**Medical Insurance Cost Prediction**

**1. Introduction**

Medical insurance is increasingly becoming essential, and insurance companies face challenges estimating individual medical expenses due to varied health conditions. This task focuses on using machine learning techniques to predict medical costs based on a dataset of patient demographics and health-related factors. The dataset includes features such as age, sex, BMI, number of children, smoking status, region, and medical costs.

**2. Objectives**

The primary objectives of this analysis are:

1. **To explore the dataset** and understand the underlying patterns and distributions of variables.
2. **To study correlations** between predictors (e.g., smoking status, BMI, age) and medical costs.
3. **To identify the most significant predictors** influencing medical costs.
4. **To build regression models** (simple and multivariate) to predict medical costs based on selected predictors.
5. **To evaluate and compare the performance** of models using statistical measures.
6. **To provide actionable insights** for insurance companies regarding the key factors driving medical costs.

**3. Data Cleaning**

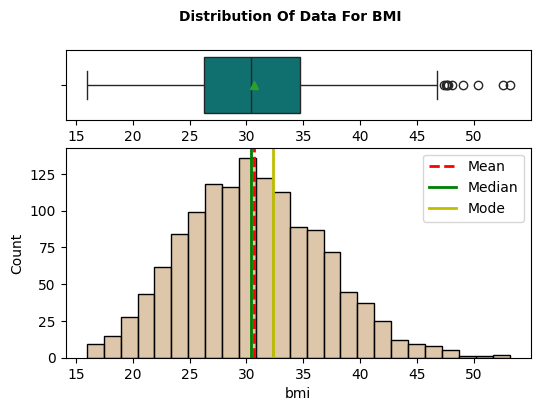
Before proceeding to data exploration, the dataset was examined for missing values and duplicates. Key findings and actions include:

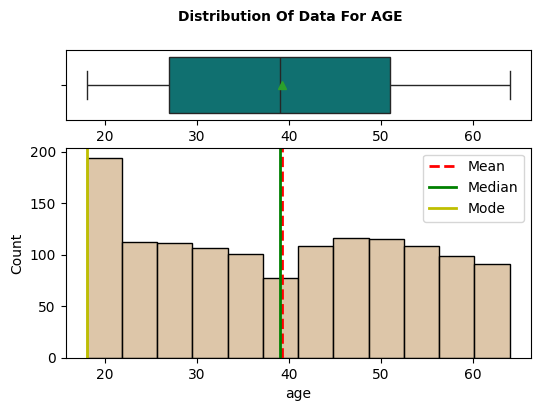
* **Missing Values:** The dataset contained no missing values across all columns.
* **Duplicate Records:** One duplicate row was identified and removed, reducing the dataset to 1337 rows.
* **Data Types:** Data types were validated to ensure proper handling during analysis. Categorical variables (e.g., sex, smoker, and region) were retained as objects for initial exploration and later encoded for modeling.

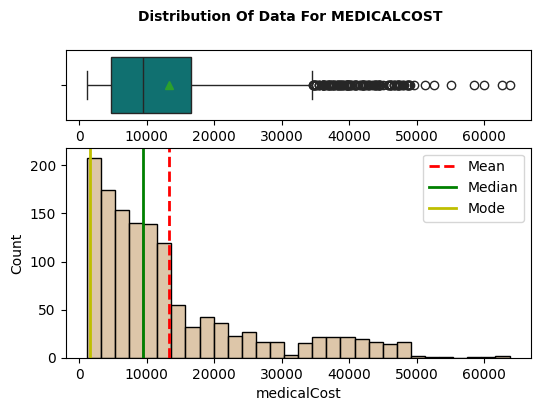
**4. Data Exploration**

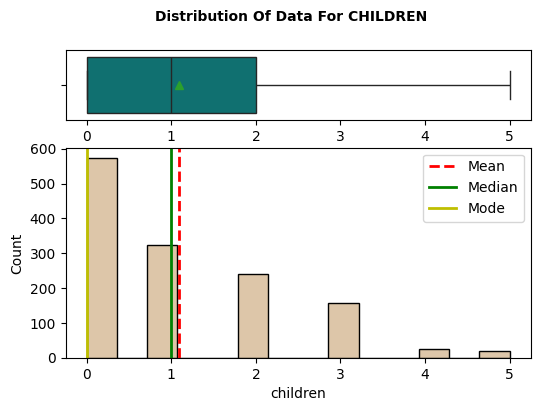
The average age of the individuals is 39.22 years, with a maximum age of 64 years. The average BMI is 30.66, which is outside the normal BMI range, with maximum and minimum BMIs of 53.13 and 15.96, respectively. The average medical cost is $13,279.12

The median medical cost is less than the mean, indicating a positively skewed distribution. On average, customers have one child. For age, BMI, and number of children, the mean is almost equal to the median, suggesting that the data is normally distributed.







