**Insurance Charges Prediction Using Regression Models**

### ****1. Introduction****

In this analysis, we are working with a dataset containing information about individuals and their insurance charges. The dataset includes features such as age, sex, BMI, number of children, smoking status, region, and the target variable, charges (the insurance charges). Our objective is to predict the insurance charges based on these features using various regression models.

**2. Data Exploration & Preprocessing**

**Dataset Overview**

The dataset contains 1,338 records and 8 columns with the following features:

* **Age**: The age of the insured.
* **Sex**: Gender of the insured.
* **BMI**: Body Mass Index of the insured.
* **Children**: Number of children/dependents covered by the insurance.
* **Smoker**: Whether the individual is a smoker (True/False).
* **Region**: The region where the insured resides.
* **Charges**: The insurance charges (target variable).

**Handling Missing Data**

No missing values were detected in the dataset. All values for the variables are complete and accounted for.

**Feature Engineering**

Categorical variables such as sex, smoker, and region were converted into numerical format using **Label Encoding**:

* **sex**: Male = 0, Female = 1
* **smoker**: Non-Smoker = 0, Smoker = 1
* **region**: Encoded with integers corresponding to the different regions.

**3. Data Visualization**

Several visualizations were created to better understand the distribution and relationships between the features and the target variable (charges):

* **Distribution of Charges**: A histogram with a Kernel Density Estimation (KDE) was plotted to show the distribution of the insurance charges.
* **Feature Relationships**: Scatter plots and bar plots were used to explore the relationships between features like age, BMI, number of children, smoking status, and region with the insurance charges.

**4. Regression Models**

Several regression techniques were applied to predict the insurance charges:

**4.1. Linear Regression**

A **Linear Regression** model was applied to predict the insurance charges based on the features:

* The **R² score** for the test data was calculated to evaluate the model's performance. The **R² score** quantifies the proportion of variance in the target variable explained by the model.
* **Intercept** and **coefficients** were extracted to understand the linear relationship between each feature and the target.

**4.2. Ridge Regression**

Ridge Regression (a regularized version of Linear Regression) was applied with an alpha value of 0.5. The regularization helps to prevent overfitting by adding a penalty term for large coefficients.

* **Intercept** and **coefficients** were obtained, and the **R² score** was calculated.

**4.3. Lasso Regression**

**Lasso Regression** was applied with an alpha value of 0.2. Lasso is another regularized linear model that can also perform feature selection by shrinking less important feature coefficients to zero.

* **Intercept** and **coefficients** were examined.
* The **R² score** on the test set was calculated to measure model performance.

**4.4. Polynomial Regression**

**Polynomial Regression** with a degree of 2 was used to capture non-linear relationships between the features and the target variable.

* **Polynomial Features** were created, and the **Linear Regression** model was applied to the transformed features.
* The **R² score** was evaluated, and the model coefficients were extracted.

**4.5. Random Forest Regression**

**Random Forest Regression** was used as an ensemble method to predict the insurance charges. Random Forest works by averaging multiple decision trees to reduce variance and prevent overfitting.

* **Mean Squared Error (MSE)** and **R² score** were calculated for both the training and test sets to evaluate model performance.

**5. Model Comparison**

| **Model** | **R² Score (Training Data)** | **R² Score (Test Data)** |
| --- | --- | --- |
| Linear Regression | 0.743 | 0.740 |
| Ridge Regression | 0.743 | 0.740 |
| Lasso Regression | 0.742 | 0.739 |
| Polynomial Regression | 0.794 | 0.768 |
| Random Forest | 0.974 | 0.831 |

**Observations:**

* The **Random Forest Regression** model performed the best in terms of **R² score** on the test data (0.831). It is capable of capturing complex relationships in the data.
* The **Polynomial Regression** model also performed well, but not as well as Random Forest (0.768 on the test data).
* **Linear Regression**, **Ridge**, and **Lasso Regression** models had similar performance, with R² scores of around 0.74, indicating that they might not capture the non-linear relationships as effectively.

**6. Model Evaluation**

* The **Random Forest** model performed the best, achieving the highest **R² score** and demonstrating strong predictive power for insurance charges.
* The **Polynomial Regression** model showed a significant improvement over linear models by capturing some non-linear relationships between features and charges.
* **Linear Regression** and its variants (Ridge and Lasso) performed similarly, achieving decent results but unable to fully capture non-linearities in the data.

**7. Conclusion**

Based on the analysis, the **Random Forest Regressor** is the most suitable model for predicting insurance charges from the given dataset. It is able to handle complex relationships and interactions between features effectively. Polynomial Regression offers an improvement over linear models, but not as much as Random Forest.

For future improvements:

* **Hyperparameter tuning** for the Random Forest and Polynomial models could further improve model performance.
* **Feature engineering**: More domain-specific features or transformations could potentially lead to better model performance.