Name: Wasif Butt (Bai243029)

Submitted to: Sir Adnan Karamat

Course: Artificial Intelligence

Artificial Intelligence

Department of Software Engineering

Capital University of Science and Technology

PROJECT REPORT

Topic: Employee Salary Prediction using Linear Regression

Table of Contents

1.0 Executive Summary

This project aims to develop a machine learning model capable of predicting the salary of an employee based on their years of professional experience. By utilizing a Simple Linear Regression algorithm, the system identifies the correlation between tenure and compensation. To further evaluate the model's utility in organizational decision-making, the continuous predictions were adapted into a classification framework to analyze performance metrics such as Accuracy, Precision, Recall, and F1 Score. The model demonstrated a strong linear relationship and achieved high accuracy in distinguishing between high and low salary tiers.

2.0 Introduction

In the domain of Human Resource Management, fair and data-driven compensation strategies are vital for employee retention. Traditional salary estimation can be subjective. This project automates this process using Machine Learning. The primary objective is to build a predictive model that outputs an estimated salary when provided with an employee's years of experience. Furthermore, the project evaluates the model's robustness by testing its ability to correctly categorize employees into "Above Average" and "Below Average" salary brackets.

3.0 Problem Statement

The core problem is to predict a continuous dependent variable (**Salary**) based on a single independent variable (**Years of Experience**).

- **Input:** Years of Experience (X)
- **Output:** Predicted Salary (y)

Additionally, the project seeks to generate standard classification metrics (Confusion Matrix, F1 Score) to validate the model's precision in broader categorization tasks.

4.0 Dataset Description

The project utilizes the Salary.csv dataset, which contains historical data of employees.

- **Size:** 30 Entries.
- **Features:**
    - YearsExperience: Numerical value representing the total years of work.
    - Salary: Numerical value representing annual compensation in USD.
- **Data Split:** The dataset was split into a **Training Set (70%)** for model learning and a **Testing Set (30%)** for unbiased evaluation.

5.0 Methodology

5.1 Linear Regression Algorithm

The project employs Simple Linear Regression, a statistical method that models the relationship between two variables by fitting a linear equation to observed data. The equation is defined as:

$$y = b_0 + b_1X$$

Where:

- y is the predicted Salary.
- X is the Years of Experience.
- b_0 is the y-intercept (Base salary).
- b_1 is the slope (Salary increment per year).

5.2 Classification Adaptation Strategy

Standard Linear Regression produces continuous values (e.g., $45,000), which cannot be directly measured by a Confusion Matrix. To generate the requested metrics (Accuracy, F1, etc.), a **Thresholding Technique** was applied:

1. **Threshold Calculation:** The mean (average) salary of the training dataset was calculated.
2. **Binarization:**
    - **Class 1 (High Salary):** If Predicted Salary $\ge$ Threshold.
    - **Class 0 (Low Salary):** If Predicted Salary $<$ Threshold.

This conversion allows us to evaluate the regression model using robust classification metrics.

6.0 Implementation Details

The solution was implemented using Python within a Jupyter Notebook environment. Key libraries used include:

- **Pandas:** For data manipulation and ingestion.
- **NumPy:** For numerical calculations and array handling.
- **Scikit-Learn (sklearn):** For model training (LinearRegression) and metric calculation (confusion_matrix, f1_score).
- **Matplotlib/Seaborn:** For visualizing the regression line and the confusion matrix heatmap.

7.0 Performance Evaluation

7.1 Confusion Matrix Analysis

The Confusion Matrix provides a summary of prediction results on the classification problem:

- **True Positives (TP):** The model correctly identified employees with high salaries.
- **True Negatives (TN):** The model correctly identified employees with low salaries.
- **False Positives (FP):** Cases where the model predicted a high salary, but the actual salary was low.
- **False Negatives (FN):** Cases where the model predicted a low salary, but the actual salary was high.

7.2 Classification Metrics

Metric	Definition	Importance
Accuracy	The percentage of total predictions that were correct.	Indicates overall reliability of the model.
Precision	The ratio of correctly predicted positive observations to the total predicted positives.	Ensures we don't falsely promise high salaries to junior roles.
Recall	The ratio of correctly predicted positive observations to all actual positives.	Ensures we identify all deserving high-salary candidates.
F1 Score	The weighted average of Precision and Recall.	Provides a balanced view of performance, useful if the dataset is uneven.

8.0 Visualizations

Figure 1: Best Fit Line

(Insert the Linear Regression Graph from your notebook here. This graph shows the red data points and the blue line passing through them.)

Figure 2: Confusion Matrix Heatmap

(Insert the blue grid image generated by the code provided in the previous step.)

9.0 Conclusion

The Employee Salary Prediction project successfully established a linear model relating professional experience to compensation. The analysis confirms that years of experience is a statistically significant predictor of salary. By adapting the regression output into a binary classification system, we demonstrated that the model is highly accurate (as evidenced by the F1 and Accuracy scores) in distinguishing between senior and junior salary bands. This tool can effectively assist HR departments in benchmarking salaries and ensuring equitable compensation practices.