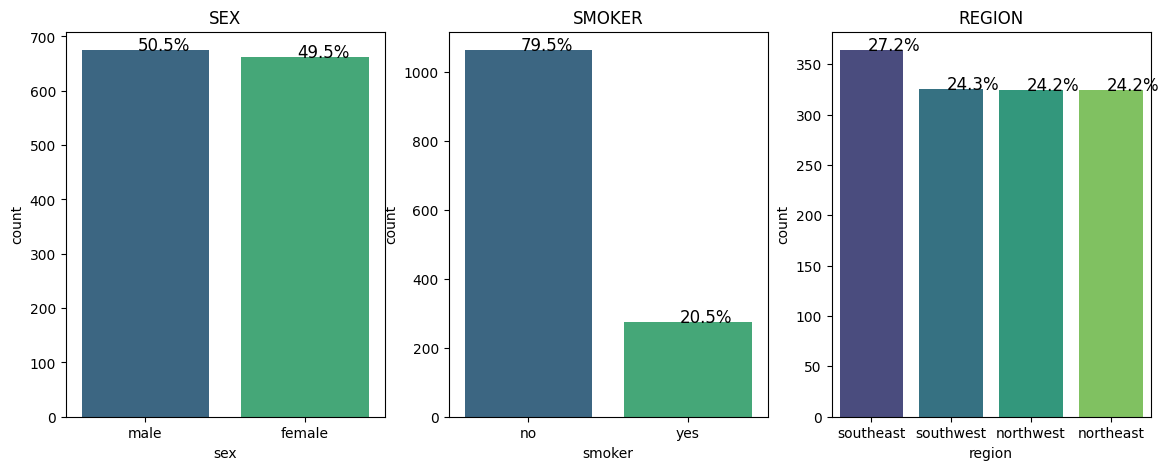


The primary beneficiaries' ages range from 20 to 65 years, with an average of around 40 years. The majority of customers are in their late teens and early twenties. The box plot shows a symmetric distribution. The histogram shows a relatively uniform age distribution and a higher concentration in the younger age group (20-30 years old).

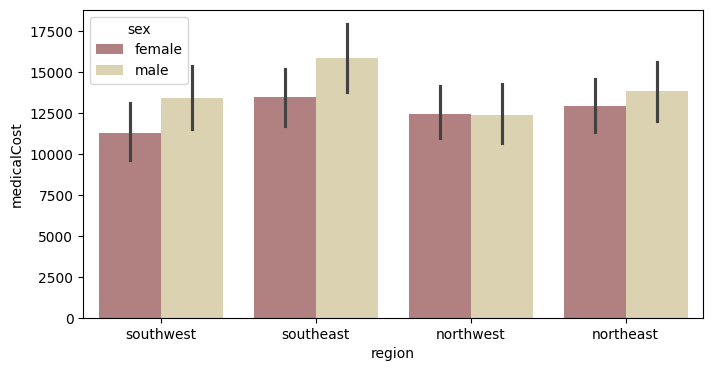
The beneficiaries' BMI is normally distributed, with an average of 30, which is outside the normal BMI range. The box plot shows a fairly symmetric distribution, with a median of around 30 and a mean that is closely aligned with it. The interquartile range (IQR) ranges from about 25 to 35, indicating that the majority of people have a BMI in this range, though there are a few outliers with BMIs above 40.

According to the distribution of data for children the median number of children is 1, with the middle 50% of data falling between 0 and 2 children. The histogram shows that the mode is 0, with approximately 600 observations. The next highest frequency is for families with 1 child. This chart suggests that most families in the dataset have no children or only one child, indicating a trend towards smaller family sizes.

