Polynomial Linear Regression Project

Introduction

This project explores Polynomial Linear Regression using a dataset from Kaggle that includes information on various positions, levels, and salaries. The primary objective is to model the relationship between experience levels and salaries and to predict salaries based on different levels of experience using polynomial regression techniques.

Dataset

- **Source**: Polynomial Linear Regression Dataset on Kaggle
- **Description**: The dataset contains three columns:
  - `Position`: Categorical data representing job titles.
  - `Level`: Integer data indicating the level of experience.
  - `Salary`: Numeric data representing the salary.

Steps and Methodology

1. **Data Exploration and Cleaning**
   - Loaded the dataset and examined its structure.
   - Handled missing values and categorical data using Label Encoding to convert job titles into numeric values.
2. **Exploratory Data Analysis (EDA)**
   - Conducted visual analysis including histograms, scatter plots, and correlation matrices.
   - Analyzed the distribution of experience levels and salaries.
   - Detected and addressed any outliers that could impact the model's performance.
3. **Data Preprocessing**
   - Encoded categorical variables and split the dataset into training and testing sets.
   - Normalized data as necessary to ensure consistency in model training.
4. **Model Selection and Training**
   - Applied Polynomial Linear Regression to model the relationship between experience levels and salaries.
   - Trained the model using the training dataset and evaluated its performance using metrics such as Mean Squared Error (MSE) and R-squared (R^2).
5. **Model Evaluation**
   - Evaluated model performance with R-squared and MSE metrics.

- Achieved a high R-squared value of 0.97, indicating that the model explains 97% of the variance in salary predictions.
- Observed a relatively high MSE, suggesting room for further model improvement.

Conclusion

The Polynomial Linear Regression model successfully captured the nonlinear relationship between experience levels and salaries, demonstrating strong predictive capabilities with an R-squared value of 0.97. However, the high Mean Squared Error highlights the need for potential model refinement to enhance prediction accuracy.