Title: Predictive Analysis of Medical Insurance Charges Using Polynomial Regression and Machine Learning Models

---

1. Introduction

This study explores the use of regression modeling and machine learning techniques to predict medical insurance charges based on personal and lifestyle factors. The dataset used, titled insurance.csv, comprises several predictor variables that are either numerical or categorical in nature. The numerical variables include age, body mass index (BMI), number of children, and the actual insurance charges. The categorical variables include sex, smoking status, and region. These variables serve as explanatory inputs to understand variations in healthcare premiums.

To prepare the data for modeling, categorical variables such as sex, smoker, and region were converted to categorical data types. This allowed for appropriate encoding and interpretation during model training. Additionally, polynomial transformations were applied to the numerical predictors to account for non-linear relationships between features and the response variable. The analytical process relied on several Python libraries, namely pandas and numpy for data processing, seaborn and matplotlib for visualization, and scikit-learn for modeling. Various regression techniques were employed, including linear regression, ridge regression, lasso regression, random forest regression, and polynomial regression, to compare performance and interpretability.

---

2. Findings and Graphical Analysis

The initial visual analysis focused on the distribution of the charges variable. A kernel density plot showed that charges were positively skewed, indicating a small number of individuals incur extremely high medical costs. A logarithmic transformation was applied to correct for this skewness, which resulted in a more symmetric distribution, suggesting this approach helps normalize the data and stabilize variance.

Subsequent visualizations analyzed the influence of geographic region on charges. Bar plots showed that individuals in the Southeast region incurred the highest total charges, and further disaggregation by sex and smoking status highlighted that male smokers in the Southeast were particularly costly to insure. When examining the number of children as a factor, the analysis revealed that while having more children did not consistently increase charges, the variability in charges did increase slightly among smokers, especially those with more children.

Three regression plots provided further insights into relationships between continuous variables and charges. A strong positive correlation was observed between age and charges, particularly among smokers. BMI also had a noticeable effect on charges, especially for those classified as smokers. The relationship between the number of children and charges was less direct, but once again, smoking status introduced significant variability in the data. Violin plots supported these findings, showing that for each level of children, the distribution of charges was wider for smokers, with significantly higher median values.

A heatmap of the correlation matrix indicated that smoking status had the highest positive correlation with charges, followed by age and BMI. This informed the feature selection strategy for modeling. Initial regression models using linear regression achieved an R^2 score of approximately

0.74 on the test data, indicating a moderate fit but failing to capture complex interactions. Ridge and lasso regression models, both using regularization, slightly improved generalization. The lasso model also introduced coefficient shrinkage, making it more interpretable.

The random forest model significantly improved predictive performance, achieving an R^2 score of around 0.86. Feature importance rankings from the random forest highlighted smoker status, age, and BMI as the most significant predictors. Finally, polynomial regression of degree two was implemented to capture non-linear relationships among variables. This model achieved the highest performance, with an R^2 score close to 0.89 on the test data. Evaluation metrics showed moderate mean absolute error and a reasonable root mean squared error, suggesting good predictive accuracy. A residual plot of predicted versus actual values showed that the residuals were centered around zero, indicating that the polynomial model did not suffer from heteroscedasticity and produced reliable predictions.

---

3. Conclusion and Policy Recommendations

The analysis reveals that smoking status, age, and BMI are the most influential factors in determining medical insurance charges. These insights have meaningful implications for healthcare policy and insurance pricing. One clear recommendation is to implement targeted programs to reduce smoking prevalence. As smoking is the strongest predictor of high medical charges, insurance providers and public health authorities should introduce incentives for quitting, such as reduced premiums or subsidized cessation programs.

Additionally, since regional disparities exist with the Southeast bearing higher costs, there is a case for creating region-specific premium models that reflect the underlying health risk patterns more accurately. Moreover, promoting healthy lifestyles to manage BMI through employer-sponsored wellness programs or government-supported fitness incentives could contribute to long-term cost reduction. Although the number of children did not significantly influence charges, there is still merit in considering family composition when designing insurance plans, especially to support families with dependent children.

Overall, the use of polynomial regression and ensemble methods like random forest offers a robust framework for predicting insurance costs and can inform more equitable and data-driven insurance policies. The findings stress the importance of preventive health behaviors and region-sensitive policy interventions in managing healthcare expenditures effectively.