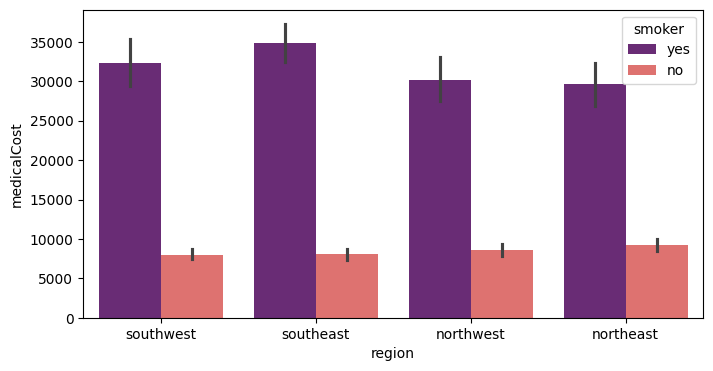
The distribution of medical costs is unimodal and right-skewed, with the average cost incurred by insurance around 13,000 and the highest charge reaching 63,770. There are many outliers at the upper end. The box plot shows that the median medical cost is approximately 9,000, with the mean slightly higher than the median. Numerous outliers exceed 30,000, with some reaching 60,000. The histogram supports the right−skewed distribution, with most data points falling below 10,000. The mode is below both the mean and the median.



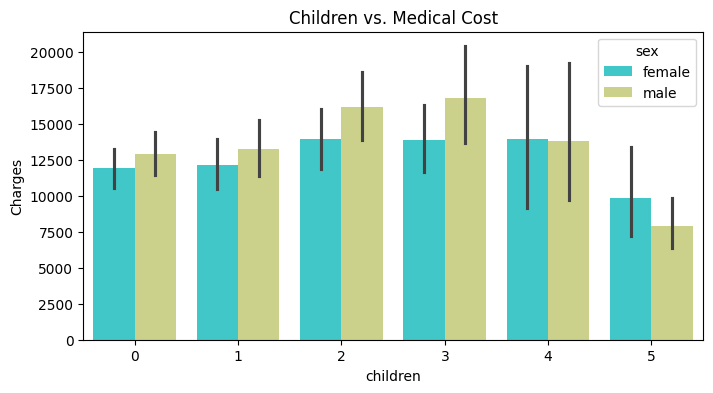
The visualizations show the distribution of three categorical variables from the dataset: sex, smoker status, and region. The first chart shows a nearly equal gender distribution, with 50.5% male and 49.5% female beneficiaries. The second graph shows that 79.5% of the beneficiaries are nonsmokers, while 20.5% smoke. The third chart depicts the regional distribution of beneficiaries, with the Southeast having the highest percentage (27.2%), followed by the Southwest, Northwest, and Northeast regions, each with approximately 24% of the beneficiaries.



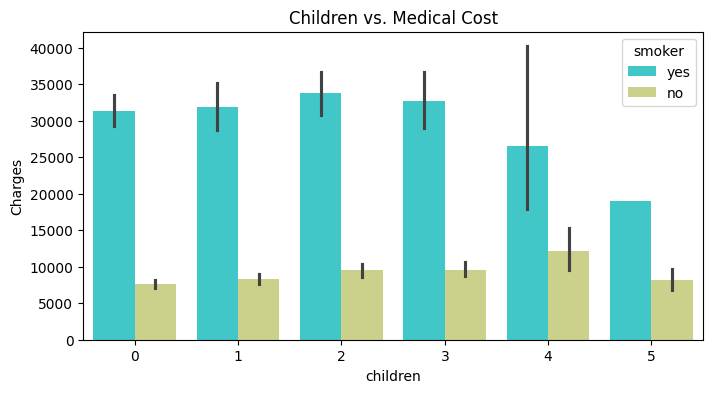
The bar chart depicts the medical costs for men and women across various regions. In the Southeast, men have the highest average medical costs, followed by women. The Southwest has comparable costs for both genders, with slightly higher costs for males. In the Northwest, medical costs are roughly equal for both genders. Overall, the Southeast region has the highest average costs for both sexes, while the other regions are more balanced or slightly lower.



