Data Analysis For Insurance Company

# Introduction

# This report provides a detailed analysis of the insurance dataset, focusing on predicting insurance charges based on various features. The analysis includes data exploration, data visualization, feature transformation, model training, hyperparameter tuning, and model evaluation. We build several regression models to predict insurance charges, including Linear Regression, Support Vector Regression (SVR), Ridge Regression, and Random Forest Regression. Hyperparameter tuning is performed to optimize the SVR model. The performance of all models is compared using R² and RMSE metrics. The aim is to predict insurance charges based on factors such as age, BMI, number of children, gender, and smoking.

# 1. Exploring the Dataset

We begin by loading the dataset and examining its structure, including column names, data types, summary statistics, and checking for null values. This step provides a comprehensive understanding of the dataset's content and quality.

The dataset consists of the following columns: age, sex, bmi, children, smoker, region, charges.

Top 10 Rows from the dataset are:

age sex bmi children smoker region charges

0 19 female 27.900 0 yes southwest 16884.92400

1 18 male 33.770 1 no southeast 1725.55230

2 28 male 33.000 3 no southeast 4449.46200

3 33 male 22.705 0 no northwest 21984.47061

4 32 male 28.880 0 no northwest 3866.85520

5 31 female 25.740 0 no southeast 3756.62160

6 46 female 33.440 1 no southeast 8240.58960

7 37 female 27.740 3 no northwest 7281.50560

8 37 male 29.830 2 no northeast 6406.41070

9 60 female 25.840 0 no northwest 28923.13692

# 2. Converting Categorical Values to Numerical

Categorical features such as sex, smoker status, and region are encoded into numerical values using Label Encoding.

This transformation is essential for modeling, as most machine learning algorithms require numerical input.

# 3. Plotting Heatmap to See Dependency of Dependent Value on Independent Features

A correlation heatmap is plotted to visualize the relationships between features. This helps identify which features are most strongly correlated with the target variable (insurance charges).

# 4. Data Visualization

Several plots are created to visualize the data:  
- Pairplot: Shows relationships between features, colored by smoking status.  
- Scatterplots with regression lines: Illustrate the relationships between age, BMI, and insurance charges, differentiated by smoking status.  
- Bar plot: Displays the number of smokers in each region.

# 5. Plotting Skew and Kurtosis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | age | bmi | children | charges |
| skew | 0.055673 | 0.284047 | 0.938380 | 1.515880 |
| kurtosis | -1.245088 | -0.050732 | 0.202454 | 1.606299 |

Skewness and kurtosis are statistical measures that help to understand the distribution and shape of the data.  
Skewness measures the asymmetry of the data distribution. A skewness value close to zero indicates that the data is fairly symmetrical. Positive skewness indicates a right-skewed distribution, while negative skewness indicates a left-skewed distribution.  
Kurtosis measures the 'tailedness' of the data distribution. A kurtosis value close to zero indicates a distribution similar to a normal distribution. Positive kurtosis indicates heavier tails and sharper peak, while negative kurtosis indicates lighter tails and flatter peak.  
In this dataset, the 'charges' feature shows the highest skewness and kurtosis, indicating a distribution with a longer tail and sharper peak compared to a normal distribution.

# 6. Data Preparation

The dataset is split into training and testing sets. Features are scaled using StandardScaler to standardize the data. Polynomial features are added to capture non-linear relationships, and hyperparameter tuning is performed for the SVR model.

# 7. Prediction Using Linear Regression

Linear regression was used to predict the insurance charges. The model performance metrics are as follows:  
R^2: 0.7833463107364538  
RMSE: 5799.587091438356

# 8. Prediction using SVR

Support Vector Regression (SVR) was applied with hyperparameter tuning using RandomizedSearchCV. The best parameters and model performance are as follows:  
Best parameters for SVR: {'kernel': 'linear', 'gamma': 'auto', 'C': 10}  
Best score for SVR: 0.6801514259688847  
Optimized SVR:  
R^2: 0.773681631587424  
RMSE: 5927.532337266519

# 9. Prediction using Ridge Regressor

Ridge regression was used to predict the insurance charges. The model performance metrics are as follows:  
R^2: 0.7830849126391299  
RMSE: 5803.084710238366

# 10. Prediction using Random Forest Regressor

Random Forest Regressor was used to predict the insurance charges. The model performance metrics are as follows:  
R^2: 0.863178803290698  
RMSE: 4608.830353821407

# 11. Performing Hyper tuning for above mentioned models

Hyperparameter tuning was performed for SVR using RandomizedSearchCV. The best parameters were found to be {'kernel': 'linear', 'gamma': 'auto', 'C': 10}, leading to improved model performance.

# 12. Plotting Graph for all Models to compare performance

The performance of different models was compared using R^2 and RMSE metrics. The following table summarizes the results:

|  |  |  |
| --- | --- | --- |
| Model | R^2 | RMSE |
| Linear Regression | 0.7833463107364538 | 5799.587091438356 |
| Ridge Regression | 0.7830849126391299 | 5803.084710238366 |
| Random Forest Regression | 0.863178803290698 | 4608.830353821407 |
| Optimized SVR | 0.773681631587424 | 5927.532337266519 |

Among the models, the Random Forest Regressor shows the highest R^2 and the lowest RMSE, indicating the best performance. The SVR model, after hyperparameter tuning, shows improved performance but still lags behind the other models in this dataset.

# Conclusion

This report presents a comprehensive analysis of the insurance dataset, from data exploration and preprocessing to model building and performance evaluation. The results indicate that different models have varying levels of predictive accuracy. Hyperparameter tuning significantly improves the performance of the SVR model. The comparison of R² and RMSE scores helps identify the best-performing models for predicting insurance charges.  
  
This detailed approach provides valuable insights and a robust framework for predicting insurance charges, which can be applied to similar datasets in the future.