Insurance Charges Prediction Model

---

1. Problem Statement

The objective of this project is to build a predictive model that estimates the **insurance charges** for individuals based on certain demographic and health-related features. The goal is to help insurance companies accurately predict the charges for new customers based on their age, sex, BMI, number of children, smoking habits, and region.

Predicting insurance charges is important for companies in determining premiums, ensuring fair pricing, and managing risk. A model that can predict insurance charges based on historical data will help insurers in offering personalized policies and improving customer satisfaction.

---

2. Dataset Overview

The dataset used for this task includes the following features:

- **age**: Age of the individual (in years).
- **sex**: Gender of the individual (male/female).
- **bmi**: Body Mass Index (BMI) of the individual.
- **children**: Number of children or dependents covered by the insurance.
- **smoker**: Whether the individual is a smoker (yes/no).
- **region**: Region where the individual resides (southeast, southwest, northeast, northwest).
- **charges**: The insurance charges (target variable).

---

3. Data Preprocessing

3.1 Encoding Categorical Features

Categorical features like **sex**, **smoker**, and **region** were transformed into numeric values using label encoding. The encoding was done manually based on domain knowledge and the values provided in the dataset.

- **sex**: Encoded as 0 for 'male' and 1 for 'female'.
- **smoker**: Encoded as 0 for 'yes' (smoker) and 1 for 'no' (non-smoker).
- **region**: Encoded with numerical values based on the region names:
  - 'southeast' = 0
  - 'southwest' = 1
  - 'northeast' = 2

    - 'northwest' = 3

3.2 Feature Scaling

Feature scaling was applied to the dataset using a **StandardScaler**, which scales the features to have a mean of 0 and a standard deviation of 1. This step is essential for algorithms like Gradient Boosting, which are sensitive to the magnitude of the input features.

---

4. Model Selection

To predict the insurance charges, several regression models were considered. The model chosen for final implementation was **Gradient Boosting** due to its superior performance on this task.

4.1 Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds multiple decision trees sequentially, with each tree trying to correct the errors of the previous one. It's a powerful model for regression tasks and works well with both numerical and categorical features.

Evaluation of Gradient Boosting Model

- **Mean Squared Error (MSE)**: 18,686,549.01
- **R-squared (R2) Score**: 0.88

The **MSE** value indicates that the model's predictions are fairly close to the actual values, as the error is relatively small. The **R2 Score** of 0.88 suggests that 88% of the variance in the insurance charges is explained by the model, which is a strong result.

---

5. Model Performance

The performance of the Gradient Boosting model was evaluated using the following metrics:

- **MSE (Mean Squared Error)**: This metric gives us the average squared difference between the actual and predicted values. A lower MSE indicates better performance.
- **R2 Score**: This is a measure of how well the model explains the variance in the target variable. An R2 score close to 1 indicates a good fit, meaning the model explains most of the variance in the data.

For this model:

- **MSE** = 18,686,549.01
- **R2 Score** = 0.88

These results indicate that the Gradient Boosting model is highly accurate and performs well on the task of predicting insurance charges.

---

6. Approach Applied

Step-by-Step Approach:

1. **Data Collection and Cleaning**:
   - The data was collected and cleaned, ensuring there were no missing or erroneous values.
2. **Feature Engineering**:
   - Categorical variables (sex, smoker, region) were manually encoded using domain-specific mappings.
   - Numerical features (age, bmi, children) were retained without modification.
3. **Feature Scaling**:
   - Feature scaling was applied to standardize the input features, which is necessary for the model to perform optimally.
4. **Model Selection**:
   - Gradient Boosting was selected as the best model after trying several regression techniques. It provided the best balance between performance and complexity.
5. **Model Training and Evaluation**:
   - The Gradient Boosting model was trained on the preprocessed data.
   - It was then evaluated on the test set, producing an MSE of 18,686,549.01 and an R2 score of 0.88.
6. **Deployment**:
   - The final trained model, encoder, and scaler were saved and can be used for making predictions on new data.

---

7. Conclusion

- **Best Model**: The **Gradient Boosting** model performed exceptionally well with an **R2 score of 0.88** and **MSE = 18,686,549.01**, which indicates a strong ability to predict insurance charges.
- **Future Work**: Although Gradient Boosting provided good results, further model optimization, such as hyperparameter tuning or trying other models like XGBoost, LightGBM, or Random Forest, may further improve the performance.
- **Deployment**: The model, scaler, and encoder were deployed using a Flask web application, enabling the user to make predictions via a user-friendly interface.

This model will be valuable for insurance companies, helping them set fair pricing and understand their customers' insurance needs based on their demographic and health-related information.