README Project Medical Cost

🩺 Medical Cost Prediction

This project aims to predict medical insurance costs based on personal factors. It uses regression models to identify the factors that most influence an individual's insurance costs. The models used are Linear Regression, Random Forest, and K-Neighbors Regressor.

📁Dataset

Source: Kaggle - Medical Cost Personal Dataset
This dataset contains 1,338 individual data points with seven input features, including age, sex, BMI, children, smoker, region, and charges.

⚙ Tools

- Python (Pandas, NumPy, scikit-learn, Matplotlib, Seaborn)
- Jupyter Notebook/VS Code

🎯 Purpose of Analysis

- Predicting an individual's medical insurance costs based on lifestyle and demographic data.
- Identifying the features that most influence insurance costs.
- Comparing the performance of different regression models.

🔎 Steps

1. Data Loading & Exploratory Data Analysis (EDA)

- Reading datasets using Pandas
- Checking missing values and descriptive statistics
- Analysing relationships between variables with heatmaps and scatter plots

2. Feature Engineering

- Create additional features, such as:
  - `BMI_Category` → categorises BMI values into Underweight, Normal, Overweight, Obese
  - `Age_Group` → grouping ages into the categories Young, Adult, Senior
  - `Smoker_BMI_Interaction` → combined effect of smoking and high BMI on insurance costs

- One-hot encoding for categorical features

3. Data Splitting & Scaling

- Split the data into train (80%) and test (20%)
- Normalise numerical features using StandardScaler

4. Modeling

Model used:

- Linear Regression
- Random Forest Regressor
- K-Neighbors Regressor

5. Evaluation

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R² Score

📊 Visualization

- Linear Regression showed the best performance with the highest R² value and the smallest error.
- The most influential features were smoker, age, and BMI.
- The scatter plot showed that smokers with high BMI had significantly higher insurance costs.

🧠Insight

1. Smokers incur much higher insurance costs than non-smokers.
2. Older age → higher insurance costs.
3. High BMI → higher insurance costs.
4. Region & number of children → less impact.
5. Combined effects (BMI_smoker & age_smoker) → amplify costs.
6. Among all the models tested, Random Forest performed best in predicting insurance charges, demonstrating its ability to identify non-linear relationships in the data. Linear regression showed the highest R² score (0.8869), indicating better overall suitability and consistency in the data. Therefore, linear regression was selected as the final model for this project due to its strong balance between readability and prediction accuracy.

🚀 How to Run

- Ensure all required packages are installed:

```
pip install pandas numpy scikit-learn seaborn matplotlib
```

- Run the notebook in Jupyter or VS Code:

  ```
  jupyter notebook Project_MedicalCost.ipynb
  ```