INSURANCE PREMIUM PREDICTION

USING MACHINE LEARNING

Name: Margi Shah
Semester: 6
Course: Data Science
Roll No: 23
Gujarat University

Introduction

In order to manage the high cost of medical care, health insurance plays a crucial role in our daily lives. However, choosing the right health insurance plan can be challenging as many factors, including age, physical condition, family situation, and location, must be taken into account.

Like many other nations around the world, India is concerned about rising healthcare costs. Growing healthcare expenses can be very demanding on patients, their families, and the healthcare system as a whole. Accurate healthcare cost projection can benefit in the allocation of resources and pricing decisions made by insurance companies and healthcare providers. In this project, we investigate the application of multiple regression models to forecast Indian healthcare costs. We examine a dataset that contains data on various elements that affect healthcare costs, including age, gender, smoking habits, location, and BMI.

Problem Definition

Choosing the best health insurance plan can be difficult because there are many variables that affect the premium cost. The goal is to develop a predictive model that can help people choose health insurance plans based on their health status and expected medical costs.

Literature Review

- Kumar Sharma and Sharma [1] aimed to develop mathematical models for predicting future premiums and validating the findings using regression models. To anticipate policyholders' choice to lapse life insurance contracts, we employed the random forest approach. Even when factoring in feature interactions, the technique beats the logistic model. Azzone et al. [2] studied how the model works; we employed global and local classification tools. The findings suggest that existing models, such as the logistic regression model, are unable to account for the variety of financial decisions.
- Premiums are determined by health insurance companies' private statistical proce dures and complicated models, which are kept concealed from the public. The goal of this study [3] is to see if machine learning algorithms can be used to anticipate the pricing of yearly health insurance premiums on the basis of contract parameters and business characteristics. The goal of this article [4] is to use a strong machine learning model to estimate the future medical costs of patients on the basis of specific parameters. Us ing the simulation results, the elements that influence individuals' medical expenditures were determined

Objective

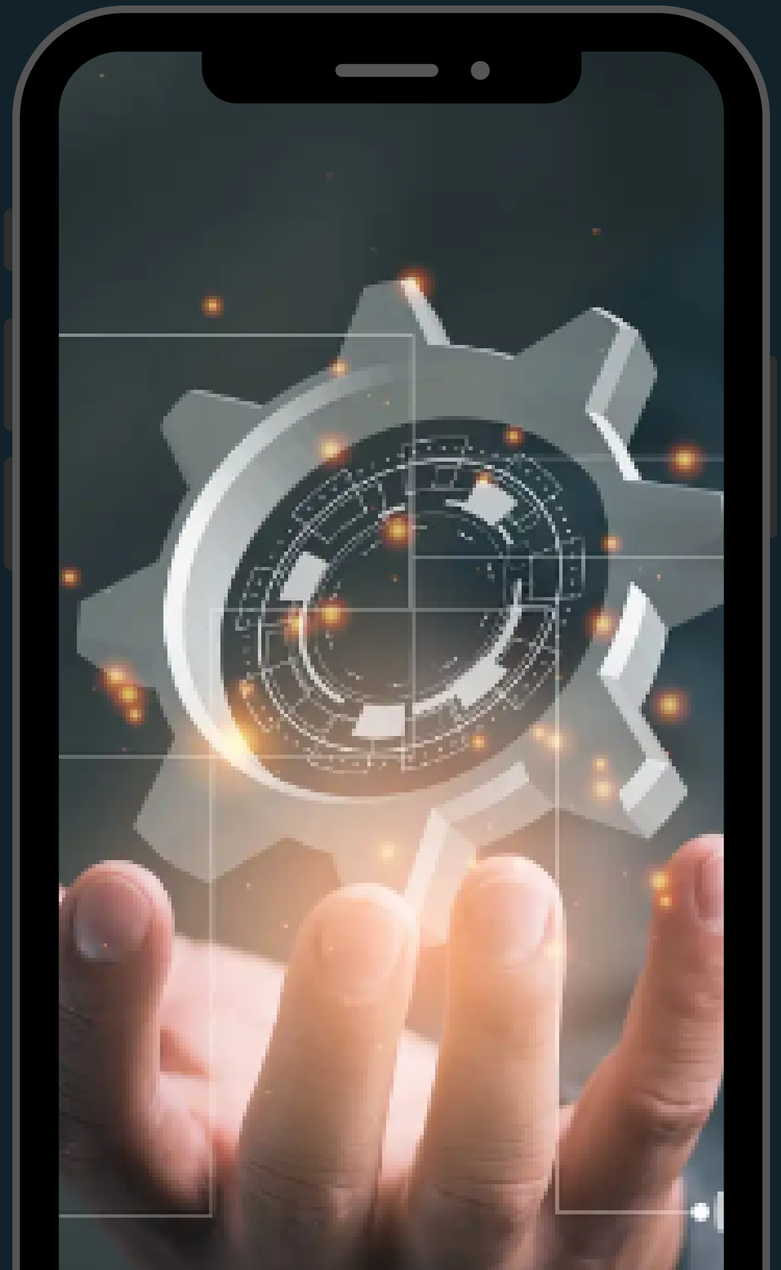
The goal of the project is to forecast an individual's premium amount using a variety of machine learning algorithms, such as multiple linear regression, polynomial regression, gradient boosting regression, and decision tree regression. We seek to provide individuals with an estimated amount they need based on their health condition, which they can take into consideration when choosing any health insurance policy, by exploring the insurance dataset and developing various models.

