**Problem Statement: Insurance Charges Prediction Using Regression Models**

**Objective:**

The objective of this project is to develop a predictive model that accurately estimates **individual insurance charges** based on personal and demographic attributes. This involves building, training, and evaluating regression models — such as Support Vector Regression (SVR) — to minimize prediction errors and gain insights into the influence of different features on insurance costs.

**Background:**

Insurance companies use predictive modeling to estimate charges for policyholders based on various risk factors. Accurate prediction helps in pricing strategies, risk assessment, and customer segmentation. Traditional pricing mechanisms may not capture complex, non-linear relationships in the data — hence, machine learning approaches are explored.

**Dataset Features:**

The dataset includes the following variables:

* age: Age of the individual
* sex: Gender of the individual
* bmi: Body Mass Index (health risk factor)
* children: Number of children covered by insurance
* smoker: Smoking status (yes or no)
* region: Residential area within the country
* charges: Annual medical insurance charges (target variable)

**Approach:**

1. **Data Preprocessing:**
   * Encoding categorical variables (e.g., sex, smoker, region)
   * Feature scaling using StandardScaler for numerical stability, especially for SVR
2. **Model Training:**
   * Train an SVR model with an RBF kernel
   * Compare with other regressors like Decision Tree, Random Forest, and Gradient Boosting (if applicable)
3. **Model Evaluation:**  
   Evaluation is performed using the following metrics:
   * Mean Absolute Error (MAE)
   * Mean Squared Error (MSE)
   * Root Mean Squared Error (RMSE)
   * R² Score (Coefficient of Determination)
4. **Model Interpretation:**
   * Analyze feature importances or sensitivity (depending on the model)
   * Visualize prediction performance with actual vs. predicted plots

**Goal:**

To achieve a reliable and interpretable model that can predict insurance charges with high accuracy and generalize well to unseen data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **MSE** | **RMSE** | **R² Score** |
| **SVR (RBF)** | 2208.43 | 19,800,276.06 | 4449.75 | 0.87 |
| **Decision Tree** | 3159.00 | 43,555,735.99 | 6599.68 | 0.72 |
| **Random Forest** | 2550.59 | 22,091,517.12 | 4700.16 | 0.86 |
| **Gradient Boosting** | 2424.53 | 18,860,900.43 | 4342.91 | **0.88** |