Report: Predicting Insurance Charges Using Regression Models

Project Overview:

The goal of this project was to predict insurance charges for individuals based on a variety of factors such as age, sex, BMI, children, smoking habits, and region. The dataset provided contained information about the demographic and health-related factors of individuals, and the objective was to build a predictive model that could forecast the charges a person might incur based on these factors.

---

1. Data Collection and Understanding

The dataset used in this project was sourced from an open insurance data repository. It included the following columns:

- **Age**: Age of the individual.
- **Sex**: Gender of the individual.
- **BMI**: Body Mass Index.
- **Children**: Number of children/dependents covered by the insurance.
- **Smoker**: Whether the individual smokes (Yes/No).
- **Region**: The region in which the individual resides.
- **Charges**: The insurance charges for the individual (target variable).

---

2. Data Preprocessing

The preprocessing steps were necessary to clean the dataset and make it suitable for modeling:

- **Handling Missing Values:** There were no missing values in the dataset, so no imputation was needed.
- **Outlier Detection:** Outliers in the BMI and Charges columns were detected and treated.
- **Categorical Variable Encoding:** The categorical variables such as Sex, Smoker, and Region were encoded using one-hot encoding.

---

3. Exploratory Data Analysis (EDA)

EDA was performed to uncover insights from the data:

- **Correlation Analysis:** A heatmap was generated to observe the correlation between various features and the target variable. It was found that smoking habits (Smoker) and BMI showed a strong positive correlation with the charges.
- **Visualizations:** Scatter plots and histograms were created to visualize relationships between features (like Age vs. Charges, BMI vs. Charges), highlighting trends that would be useful for prediction.

Report: Predicting Insurance Charges Using Regression Models

4. Feature Engineering

- **Interaction Features:** Created interaction terms such as Age * BMI to capture relationships between different features.
- **Feature Scaling:** The BMI and Age features were scaled using standardization, ensuring that they contributed equally to the model's performance.

5. Model Building

Several regression models were implemented and evaluated to predict the insurance charges:

- **Linear Regression:** A baseline model was built using linear regression, providing a simple yet effective prediction.
- **Lasso Regression:** Applied Lasso to regularize the model and reduce overfitting.
- **Random Forest Regression:** Used Random Forest to capture non-linear relationships and interactions between features.

6. Model Evaluation

The models were evaluated using the following metrics:

- **R-squared:** The linear regression model achieved an R-squared of 0.79, indicating that the model explained about 79% of the variance in the target variable.
- **Mean Absolute Error (MAE):** The Random Forest model had the lowest MAE, showing better accuracy in prediction compared to other models.
- **Cross-Validation:** Cross-validation was performed to ensure that the model generalizes well on unseen data.

7. Results and Insights

The Random Forest model was selected as the best model due to its superior predictive performance and ability to handle non-linear relationships in the data. Key insights include:

- **Smoking** significantly impacts insurance charges, with smokers having substantially higher charges compared to non-smokers.
- **BMI** also has a significant positive relationship with charges, as higher BMI typically correlates with higher health risks.

Tools Used:

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn
- **Modeling Algorithms:** Linear Regression, Lasso Regression, Random Forest

Keywords: Insurance Charges Prediction, Regression Models, Data Preprocessing, Feature Engineering, Exploratory Data Analysis, Machine Learning.