About Data

There are 1338 observations and 7 features in the insurance.csv dataset. The dataset includes four numerical features—age, bmi, children, and expenses—as well as three nominal features—sex, smoker, and region—that have been converted into factors with numerical values assigned for each level. The data has been pre-processed to ensure data quality and consistency

Methodology

- Data Collection
- Data Cleansing
- Exploratory Data Analysis
- Data Visualization
- Feature Engineering
- Model Training
- Model Evaluation



Process

The first step in the methodology was data exploration, where we explored the dataset to understand the distribution of variables, identify missing values, and detect any outliers. We used various data visualization techniques to understand the relationship between variables and their impact on insurance expenses.

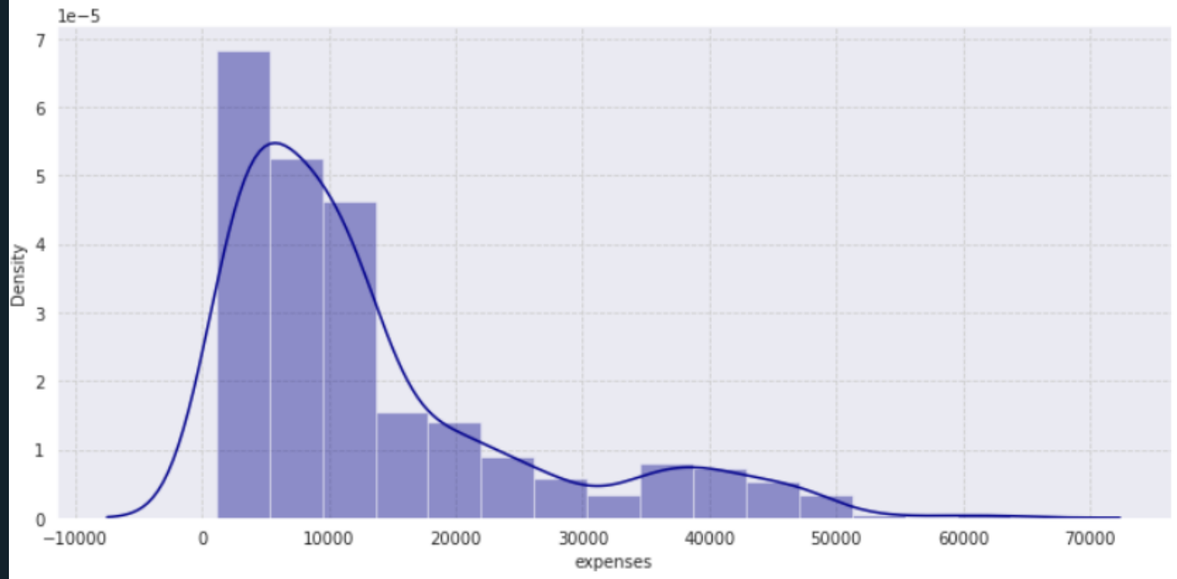
After data exploration, we performed data cleaning to handle any missing or duplicate values, and normalize the dataset. We also performed feature engineering to extract more useful information from the existing variables, such as creating dummy variables for categorical variables like sex and region.

