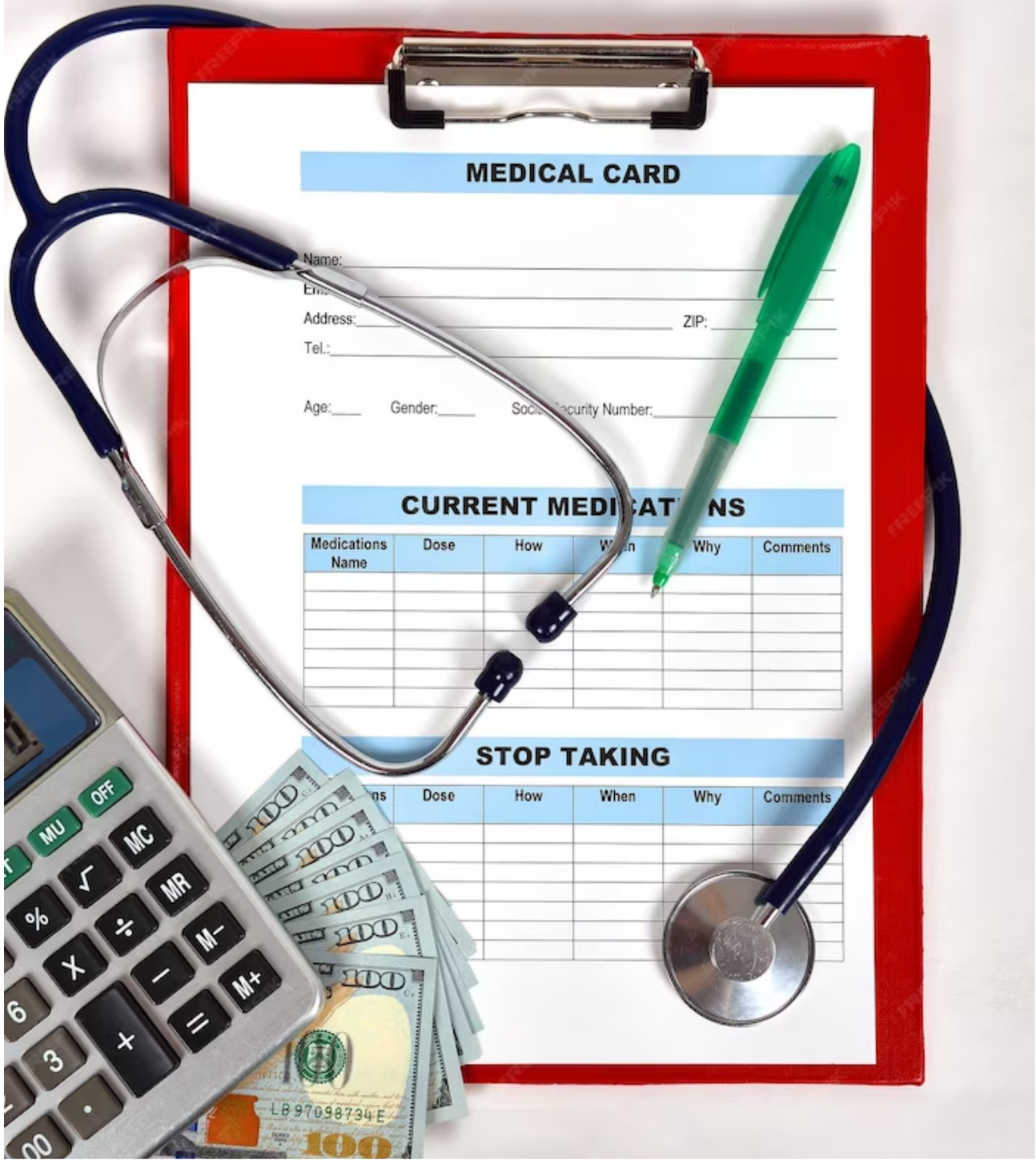
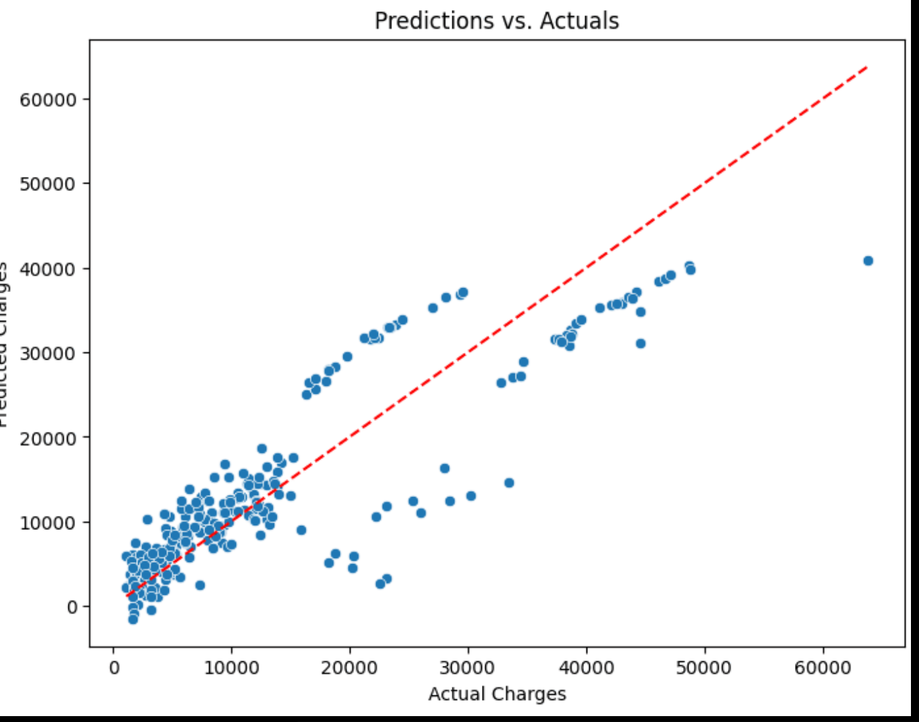
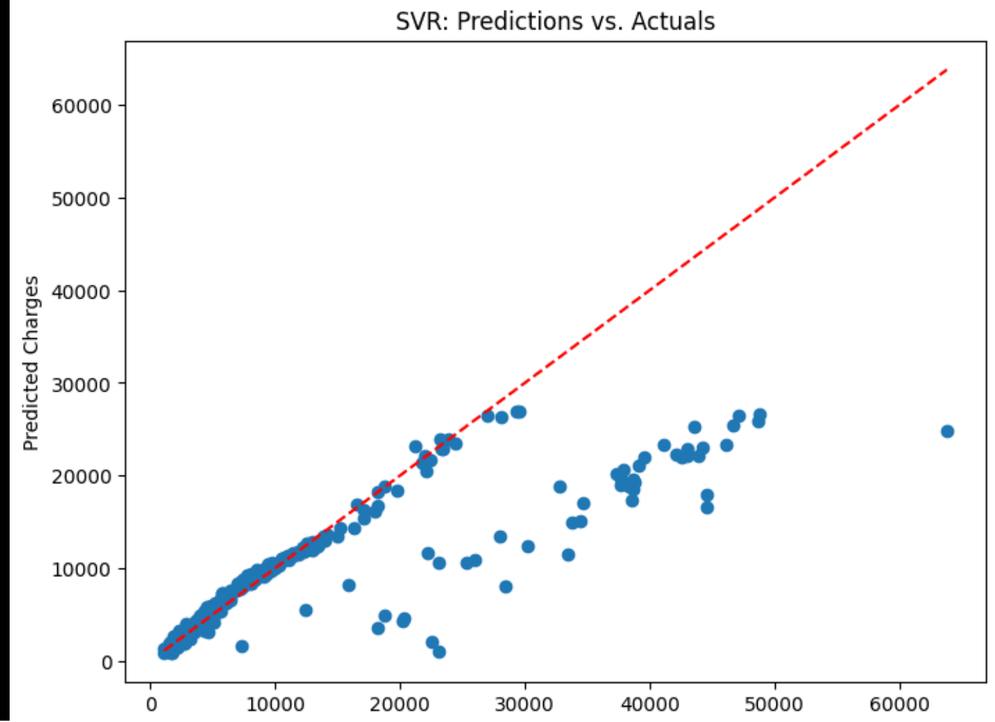
Unveiling Insights: A Data-driven Analysis of Health Insurance Charges through Predictive Modeling using Linear Regression and Support Vector Regression

