Salary Prediction Project Report

1. Introduction

This project aims to build a machine learning model to predict salaries of employees based on various features such as experience, education level, job role, location, and industry. The dataset used here is synthetic but designed to mimic real-world trends.

2. Dataset Description

The dataset (salaries.csv) consists of 500 records with the following columns: - Experience (Years) - Education (High School, Bachelors, Masters, PhD) - Job Role (Data Scientist, Software Engineer, Manager, Analyst) - Location (New York, San Francisco, Austin, Remote) - Industry (Tech, Finance, Healthcare, Education) - Salary (Target variable, in USD)

3. Methodology

The project follows a typical data science workflow: 1. Data Loading & Exploration (EDA) 2. Data Preprocessing (handling categorical data, scaling) 3. Model Training (Linear Regression, Random Forest) 4. Model Evaluation (R^2 Score, MAE, RMSE) 5. Visualization of predictions

4. Exploratory Data Analysis (EDA)

- Salary distribution is approximately normal with some variance due to added noise. - Higher education levels generally correspond to higher salaries. - Job roles like Managers and Data Scientists have higher average salaries compared to Analysts. - Locations such as San Francisco and New York have higher salaries compared to Remote jobs.

5. Models Used

Two models were trained and evaluated: - Linear Regression: A simple baseline model for salary prediction. - Random Forest Regressor: An ensemble model that performed better due to handling non-linear relationships.

6. Evaluation Metrics

The models were evaluated using: - R^2 Score (explains variance in data) - Mean Absolute Error (MAE) - Root Mean Squared Error (RMSE) Random Forest performed better compared to Linear Regression, achieving higher R^2 and lower error values.

7. Conclusion

This project demonstrates how data science techniques can be applied to predict salaries. Although the dataset is synthetic, the methodology can be extended to real-world datasets for HR analytics, compensation planning, and career guidance applications. Future improvements: - Use advanced models like XGBoost or Neural Networks. - Include additional features such as company size, certifications, or skills. - Deploy the model as a web application (Streamlit/Flask) for interactive use.