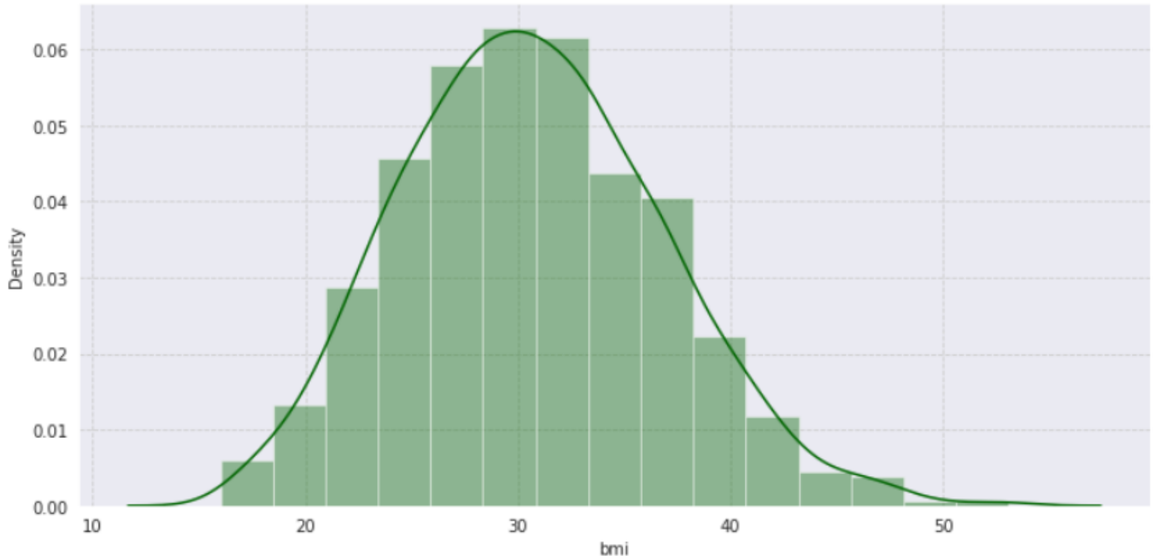
Next, we built various regression models using multiple linear regression, polynomial regression, decision tree regression, and gradient boosting regression. We used the Scikit-learn library in Python to build these models and evaluated their performance using mean absolute error (MAE), mean squared error (MSE), and R-squared metrics.

