Predicting Insurance Charges and Classifying High-Cost Customers using Machine Learning

1. Problem Statement

The objective of this project is two-fold:

1. To build a **classification model** that predicts whether an individual is likely to incur high insurance charges based on demographic and health features.
2. To build a **regression model** that predicts the **actual insurance cost** for an individual using the same set of features.

2. Dataset Description

Source: https://www.kaggle.com/datasets/mirichoi0218/insurance

Feature	Description
`age`	Age of the primary beneficiary
`sex`	Gender (male/female)
`bmi`	Body Mass Index
`children`	Number of children covered by health insurance
`smoker`	Smoking status (yes/no)
`region`	Residential area in the US
`charges`	Final insurance cost billed to the customer

Derived Feature:

- `high_cost`: A binary variable created by labeling individuals whose `charges` exceed the median as `1` (high cost), and others as `0` (low cost). Used as the target for classification.

3. Exploratory Data Analysis (EDA)

- No missing values were found in the dataset.
- Categorical features were inspected using count plots.
- Numerical features such as `age`, `bmi`, and `charges` were visualized using histograms.
- Class balance in the new `high_cost` variable was checked and found to be balanced (669 high-cost, 669 low-cost).

4. Data Preprocessing

- Categorical variables (`sex`, `smoker`, `region`) were encoded using **Label Encoding**.
- Numerical features (`age`, `bmi`, `children`) were **standardized** using `StandardScaler`.
- Two separate train/test splits were prepared:
  - For classification: `y = high_cost`
  - For regression: `y = charges`

5. Model Building

A. Classification Model

- **Target Variable**: `high_cost`
- **Model Used**: Logistic Regression
- **Metrics Used**: Accuracy, Precision, Recall, F1-score, Confusion Matrix

Results:

Metric	**Score**
Accuracy	91%
Precision	90%
Recall	92%
F1-Score	91%

Confusion Matrix:

```
[ [120  14]
  [ 11 123] ]
```

The model performs well with a balanced precision and recall, and minimal misclassifications.

B. Regression Model

- **Target Variable**: `charges`
- **Model Used**: Linear Regression
- **Metrics Used**: MAE, MSE, RMSE, R^2 Score

Results:

Metric	**Value**
MAE	4186.51
MSE	33,635,210
RMSE	5799.59
R^2 Score	0.78

The regression model explains approximately 78% of the variance in insurance charges. Errors are reasonable given the scale of the target variable.

6. Conclusion

This project successfully demonstrates the application of machine learning techniques on real-world health insurance data for both classification and regression tasks. The models built were:

- A **Logistic Regression classifier** to identify high-cost customers with strong performance (F1-Score: 91%)
- A **Linear Regression model** to predict actual charges with good predictive power (R^2 Score: 0.78)

Further improvement can be made by exploring:

- Advanced models like Random Forest or Gradient Boosting
- Hyperparameter tuning with GridSearchCV
- Feature engineering or polynomial features