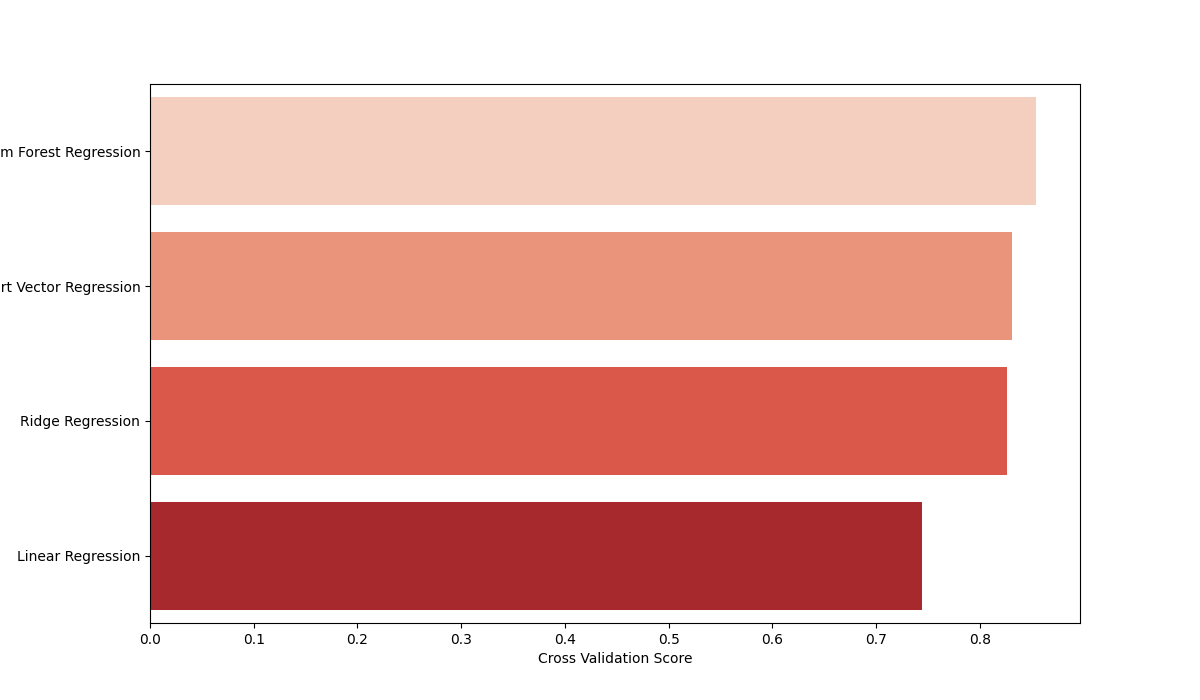
|  |  |  |  |
| --- | --- | --- | --- |
| Model | RSME | R2 Score (Train) | R2 Score (Test) |
| Linear Regression | 0.446 | 0.744 | 0.731 |
| Random Forest Regression | 0.418 | 0.860 | 0.849 |
|  |  |  |  |

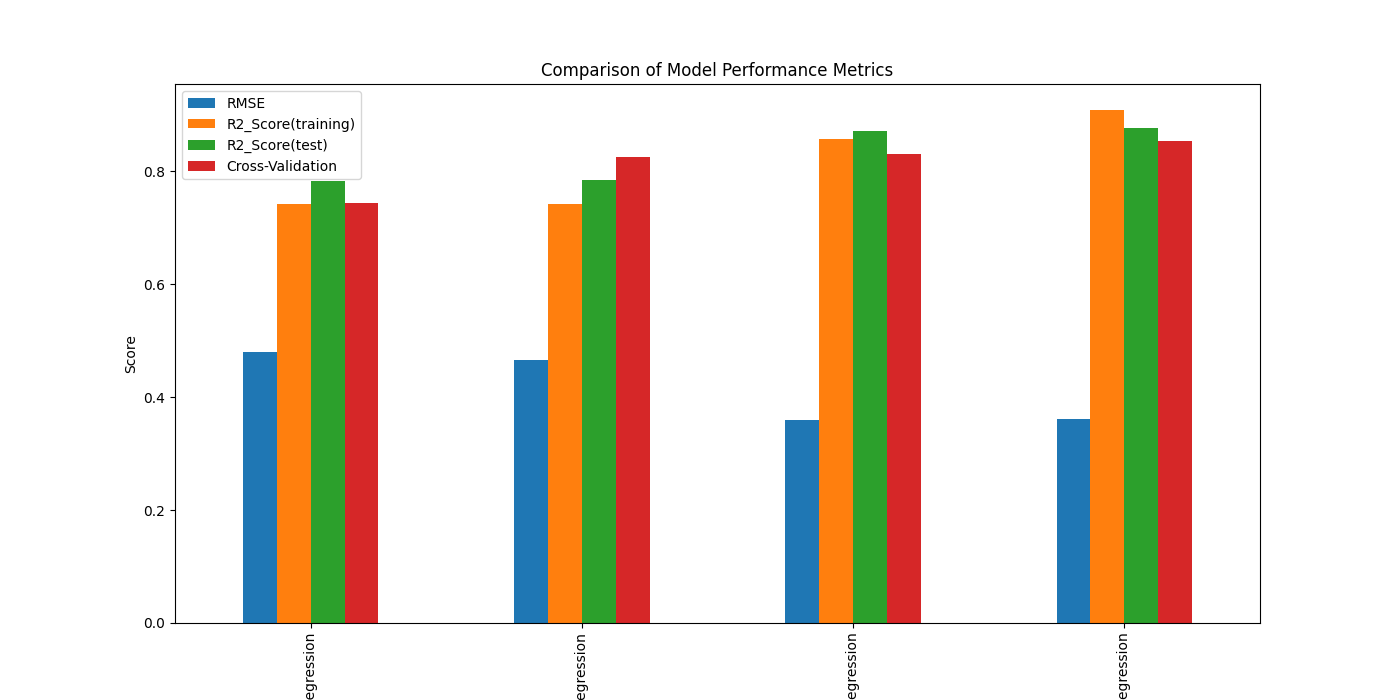
The Random Forest Regressor outperformed all other models, achieving the highest R² score and the lowest RMSE, indicating that it is the best model for predicting health insurance costs in this context.

