**PROBLEM STATEMENT:**

**This is a regression problem where the goal is to predict insurance charges based on various parameters provided in the dataset. Typical features in such datasets include age, BMI, smoking status, number of dependents, etc.**

**Step 2: Dataset Information**

Total Rows: 1,338

Total Columns: 7

* Columns: age, sex, bmi, children, smoker and charges.

Initial Observations:

* Insurance charges is the target variable.
* Some columns (like sex, smoker) are categorical and need encoding.

**Step 3: Pre-processing Data:**

Handling Categorical Variables: Use one-hot encoding for sex & smoker

**4. Model Development**

**Train four models (Multiple Linear Regression, Support Vector Machine, Decision Tree and Random Forest) and evaluate using the R² score to choose the best model.**

**5. R² Scores**

| **Model** | **R² Score** |
| --- | --- |
| **Multiple Linear Regression** | **0.78** |
| **Random Forest** | **0.87** |
| **Decision Tree** | **0.78** |
| **Support Vector Regressor (SVR)** | **-0.88** |

**6. Final Model and Justification**

**The Random Forest model achieved the highest R² score (0.87) and is selected as the final & best model. The Random Forest model achieved the highest R² score among all the models tested. This makes it the best choice for predicting insurance charges in our case.**

**High Accuracy: Random Forest consistently outperformed other models with the highest R² score, indicating its ability to explain the variability in the data.**

1. **Non-linear Relationships: It can handle complex and non-linear patterns in the data, which are likely present in real-world scenarios like insurance charges.**
2. **Robustness: Random Forest reduces the risk of overfitting due to its ensemble nature (combining multiple decision trees), making it more reliable on unseen data.**

**For these reasons, the Random Forest model is selected as the final model.**