Title – Salary Prediction

1. Introduction

1.1 Project Overview

This project aims to predict the Salary using the Work Experience. The objective is to help in assigning the salary to new employees based on their work experience and helps in defining the salary as the job experience increases.

1.2 Dataset Description

The dataset used includes Delivery time and sorting time of first 29 records. Key attributes include ‘Years Exp’ and ‘Salary’. Initial pre-processing steps involved data cleaning, pattern analysing, studying insights and finding the right model.

2. Solution Architecture

2.1 Data Ingestion

Data was loaded using pandas. The ingestion process included handling large datasets, connecting to the database.

2.2 Data Processing

The data processing pipeline included:

* Cleaning: Addressing missing values by fillna().
* Transformation: Feature scaling and encoding were performed using Z-score normalization.

2.3 Model selection

The models considered included Linear Regression. The final model was selected based on patterns that were made by analysing the dataset.

2.4 Evaluation Metrics

Model performance was evaluated using mean square error, r2 score, train test split. These metrics were used to improve the strength or make the predicted value more accurate towards the original values.

3. Methodology

3.1 Exploratory Data Analysis (EDA)

EDA revealed that when there is increase in time of sorting the delivery time also gets increased. Visualisations such as regression plot was used to understand the data distribution.\

3.2 Modeling Process

The model was built was using Linear regression. Training involved splitting the data by the ratio of 1:4, then 4th ratio was trained using Linear Regression and then the model was tested on 1st ratio.

4. Implementation

4.1 Code Breakdown

Key sections of the code include:

* Data loading
* Feature engineering
* Model training

4.2 Libraries and Tools

The project was implemented by using pandas, matplot, seaborn, math, numpy.

4.3 Challenges and Solution

Challenges faced included Data sorting, finding the necessary patterns, selecting the right model and improve its accuracy score. These were addressed by using pandas for data sorting, seaborn and matplot for finding the right patterns, linear regression for selecting right model and r2 score for checking its accuracy.

5. Result and Conclusion

5.1 Final Model Performance

The final model achieved 95.7% of accuracy with error of 5592.044 currency units.

5.2 Business Impact

The solution provides help in assigning the salary to new employees based on their work experience and helps in defining the salary as the job experience increases.

5.3 Future Work

Future improvements could include adding new features, testing more models based on future dataset values.

6. References

Kaggle dataset, ChatGPT, stackoverflow, google search, regression documentation.