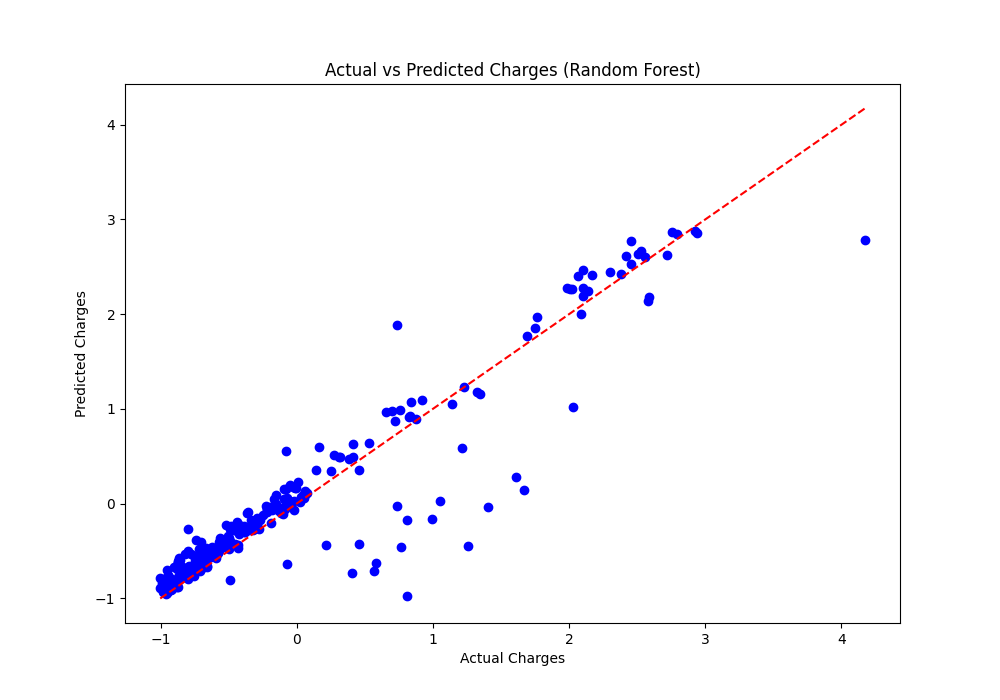
1. *Visualization of Model Performance*

A bar plot visually compares the cross-validation scores of different models, highlighting the Random Forest Regressor's superior performance. This visual aid makes it easier to grasp the differences in model effectiveness.

* *Bar Plot of Cross-Validation Scores:* This plot compares the cross-validation scores of various models, emphasizing the superior performance of the Random Forest Regressor.



* *Bar Plot of Performance Metrics:* This plot provides a comprehensive view of how each model performs across multiple metrics, including RMSE, R² scores for training and testing.
* *Actual vs. Predicted Values:* A scatter plot comparing actual vs. predicted values illustrates the model's accuracy. The x-axis shows actual values, the y-axis displays predicted values, and a diagonal line (y=x) highlights perfect prediction alignment.



1. **Prediction Example**

To demonstrate the predictive capability of the best-performing model, the following inputs were used for prediction:

* Age:36
* Sex (0 for male, 1 for female): 0
* BMI: 23.6
* Number of Children: 1
* Smoker Status (0 for no, 1 for yes): 1
* Region (0 for northwest, 1 for northeast, 2 for southeast, 3 for southwest): 3

Based on these inputs, the predicted insurance cost is **17,628.34**.

1. **Conclusion**

The conclusion summarizes the key findings, emphasizing that the Random Forest Regressor was the most effective model. It highlights the importance of model selection and tuning in achieving high predictive accuracy. The section also suggests future work, such as testing more advanced models like Gradient Boosting Machines or Neural Networks and exploring additional features that might further enhance prediction accuracy.

1. **Appendix**

The appendix includes additional resources such as:

* ***Graphs:*** Visualizations generated during the data exploration and model evaluation phases.
* ***Code:*** Python code used for data preprocessing, training, and evaluation, providing transparency and enabling others to replicate the results.