Simple Linear Regression For Sales dataset

[Machine Learning mini project]

Problem Statement:

Organizations and HR departments often aim to understand how years of experience influence employee salaries in order to make informed hiring decisions and offer fair compensation. This project focuses on building a predictive model using Simple Linear Regression to estimate an individual's salary based on their years of experience.

The objectives of this project are to:

- Explore the linear relationship between experience and salary
- Develop a regression model that can accurately predict salaries
- Support data-driven decision-making in HR and workforce planning

Data Exploration and Preprocessing:

Before building the regression model, a thorough exploration and preprocessing of the dataset was carried out to ensure data quality and extract meaningful insights.

1. Exploratory Data Analysis (EDA):
    - The dataset contains two primary features:
        - YearsExperience (independent variable)
        - Salary (dependent variable)

- A scatter plot was generated to visualize the relationship between experience and salary. The plot indicated a strong positive linear correlation, suggesting that linear regression would be a suitable modeling technique.
- Basic statistical summaries (mean, min, max, standard deviation) were calculated to understand the scale and distribution of the data.
- The correlation coefficient between years of experience and salary was computed and found to be close to 1.0, further confirming a strong linear relationship.

2. Data Preprocessing:

- The dataset was checked for missing values or outliers. No missing data was found.
- Data types were verified to ensure numeric types for modeling.
- The data was split into:
    - Training set (typically 80%)
    - Testing set (typically 20%) using train_test_split from Scikit-learn to evaluate model performance objectively.

This clean and well-structured dataset required minimal preprocessing, making it ideal for demonstrating the fundamentals of linear regression.

Model Building and Evaluation:

After exploring and preparing the data, the next step was to build a predictive model using Simple Linear Regression. This section outlines the process used to train and assess the performance of the model.

- Model Building:
    - Data Split:The dataset was divided into training and testing sets (typically 80% for training and 20% for testing) using Scikit-learn's train_test_split function. This separation ensures that the model can be evaluated on unseen data.

- Model Training: Using the training data, a Simple Linear Regression model was constructed. The model learns the relationship between YearsExperience (independent variable) and Salary (dependent variable) by estimating the best-fitting line through the data.
- Implementation: The model was implemented using Scikit-learn's LinearRegression class in Python. After fitting the model on the training data, the learned parameters (slope and intercept) define the linear relationship that can be used for prediction.

- Model Evaluation:
    - Predictions: The trained model was used to predict salaries on the test set, allowing for the assessment of its performance on new data.
    - Performance Metrics:
        - Mean Squared Error (MSE): This metric measures the average squared difference between the predicted and actual salary values. A lower MSE indicates a model with better predictive accuracy.

        - R^2 Score (Coefficient of Determination): The R^2 value explains the proportion of the variance in the dependent variable that is predictable from the independent variable. An R^2 close to 1.0 indicates that the model explains most of the variability in the salary data.

- Visualization:
    - A scatter plot of the test data points was created and overlaid with the predicted regression line.

- This visual representation helps in confirming the adequacy of the linear model and in identifying any potential misfits.

Overall, the model was able to capture a strong linear relationship between years of experience and salary, and evaluation metrics demonstrated that the model achieves high accuracy in predicting salary values based on experience.

Model Performance:

The linear regression model showed a strong performance with high R^2 and low MSE, making it a reliable model for predicting salaries based on years of experience.