Machine Learning Final Project

---

1. Dataset Selection

The dataset used for this project was the Medical Insurance Dataset, obtained from Kaggle.

Each row in the dataset represents an individual, and each column represents a factor that may influence medical insurance costs. The dataset includes both numerical and categorical variables, making it suitable for regression analysis and data preprocessing techniques.

Given Features

- `age` – Age of the individual
- `sex` – Gender
- `bmi` – Body Mass Index
- `children` – Number of dependents
- `smoker` – Smoking status
- `region` – Geographic region
- `charges` – Medical insurance cost (target variable)

---

2. Problem Statement and Model Choice

Objective :

The goal of this project was to predict medical insurance charges based on individual health, demographic, and lifestyle factors.

Why Linear Regression?

Linear regression was chosen because:

- The target variable (`charges`) is continuous
- Multiple independent variables influence the outcome
- The model provides interpretability, allowing analysis of how each factor affects insurance costs
- It is appropriate for real-world numerical prediction problems

---

3. Data Preprocessing and Feature Engineering

Significant preprocessing was required before training the model.

Data Cleaning

- Missing values were removed using **`dropna()`**
- All categorical values were standardized by converting them to lowercase

Categorical Encoding

Different encoding methods were used based on the nature of each variable:

- **Ordinal Encoding**
  - Applied to binary categorical variables:
    i. `sex`
    ii. `smoker`

- **One-Hot Encoding**
  - Applied to **region**, since it has multiple categories with no inherent order

Feature Engineering

To improve model performance, interaction features were created:

- `bmi_smoker` = BMI × smoker status
- `bmi_age` = BMI × age
- `age_smoker` = age × smoker status

---

4. Model Training and Evaluation

The dataset was split into 80% training data and 20% testing data.