Health Insurance Charges

Understanding the factors influencing **health insurance charges** is crucial for informed decision-making. This analysis aims to provide valuable insights into the complex nature of these charges.





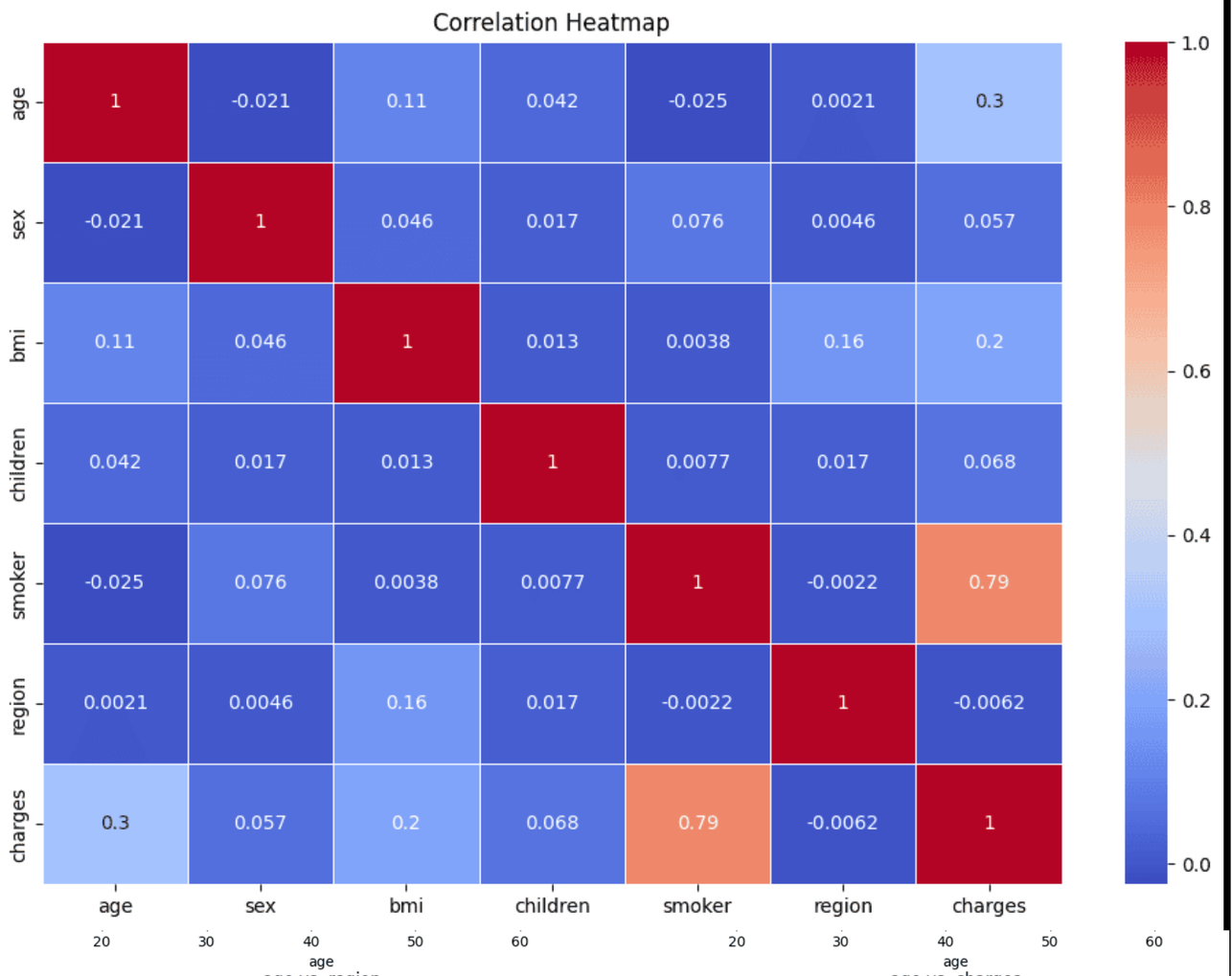


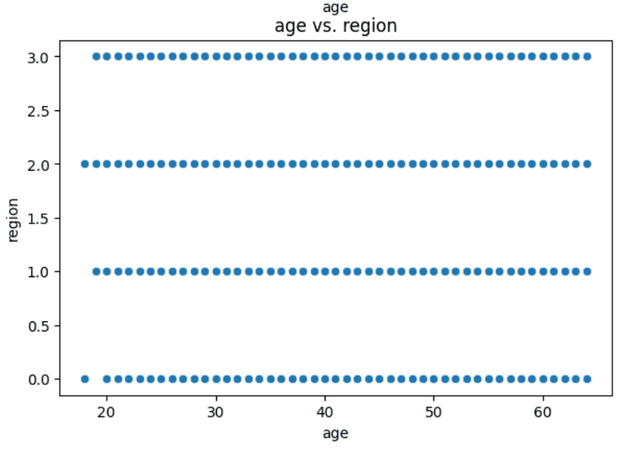
Predictive Modeling

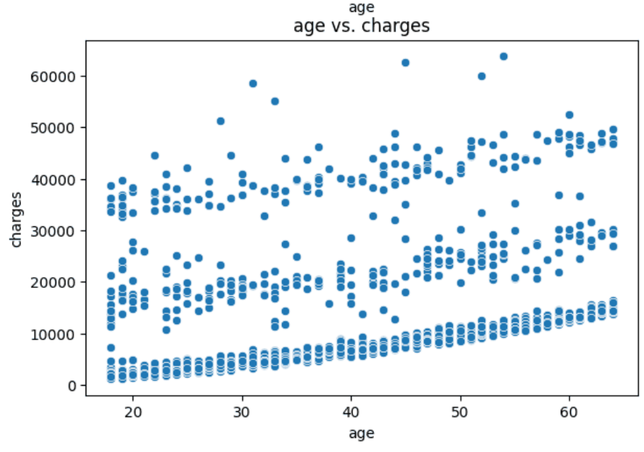
Linear Regression:
Linear Regression is a fundamental machine learning technique used to establish relationships between independent variables and a target variable. It works by fitting a linear equation to the data, allowing us to understand how changes in the independent variables affect the target variable. In the context of predicting health insurance charges, Linear Regression helps us quantify the impact of factors like age, BMI, and smoking status on the final insurance cost. It's particularly useful for uncovering linear associations between variables.Support Vector Regression (SVR):
Support Vector Regression is a versatile regression technique that can capture both linear and non-linear patterns in data. Unlike traditional Linear Regression, SVR doesn't assume a strict linear relationship between variables. It's well-suited for complex datasets where the relationship between variables may not be straightforward. In the context of health insurance charges, SVR can handle intricate relationships, accommodating a wide range of data structures and complexities. It's especially valuable when dealing with non-linear patterns and outliers in the data.