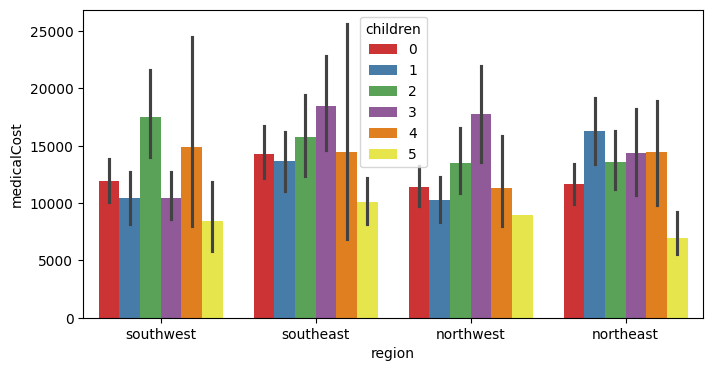
The chart shows the average medical costs for smokers and non-smokers across different regions. For non-smokers, the medical costs are relatively consistent across all regions. However, for smokers, there is significant variation. The Southeast region has the highest medical costs for smokers, while the Northwest and Northeast regions have nearly identical average medical costs for smokers



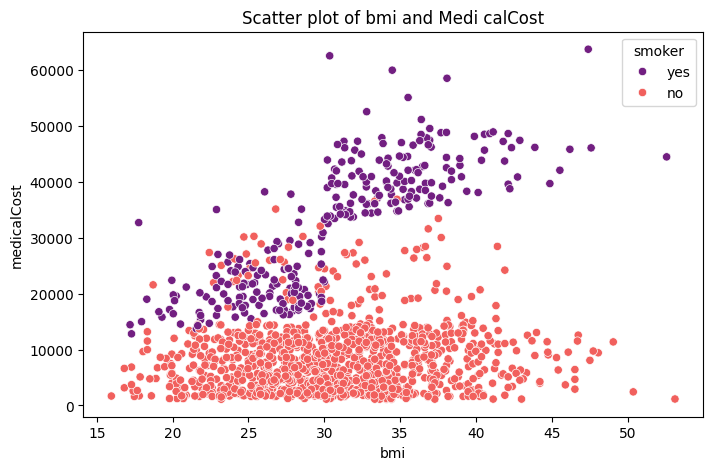
The chart illustrates the medical costs for males and females based on the number of children they have. Males with three children show the highest medical costs, while males with five children have the lowest medical costs. Additionally, individuals with no children incur higher medical costs than those with five children.

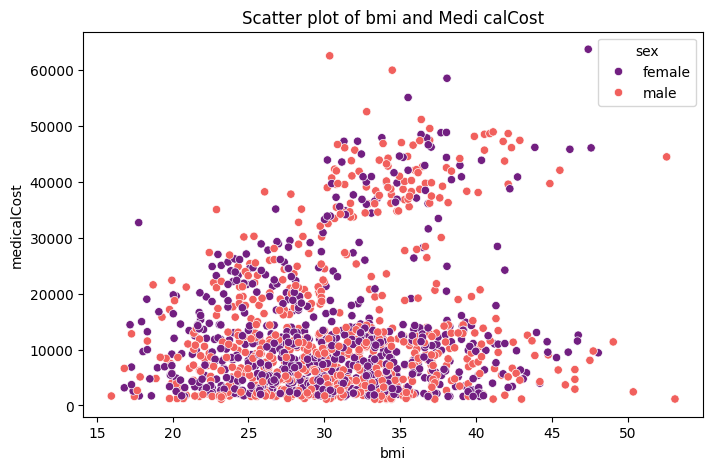


The chart depicts the medical costs associated with smokers and non-smokers based on the number of children they have. Smokers consistently incur significantly higher medical costs compared to non-smokers. The graph indicates that people with 5 childrens smokes less than others.



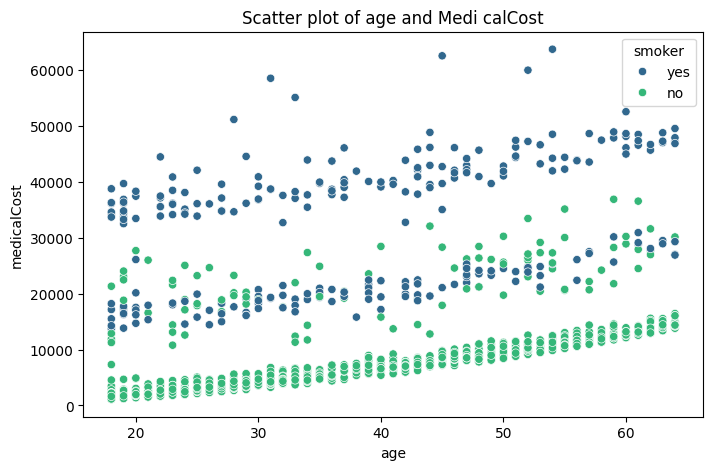
The chart depicts the average medical costs in various regions, categorized by the number of children. Medical costs vary significantly by region and number of children. Generally, individuals with no children tend to have lower medical costs compared to those with children. The Southeast region exhibits the greatest variability, with costs peaking for those with one or three children. In contrast, the Southwest, Northwest, and Northeast regions show a more consistent pattern, though medical costs continue to rise with the number of children. This suggests that medical costs are influenced by both the region and the number of children.

