**Health Insurance Prediction Model:**

**Introduction**

The goal of this project is to build a predictive model that will use pertinent information like age, gender, BMI, and smoking status to estimate the costs of individual medical insurance. The dataset used in this study includes detailed data on health insurance claims for a predetermined population, which allows for a more thorough examination of the complex interactions between age, gender, BMI, smoking status, and related healthcare costs. The goal of this dataset exploration is to create a strong predictive model that will not only identify subtle patterns in the data but also offer insightful information about the variables affecting health insurance premiums. This will help to advance knowledge of the dynamics in the healthcare industry.

**Methodology:**

1. **Data Collection and Preprocessing**

The dataset used in this project was obtained from Kaggle. It contains 1338 records and 7 columns, including age, sex, BMI, children, smoker, region, and charges. We removed the region column since it did not provide significant insights for our analysis. We also converted the sex and smoker columns into binary variables (0 or 1) for the ease of our analysis.

We also checked for missing values, and there were none in the dataset. We then split the data into training and testing sets with a 70:30 ratio.

1. **Model Building**

In this project, we used a machine learning algorithm called linear regression to describe the relationship between the dependent variable—medical insurance costs—and a number of independent variables, such as age, sex, BMI, number of children, and smoking status. These relationships are modeled using a linear framework via linear regression. To act as a comparative model, a decision tree was also used as a regressor. By comparing the decision tree model's performance and efficacy to that of the linear regression model, this method makes it possible to assess how well each algorithm predicts medical insurance costs based on the given independent variables.

When we first started the project, we used the training data to create a linear regression model. We used the model to forecast health insurance costs for the test data after it was created. Then, using critical metrics like mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE), the model's performance was carefully assessed. We also included a visual element to this research by creating a scatter plot that compares expected and actual numbers. We expanded our analysis by adding a decision tree model in addition to the linear regression method. An overall view of the efficiency of the linear regression and decision tree models in predicting medical insurance costs based on the given variables was provided by this supplementary analysis, which made it possible to compare their predictive powers. This thorough assessment methodology seeks to provide light on each model's advantages and disadvantages in order to promote a more nuanced comprehension of each model's suitability for use in the context of estimating health insurance costs.

1. **Results**

Our linear regression model performed well on the test data, with an MAE of 1171.00, MSE of 5395627.336, and RMSE of 2322.8490. On the other hand, the performance of the decision tree approaches with an MAE of 1287.57, MSE of 10811491.25 and RMSE of 3288.08. By comparing the R2 value for each model we have come to a conclusion that the linear regression model is performing better than the decision tree. The scatter plot of the predicted versus actual values showed a strong linear relationship, indicating a good fit of the model.

The program, featuring both a linear regression model and a decision tree model, has practical applications in addressing real-world challenges within the healthcare and insurance industries. Such as:

**Precision in Cost Estimation:**

* The program, employing both linear regression and decision tree models, excels in predicting medical insurance costs with a higher degree of accuracy.
* Accurate cost estimation allows insurance providers to better assess and manage risk, leading to more precise pricing strategies.

**Fairness and Transparency**:

* By incorporating a range of factors and leveraging sophisticated algorithms, the program mitigates biases in insurance pricing.
* This contributes to a fairer distribution of premiums, ensuring that insurance costs align more objectively with individual healthcare needs.

**Equitable Insurance System:**

* The program's ability to provide accurate cost estimates fosters fairness in the overall insurance system.
* More equitable pricing promotes a balanced and inclusive approach, addressing disparities and enhancing the accessibility of insurance coverage.

**Empowering Consumers:**

* Individuals can leverage the predictive models to make informed decisions about their insurance coverage.
* The estimated costs empower consumers to plan for future medical expenses, facilitating a proactive and strategic approach to managing healthcare expenditures.

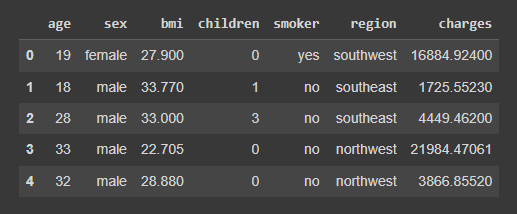
**Informed Decision-Making:**

* Both insurance providers and individuals benefit from the program's insights.
* The program equips stakeholders with valuable information, enabling them to make informed decisions that align with their financial and healthcare objectives.

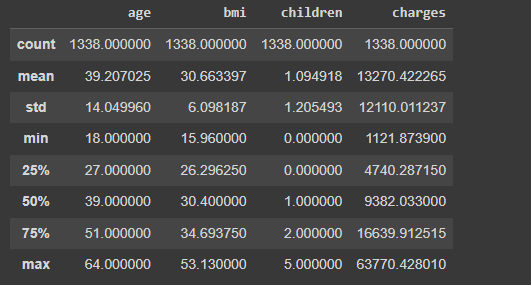
**Conclusion**

To summarize, our machine learning project successfully used linear regression to predict health insurance premiums while taking age, gender, BMI, and smoking status into account. On the test data, the model performed admirably and closely matched our predicted results. Because of its ease of use and efficiency, linear regression is a machine learning technique that can be used for a wide range of situations, most notably the prediction of health insurance costs.