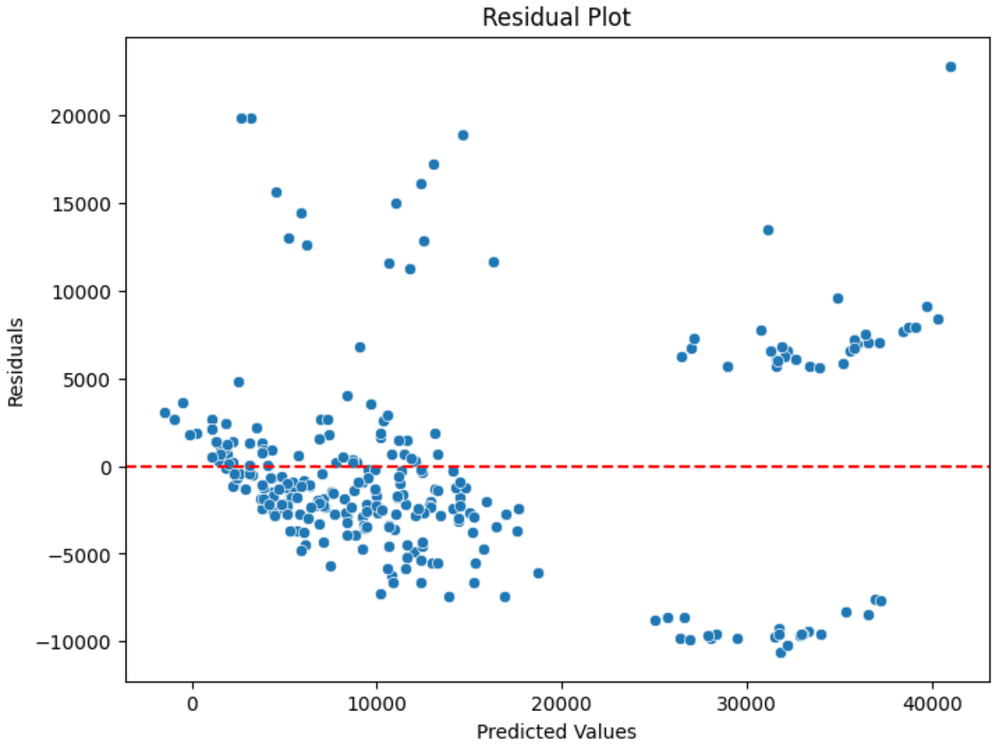
Data-driven Analysis

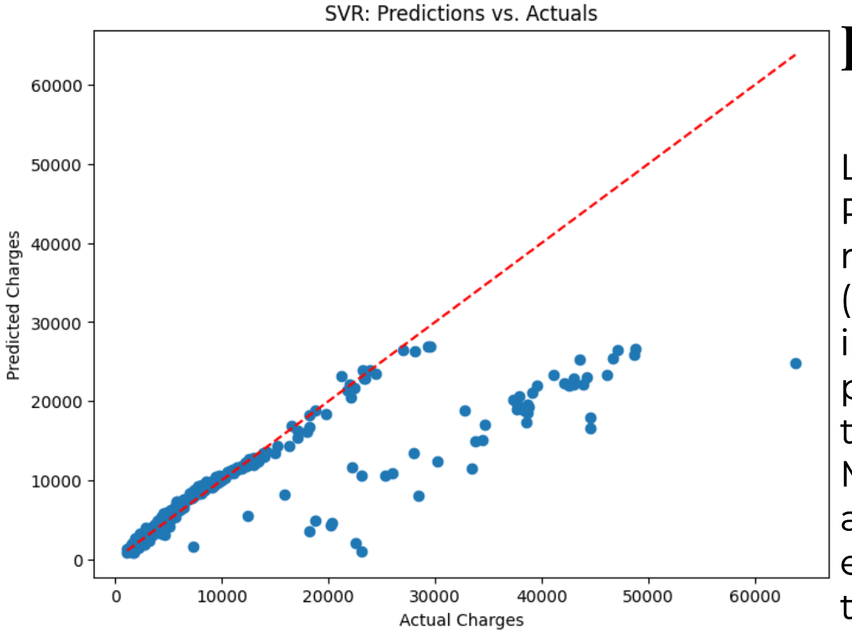
Age (0.299): Age has a moderate positive correlation with insurance charges. Older individuals tend to have higher insurance costs, likely due to increased health risks with age.Smoker (0.787): Smoking status exhibits a strong positive correlation with insurance charges. Smokers face significantly higher insurance costs, reflecting the health risks associated with smoking.BMI (0.198): BMI is moderately positively correlated with insurance charges. Higher BMI values are linked to increased insurance expenses, possibly due to obesity-related health risks.Children (0.068): The number of children has a weak positive correlation with insurance charges. Individuals with more children may have slightly higher costs, likely due to family coverage.Sex (0.057): Sex shows a weak positive correlation. Males, on average, may have slightly higher insurance charges than females, but the correlation is not strong.Region (-0.006): The region of residence has almost no significant correlation with insurance charges, indicating that location has minimal impact on costs.In summary, smoking status has the most substantial influence on insurance charges, followed by age and BMI. Other factors like sex, number of children, and region have weaker correlations, suggesting less direct impact on insurance costs.