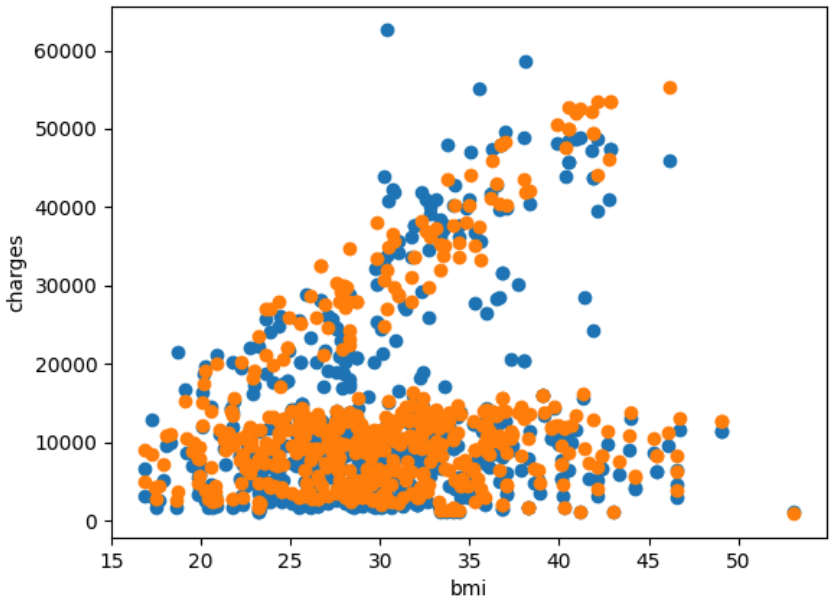
Model Performance

- **Training R^2 =** 0.8420830158817929
- **Testing R^2 =** 0.8348919181822081

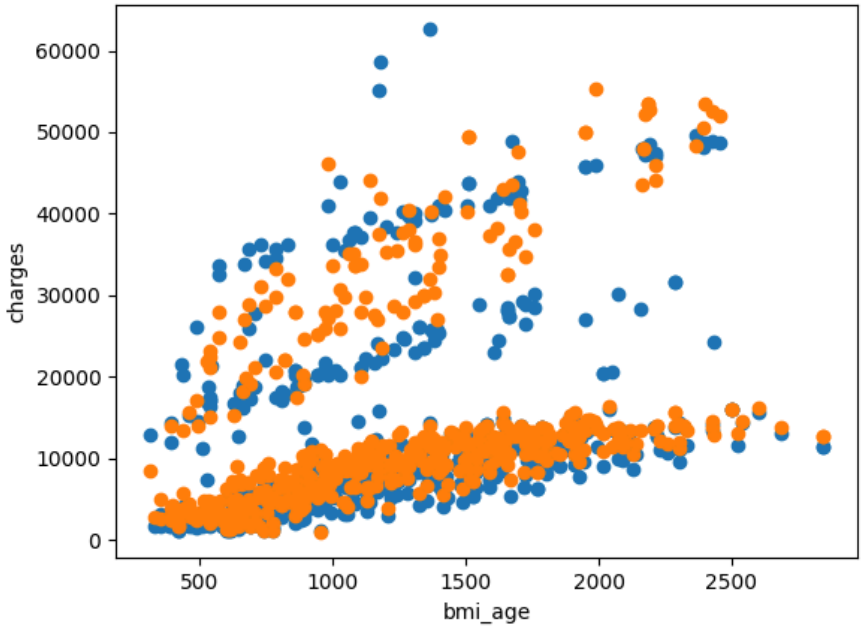
---

5. Data Visualization

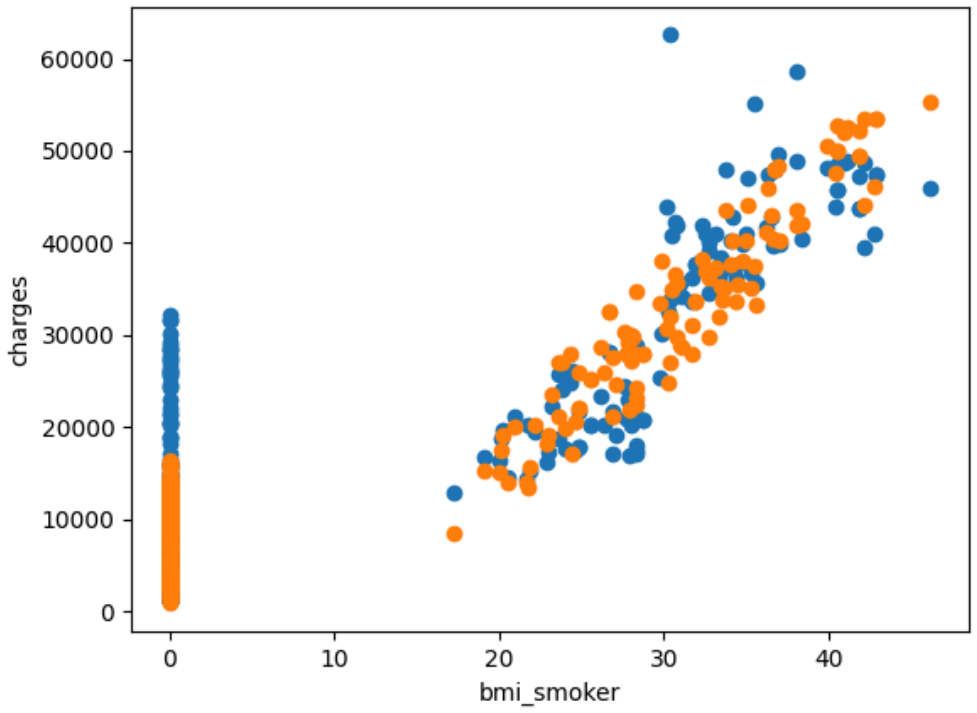
The **BMI vs charges** graph shows that insurance charges generally increase as BMI increases, but the data is widely scattered. This indicates that BMI alone does not fully explain insurance costs and that other factors influence the outcome.



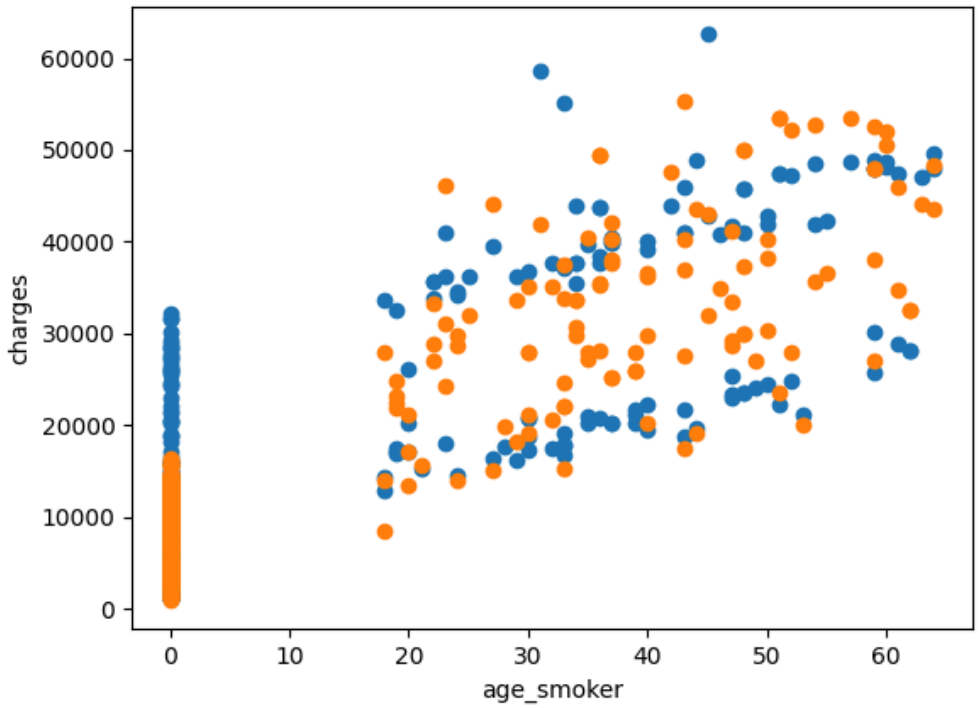
The **BMI × Age** graph shows a clearer upward trend, suggesting that the effect of BMI becomes stronger as age increases. This supports the idea that older individuals with higher BMI tend to have significantly higher insurance charges.



