Midterm Project Report : Salary Prediction using Multiple Linear Regression

1. Problem Description

The goal of this project is to develop a machine learning model that predicts employee salaries based on relevant experience and certification data. Salary prediction is a valuable task in HR analytics, helping organizations set fair compensation and helping job seekers understand expected salary ranges. I chose this problem because salary is a continuous variable, and it fits well with regression modeling, which is the focus of this midterm.

2. Dataset

The dataset was downloaded from Kaggle and contains 1,000 rows and 5 columns:

- Total Experience
- Team Lead Experience
- Project Manager Experience
- Certifications
- Salary (Target)

All features are numeric and directly related to work experience or qualifications. I used all four predictor variables to estimate salary.

Before modeling, I checked for missing values and duplicates, but none were found. Since the features have different scales (e.g., years vs. count of certifications), I applied StandardScaler to normalize the data.

3. Modeling Approach

I started with Multiple Linear Regression using scikit-learn. After splitting the data into training and test sets (80/20), I scaled the features using StandardScaler fitted only on the training set to avoid data leakage.

I also explored Polynomial Regression (degree=2) to capture potential nonlinear relationships, but the increase in performance was minimal and not worth the added complexity for this dataset.

4. Evaluation

I used the following metrics to evaluate the model:

- R^2 Score: 0.94
- Mean Squared Error (MSE): 20819427
- Mean Absolute Error (MAE): 3733

The R^2 score indicates that about 94% of the variation in salary is explained by the model. The MSE and MAE show that on average, the model's salary prediction is off by around 3733 units.

5. Reflection

One challenge was interpreting the effect of multicollinearity—some experience columns may be correlated (e.g., Team Lead Experience and Total Experience). A correlation heatmap helped reveal these relationships(0.65).

I learned how to preprocess real-world numeric data, apply feature scaling, and evaluate regression models using multiple metrics. If I were to improve this project, I would:

- Add categorical features like education level, location, or industry
- Try regularization (Ridge or Lasso) to reduce potential overfitting
- Use cross-validation for a more robust evaluation

This project helped reinforce core concepts in regression modeling and data handling.