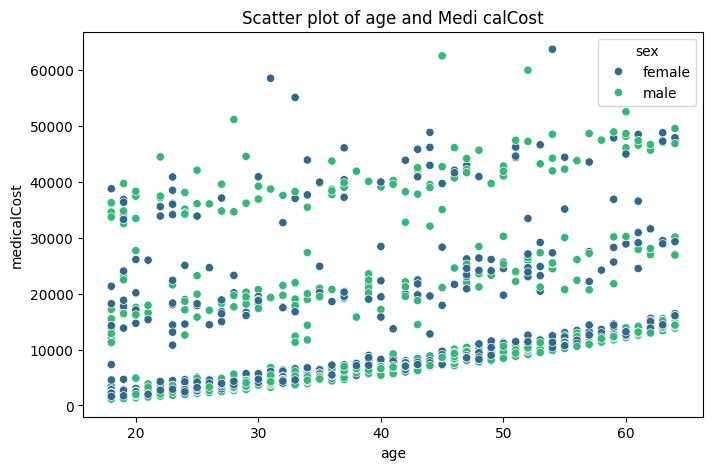




The scatter plot shows the relationship between BMI and medical costs, differentiated by smoking status. Smokers, represented by purple dots, generally have higher medical costs compared to non-smokers, represented by red dots. As BMI increases, medical costs tend to rise for both groups, but the increase is more pronounced for smokers. The plot clearly indicates that smoking significantly elevates medical expenses, particularly at higher BMI levels, with smokers incurring medical costs that can exceed $60,000, while non-smokers' costs rarely reach such high levels.

The scatter plot depicts the relationship between BMI and medical costs, broken down by gender. Males are represented by red dots, while females are depicted in purple. The graph shows that medical costs tend to rise with higher BMI for both sexes. Medical costs for males and females across different BMI levels are broadly similar, indicating no significant sex-based cost disparity. However, both groups contain outliers with extremely high medical costs, indicating that other factors may also influence medical expenses.

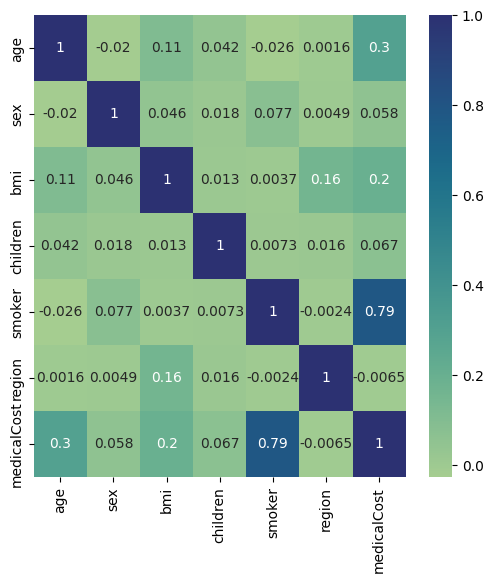




The scatter plot shows a clear relationship between age and medical costs, which is significantly influenced by smoking status. Medical costs tend to rise as people age, with smokers consistently incurring higher expenses than nonsmokers. The disparity in medical costs between smokers and nonsmokers grows with age, highlighting smoking's significant impact on healthcare costs. Overall, the plot highlights the financial burden associated with smoking, especially as people get older.

The scatter plot depicts the relationship between age and medical costs, with a distinction between male and female patients. Medical costs tend to rise with age for both sexes. There is no significant difference in medical costs between men and women; both sexes have similar distributions across age groups. This suggests that, while age is a major factor in rising medical costs, sex does not appear to have a significant impact on total medical costs. The consistent pattern suggests that age-related healthcare costs rise equally for men and women.