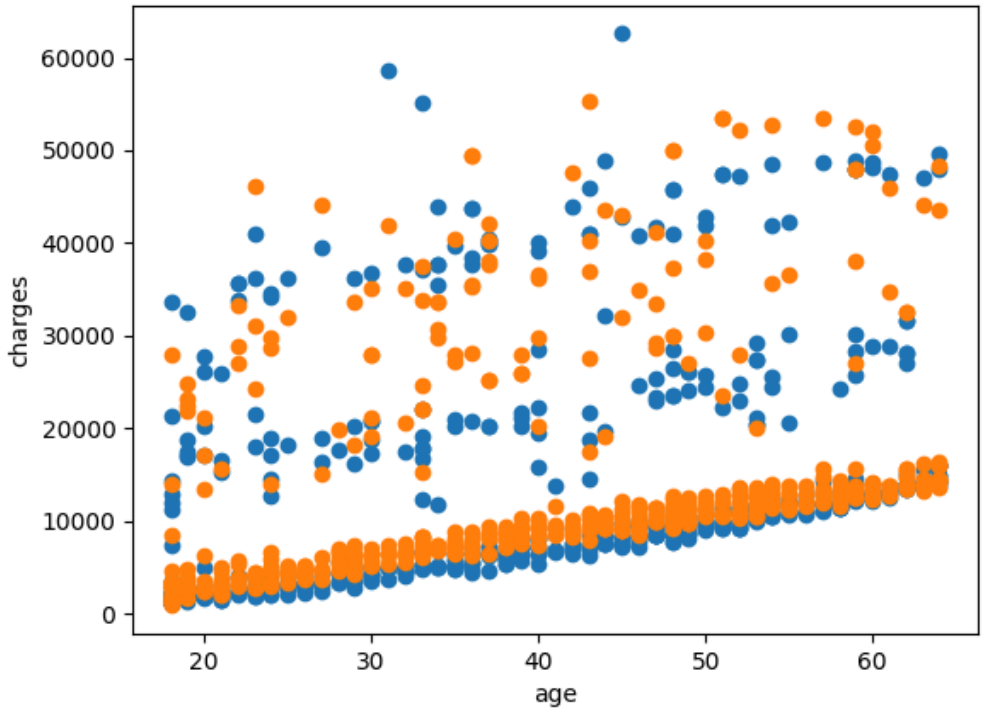
The **BMI × Smoker** graph shows a strong separation between smokers and non-smokers. Non-smokers cluster near zero, while smokers show a clear increase in charges as BMI increases. This demonstrates that smoking greatly amplifies the effect of BMI on insurance costs.



Similarly, the **Age × Smoker** graph shows that insurance charges rise much more rapidly with age for smokers than for non-smokers. This confirms that smoking intensifies the impact of age on medical expenses.



Finally, the **Age vs charges** graph shows a general upward trend but with significant variation. When compared to the interaction graphs, it becomes clear that age predicts charges more accurately when combined with smoking status or BMI.



Based on the graphs, smoking status has the strongest effect on insurance charges, as both the BMI × smoker and age × smoker plots show much higher and more rapidly increasing costs for smokers compared to non-smokers. Age is the next most important factor, since charges generally increase as age increases, especially for smokers. BMI also affects insurance charges, but its impact is moderate on its own and becomes much more significant when combined with age or smoking. Overall, the interaction between smoking, age, and BMI explains insurance costs far better than any single variable alone.