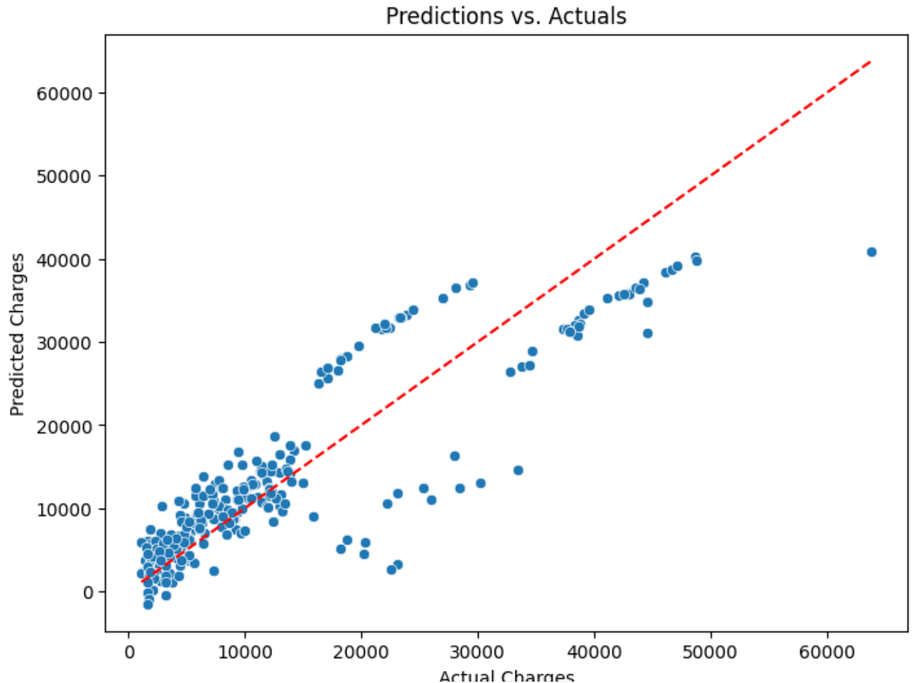


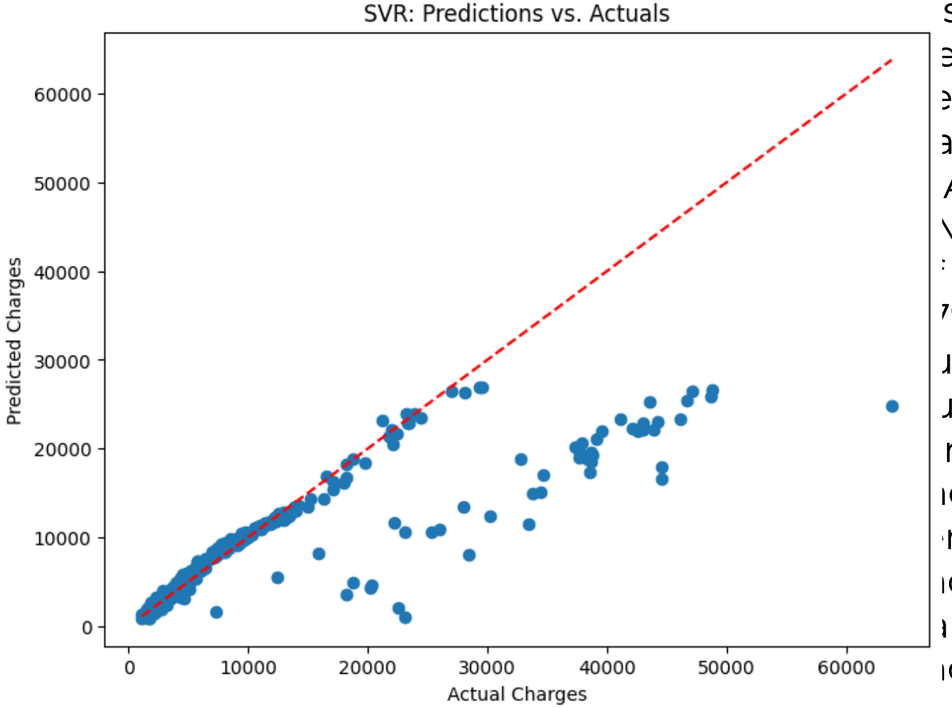








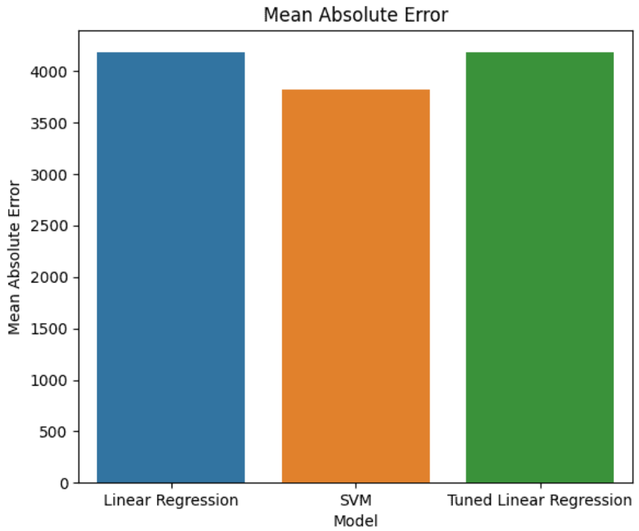


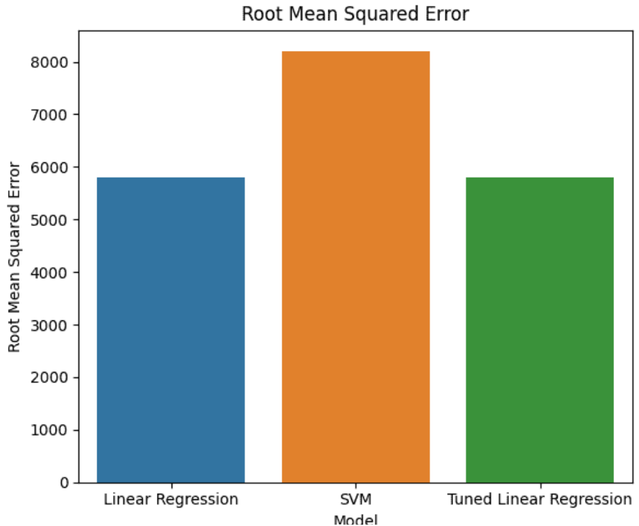


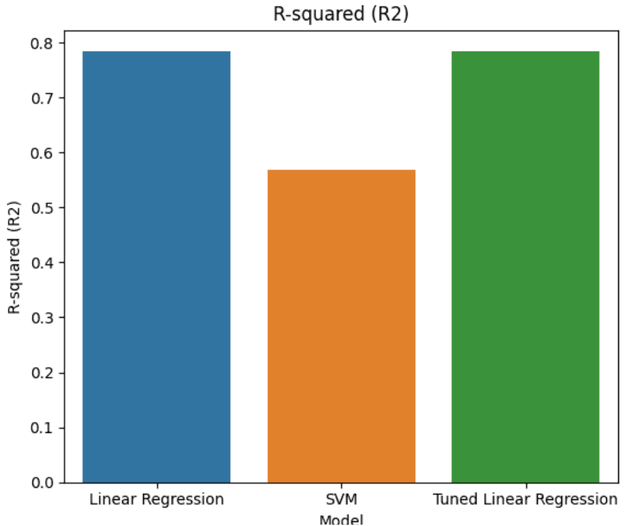
Results

Linear Regression: Our Linear Regression model showed promising results. The Mean Absolute Error (MAE) was approximately 4187, indicating that, on average, our predictions were off by this amount in terms of insurance charges. The Root Mean Squared Error (RMSE) was around 5800, reflecting the spread of errors in our predictions. Moreover, the R-squared (R2) value of 0.78 demonstrated that our model explained 78% of the variance in surance charges.Support Vector egression (SVR): SVR, known for its exibility in handling complex atterns, achieved decent results. The AE was approximately 3826, and the MSE was around 8189. The R2 value 0.57 indicated that SVR explained % of the variance.Hyperparameter uning for Linear Regression: urprisingly, hyperparameter tuning r Linear Regression didn't improve ıe model's performance. The results mained consistent with the initial odel, with MAE, RMSE, and R2 alues identical to the untuned odel.

Fit or Unfit







Linear Regression outperformed Support Vector Regression (SVR) in predicting health insurance charges. While Linear Regression achieved an R-squared (R2) value of 0.78, indicating that 78% of the variance in charges was explained by the model, SVR lagged behind with an R2 of 0.57. Additionally, Linear Regression had a lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), suggesting more accurate predictions. Despite hyperparameter tuning, Linear Regression maintained its edge, with tuned and untuned models yielding identical results. Therefore, Linear Regression stands as the preferred choice for this insurance charge prediction task due to its superior performance.

Model Suitability

Linear Regression suits the data well, demonstrating good predictive performance with an R-squared value of 0.78. In contrast, Support Vector Regression (SVR) is less suitable for this dataset, as it lags with an R-squared value of 0.57.

Conclusion

In conclusion, our analysis of health insurance charges has revealed important insights into the factors influencing insurance costs. Linear Regression and Support Vector Regression (SVR) were employed to model and predict these charges. While Linear Regression exhibited a strong performance with an R-squared value of 0.78, indicating its suitability for this dataset, SVR had a lower R-squared value of 0.57, suggesting a less favorable fit. To enhance predictive accuracy, hyperparameter tuning was applied to Linear Regression, yielding consistent results. These findings emphasize the significance of model selection and parameter optimization in achieving accurate predictions for health insurance charges. Ultimately, this study underscores the potential for data-driven insights to inform decision-making within the insurance industry and beyond.