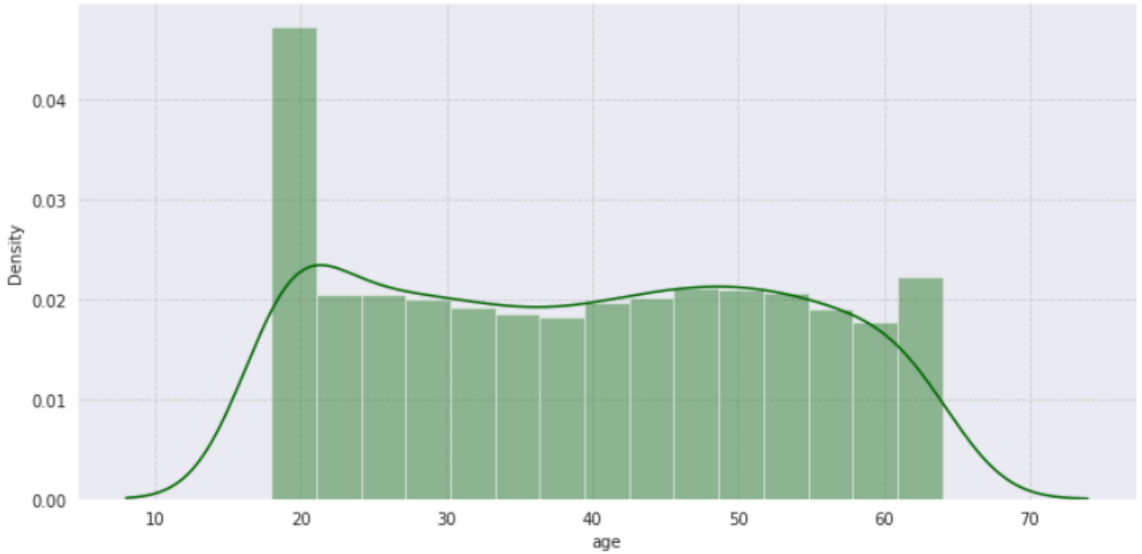
Data Collection

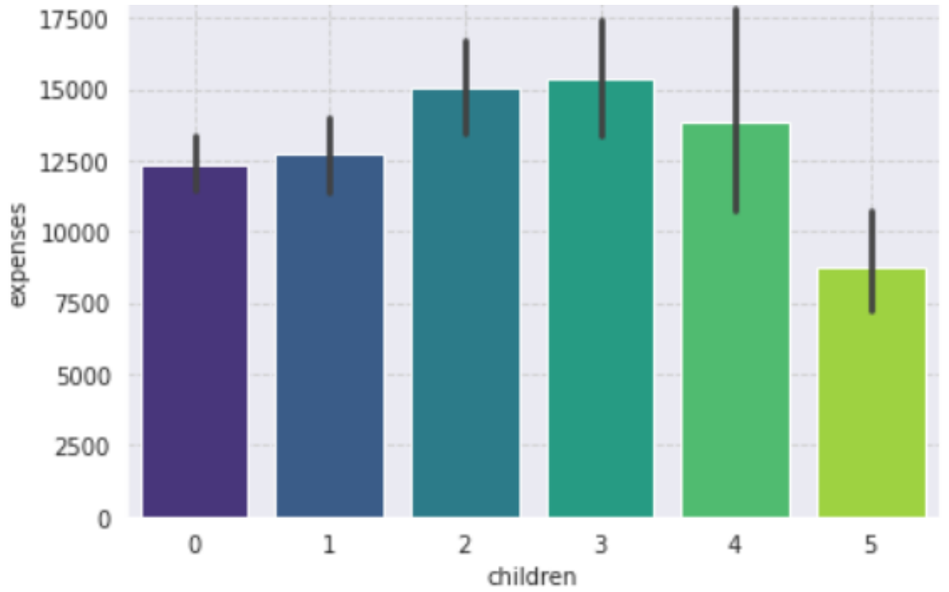
	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86

