**Linear Regression Project Report**

**1. Project Title:**

Predicting Insurance Premium Based on Age using Linear Regression

**2. Introduction**

This project applies **Linear Regression**, a supervised learning algorithm, to predict insurance premiums based on a person's age. By analyzing a given dataset, the model learns the relationship between age and premium and predicts premium amounts for new ages.

**3. Dataset Description**

The dataset (custom\_data.csv) consists of two columns:

* **Age (Independent Variable - X):** Represents the age of the insured person.
* **Premium (Dependent Variable - Y):** Represents the insurance premium to be paid.

**4. Methodology**

**Step 1: Data Loading and Exploration**

* Used pandas to read and display the dataset.
* Verified the dataset structure and checked for missing values.

**Step 2: Data Visualization**

* Used seaborn.lmplot() to visualize the relationship between age and premium with a regression line.

**Step 3: Model Training**

* Used sklearn.linear\_model.LinearRegression() to create a linear regression model.
* Trained the model using the fit() method with age as the input feature and premium as the target variable.

**Step 4: Prediction**

* Used reg.predict([[21]]) to predict the insurance premium for a person aged **21**.

**5. Results & Interpretation**

* The trained model successfully learned the pattern between age and premium.
* The predicted premium for age 21 gives an estimated value based on the regression model.
* The visualization confirms a **linear relationship**, validating the use of linear regression.

**6. Conclusion**

* **Linear Regression** is effective for predicting continuous numerical values.
* This model can be expanded with more features (e.g., income, health status) for better accuracy.
* Further improvements can include **feature scaling, outlier detection, and hyperparameter tuning**.

**7. Future Enhancements**

* Testing the model on a larger dataset.
* Using **Polynomial Regression** if the relationship is non-linear.
* Evaluating model performance with **Mean Absolute Error (MAE) and R-squared score**.

**Project Completed by:** M.Iqbal  
**Tools Used:** Python, Pandas, Seaborn, Scikit-Learn