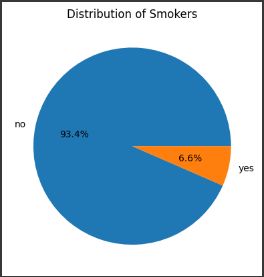
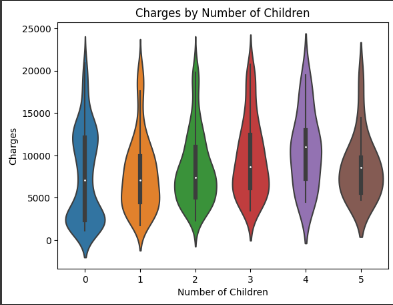
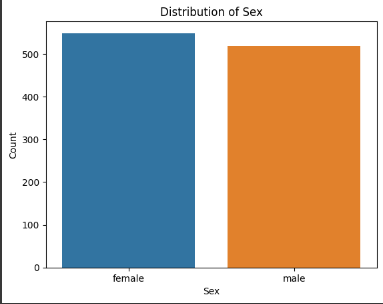
**Appendix:**

****

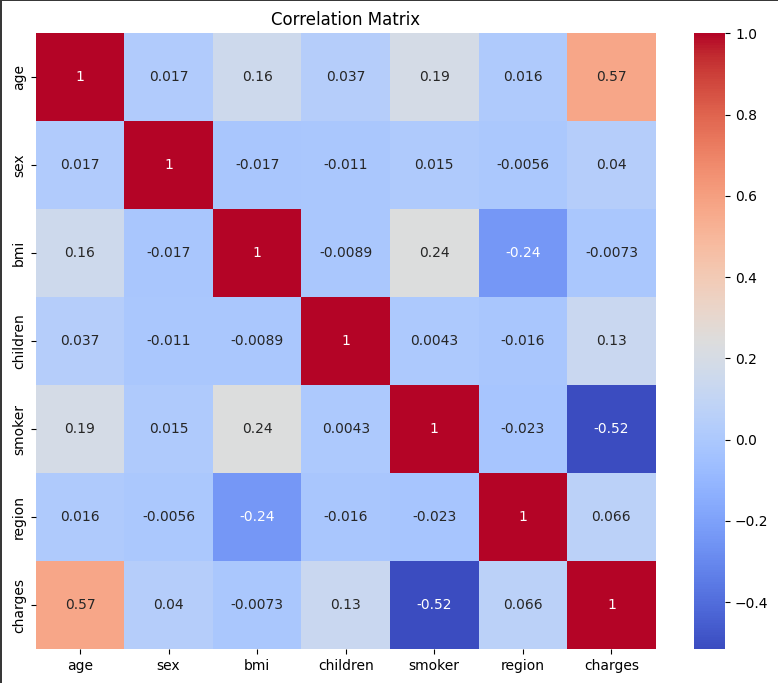
**Fig 1:** Dataset



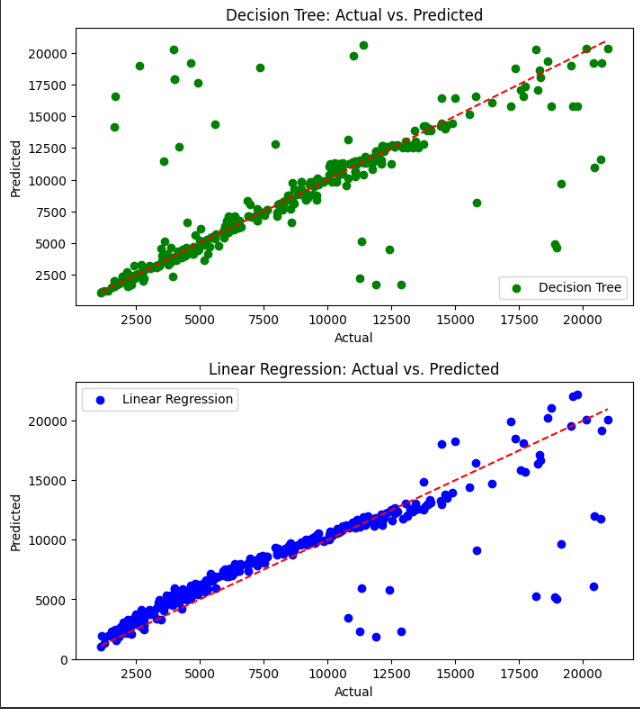
**Fig 2.1**: Dataset Analysis

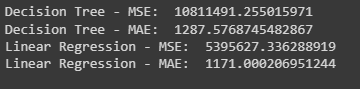
**Fig 2.2**: Data Visualization

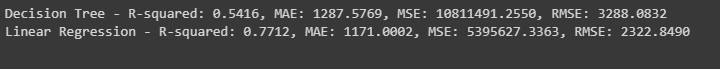


**Fig 3**: Correlation Matrix



**Fig 4**: Result Analysis





**Fig 5:** Accuracy Analysis