Exploratory Data Analysis

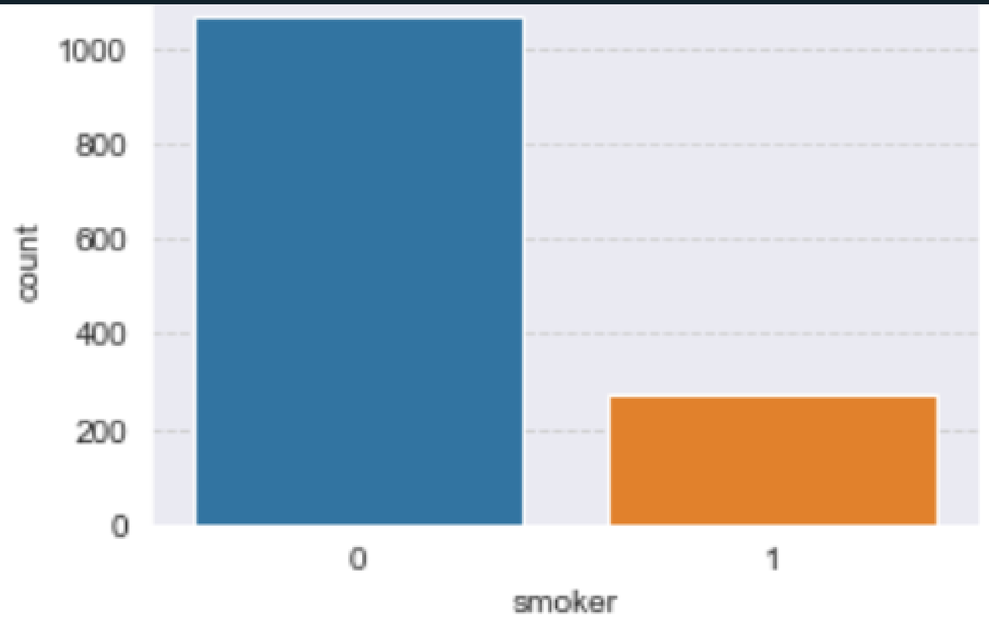


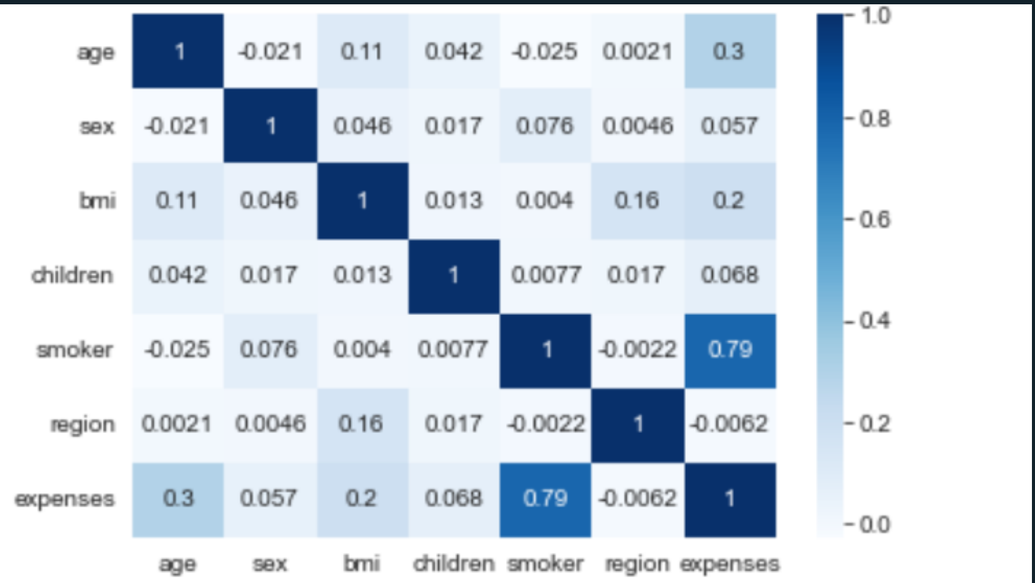


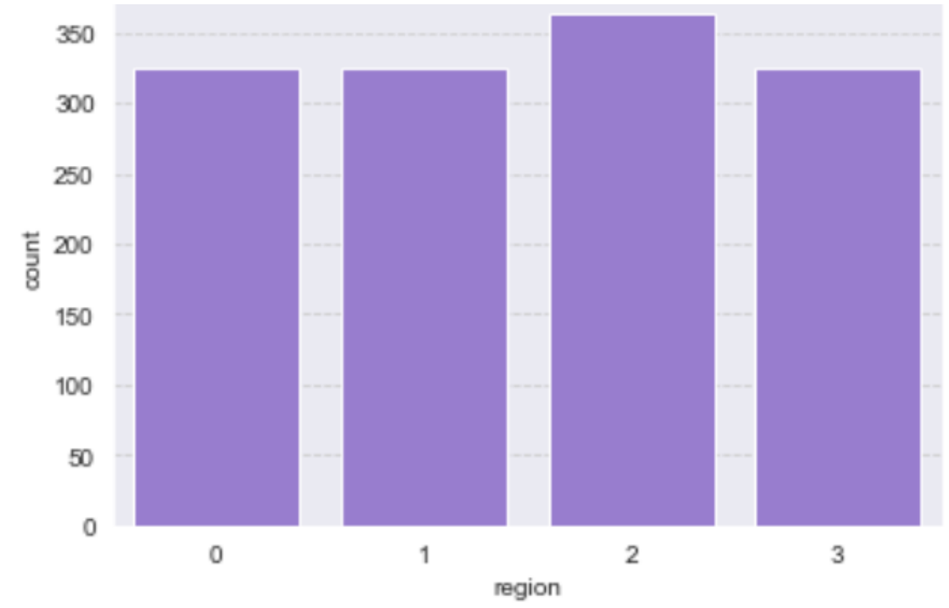


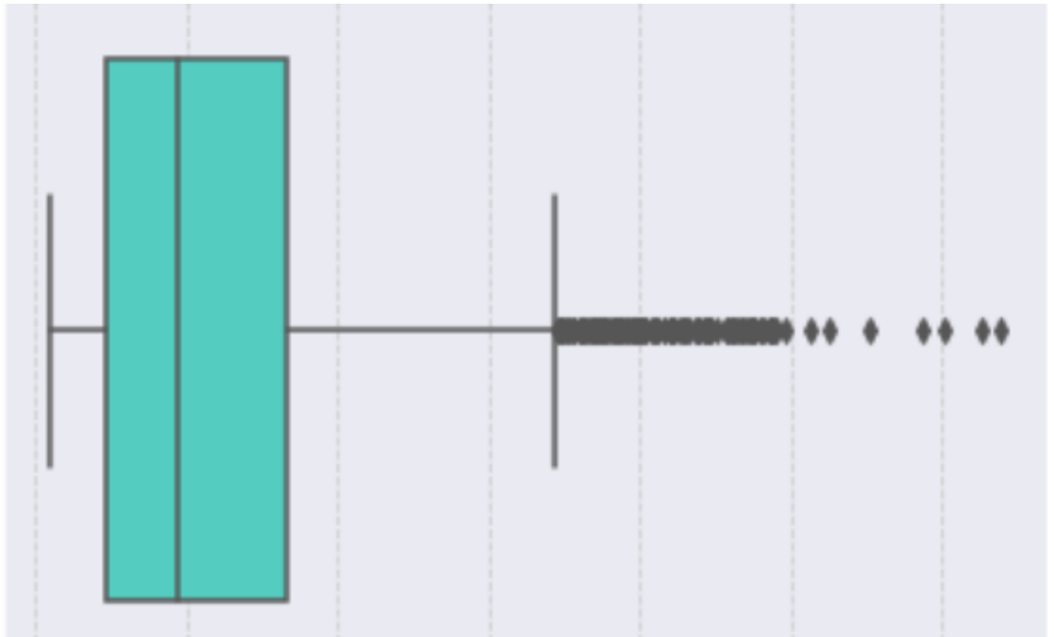


Exploratory Data Analysis









Algorithms Used

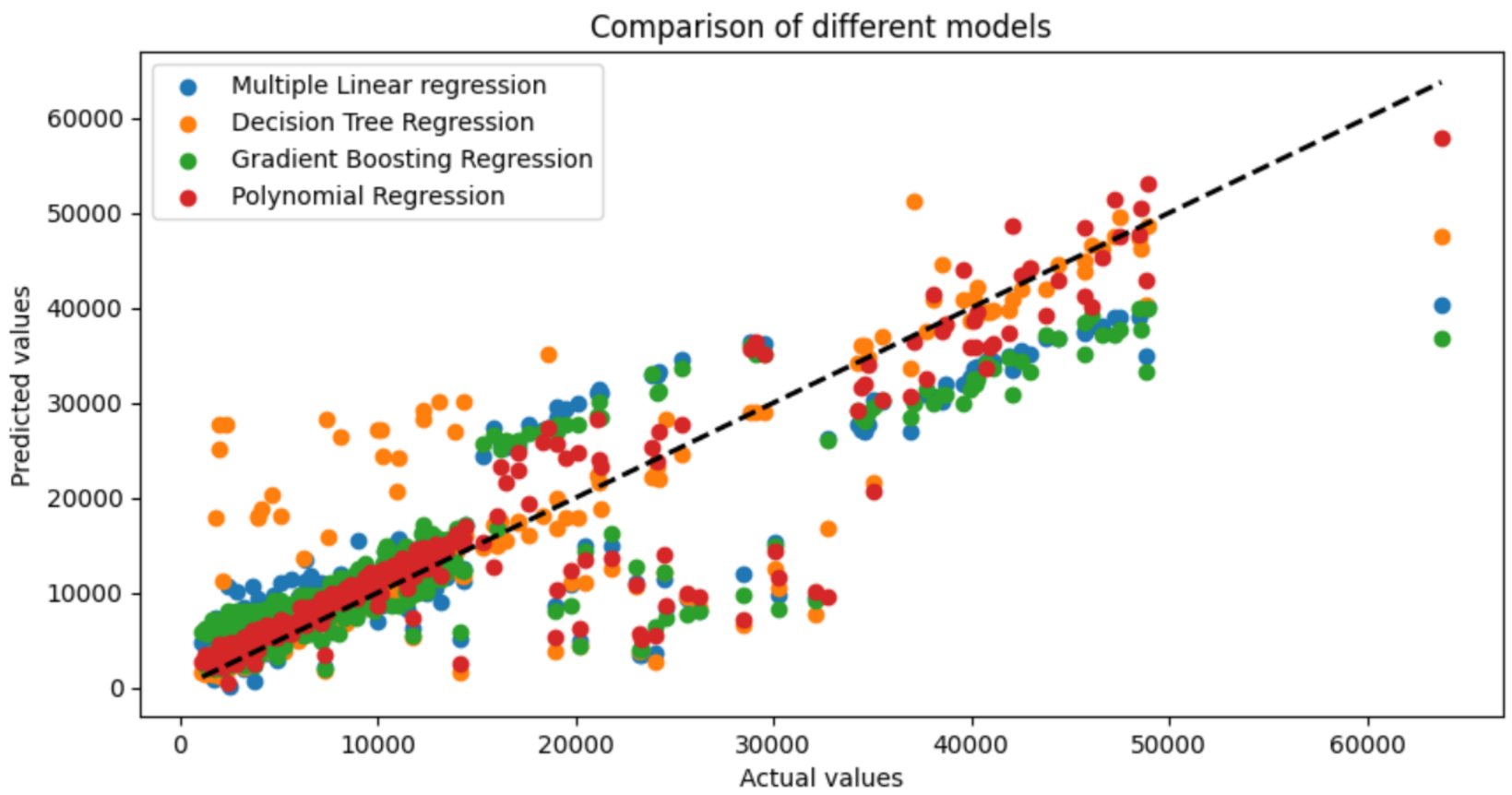
- Multiple Linear Regression
- Polynomial Regression
- Decision Tree Regression
- Gradient Boosting Regression

```
Multiple Linear regression
MAE:  4085.232059801649
MSE:  31904777.378466826
R^2:  0.7714568938909903

Decision Tree Regression
MAE:  2696.7829291044777
MSE:  31754482.804984424
R^2:  0.7725334971923561

Gradient Boosting Regression
MAE:  4198.558328529453
MSE:  32644813.52037938
