Project Report: Employee Salary Prediction

Submitted to:

Edunet Foundation
IBM SkillsBuild Program

Submitted by:

Name : [Richa Sonkar]
University Name : [Indira Gandhi Delhi Technical University For Women]
Course Name: [B.Tech in Computer Science And Engineering With Specialisation in Artificial Intelligence]
Submission Date: [30th July 2025]

1. Introduction

Salary prediction is an essential task in human resource management. Inaccurate or biased compensation decisions can lead to employee dissatisfaction, attrition, and legal complications. Using data science techniques, particularly machine learning (ML), organizations can forecast employee salaries more objectively and fairly.

This project focuses on building a machine learning model that predicts employee salary based on several features such as experience, education level, job title, and location. This kind of system can support HR departments in:

- Benchmarking salaries,
- Identifying outliers,
- Promoting transparency and fairness in compensation.

2. Objectives

The main objectives of the Employee Salary Prediction project are:

- To apply machine learning algorithms to estimate the salary of an employee.
- To understand how features like experience, education, job title, and location influence salary.
- To evaluate the performance of multiple regression models and choose the best-performing one.
- To practice real-world implementation of data science concepts such as preprocessing, model training, evaluation, and deployment.

3. Tools and Technologies

The following tools, libraries, and platforms were used throughout the project:

- Programming Language: Python
- IDE / Platform: Google Colab / Jupyter Notebook
- Libraries:
    - pandas – for data manipulation
    - numpy – for numerical operations
    - matplotlib, seaborn – for data visualization
    - scikit-learn – for machine learning models
    - xgboost – for gradient boosting
- Visualization: Charts and graphs created using matplotlib and seaborn
- Documentation: MS Word / Google Docs for report writing

4. Dataset Description

As no real-world dataset was provided, a synthetic dataset was created to simulate a real HR salary dataset. The dataset contains the following features:

Feature Name	Description
Experience	Number of years of experience (numeric)
Education	Categorical: Bachelor's, Master's, PhD
Job Title	Categorical: Data Scientist, Developer, Analyst, etc.
Location	Categorical: Delhi, Mumbai, Bangalore, etc.
Salary	Target variable (numeric, in INR)

The dataset contains 1000+ entries, ensuring diversity across roles and locations.

5. Methodology

5.1 Data Preprocessing

- Handling Categorical Variables: Used Label Encoding and One-Hot Encoding to convert textual data (like education level and job title) into numerical format.
- Feature Scaling: Applied StandardScaler for normalizing features like experience.

- Splitting Dataset: Divided data into training and testing sets (typically 80:20).

5.2 Model Selection & Training

Three regression models were trained and evaluated:

1. Linear Regression – Baseline model for comparison
2. Random Forest Regressor – An ensemble model using decision trees
3. XGBoost Regressor – Gradient boosting technique for optimized predictions

5.3 Model Evaluation

The models were evaluated using the following metrics:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R² Score (Coefficient of Determination)

6. Results

After training and testing, the following results were obtained:

Model	MAE (Rs)	RMSE (Rs)	R² Score
Linear Regression	4223.55	5081.84	0.89
Random Forest	4644.34	5794.22	0.86
XGBoost	5115.35	6275.09	0.84

- Linear Regression performed the best with an R² Score of 0.89, meaning it explained 89% of the variance in salary prediction.

7. Key Learnings

- Gained hands-on experience in real-world regression tasks.
- Understood the role of data preprocessing in model accuracy.
- Learned to compare model performance using standard evaluation metrics.
- Discovered that even simple models (like linear regression) can perform remarkably well on structured datasets.

8. Conclusion

This project successfully demonstrated how machine learning can be leveraged to predict employee salaries. It highlights the importance of clean data, careful model selection, and evaluation in building effective predictive systems. HR departments can benefit from such tools for data-driven compensation management, ultimately leading to better talent retention and satisfaction.

9. Attachments

- employee_salary_prediction.py– Code & output
- Employee Salary Prediction .1 – Presentation
- Employee_Salary_Prediction_Report.pdf – This report (PDF)
- README.md – Project description for GitHub

10. Acknowledgements

I would like to thank:

- Edunet Foundation for organising this opportunity
- IBM SkillsBuild for providing learning resources and certification
- Mentors and coordinators for their continuous support and guidance