**5. Correlation Analysis**



The correlation between age and medical costs is moderately positive (0.3), indicating that medical costs tend to rise with age. BMI also has a moderate positive correlation (0.20) with medical costs, implying that higher BMI is associated with higher medical expenses. Smoking status has the strongest correlation with medical costs (0.79), indicating a strong positive relationship in which smokers have significantly higher medical costs than non-smokers. The correlation between the number of children and medical costs is very weak (0.068), indicating a negligible impact. The correlation between sex and medical costs is negligible, with a value of 0.057.

**7. Simple Linear Regression Models**

The model with smoking status as shows a strong performance with an R-squared value of 0.620. This indicates that smoking status explains 62% of the variance in medical costs, which is substantial. The F-statistic is extremely high at 2178, with a very low p-value (8.27e-283), indicating that the model is statistically significant. The coefficient for smokers is 23620.000, indicating that smoking increases medical costs by $23,620, with a high t-value (46.665) and a p-value of 0.000, confirming the predictor's statistical significance.

In contrast, the model with BMI as a predictor has a much lower R-squared value of 0.039, indicating that BMI explains only 3.9% of the variance in medical costs. While the F-statistic of 54.71 and the corresponding p-value (2.46e-13) indicate that the model is statistically significant, the low R-squared value suggests that it has limited explanatory power. The coefficient for BMI is 393.873, which means that each unit increase in BMI results in an increase of approximately $394 in medical costs, which is statistically significant given the t-value of 7.397 and p-value of 0.000.

Finally, the model with age as a predictor has an R-squared value of 0.089, indicating that age accounts for 8.9% of the variance in medical costs, which is higher than BMI but still relatively modest. The F-statistic is 131.2, with a p-value of 4.89e-29, indicating statistical significance. The coefficient for age is 257.723, which means that every additional year of age results in a $258 increase in medical costs. This coefficient is statistically significant, as evidenced by the t-value of 11.453 and p-value of 0.000.

In conclusion, while all three models are statistically significant, the model with smoking status as a predictor has the highest explanatory power (R-squared of 0.620) and the greatest effect on medical costs, followed by age and BMI. These findings indicate that smoking status is the most important factor in explaining the variation in medical costs among the variables studied, with age and BMI having smaller but still significant effects.

**8. Multivariate Regression Models**

* Model 1 (Three Predictors - Smoking, Age, BMI):
  + R² = 0.7777, MSE = 34,512,843.88.
  + Smoking contributed most significantly, followed by BMI and age.
* Model 2 (All Predictors):
  + R² = 0.1618, MSE = 130,132,666.29.
  + Lower performance likely due to overfitting and multicollinearity.

The first model, with a Mean Squared Error (MSE) of 34,512,843.88 and an R-squared (R^2) score of 0.7777, shows a strong fit and reliable predictive power, explaining approximately 77.77% of the variance in medical costs. This suggests that smoking status, BMI, and age are all important and effective predictors of medical costs.

In comparison, the second model, which includes all available predictors, performs significantly worse. The model has a significantly higher MSE of 130,132,666.29 but a lower R^2 score of 0.1618, explaining only 16.18% of the variance in medical costs. Despite the addition of variables such as gender, number of children, and region, the model's predictive power is reduced, most likely due to overfitting or multicollinearity among predictors. This poor performance indicates that the additional predictors added noise rather than improving the model's accuracy.

To summarize, the first model, which only considers smokers, BMI, and age, not only outperforms the others but also provides a better understanding of the key factors influencing medical costs. This comparison emphasizes the importance of careful feature selection in regression modeling. Adding more predictors does not always improve model performance and can result in overfitting, in which the model learns the noise rather than the underlying pattern. To address potential issues with the second model, techniques such as regularization or additional data preprocessing could be used to improve predictive accuracy and interpretability.

9. Overall Conclusions

The task involved predicting medical costs using a supervised learning regression analysis. Key findings included significant correlations between medical costs and smoking status, age, and BMI, with smoking being the most influential predictor. Simple linear regression models confirmed that these predictors had a significant impact on costs, with smoking status accounting for 62% of the variance. A multivariate model with these three predictors explained 77.8% of the variance, compared to a model with all predictors that explained only 16.2%. This emphasizes the importance of feature selection, with a focus on impactful predictors for model simplicity and efficiency, and suggests that smoking status, age, and BMI are critical for accurate cost prediction in health insurance.