R^2:  0.766155801941677

Polynomial Regression
MAE:  2866.3114499216335
MSE:  22345397.16245026
R^2:  0.8399334866322379
```



Results

The results of the algorithms have been evaluated based on the MAE, MSE, and R^2 values. The Polynomial Regression algorithm outperformed the other algorithms with an MAE of 2866.31, MSE of 22345397.16, and R^2 of 0.84. The other algorithms like Multiple Linear Regression, Gradient Boosting Regression, and Decision Tree Regression also provided reasonable results, but they were not as accurate as Polynomial Regression.

References

1. Kumar Sharma, D.; Sharma, A. Prediction of Health Insurance Emergency Using Multiple Linear Regression Technique. Eur. J. Mol. Clin. Med. 2020, 7, 98–105.
2. Azzone, M.; Barucci, E.; Giuffra Moncayo, G.; Marazzina, D. A Machine Learning Model for Lapse Prediction in Life Insurance Contracts. Expert Syst. Appl. 2022, 191, 116261. [https://www.sciencedirect.com/science/article/abs/pii/S095741742101570 0?via%3Dihub]
3. Lui, E. Employer Health Insurance Premium Prediction. Available online: http://cs229.stanford.edu/proj2012/Lui-EmployerHealthInsurancePremiumPrediction.pdf (accessed on 17 May 2022).
4. Prediction of Health Expense—Predict Health Expense Data. Available online: https://www.analyticsvidhya.com/blog/2021/0 5/prediction-of-health-expense/ (accessed on 9 May 2022)

THANK YOU