PROJECT REPORT

Name: Shikhar Srivastava
Roll Number: 2300100100368
Class & Division: 3rd Year
Project Guide: R Dev Nath
Project Title: Salary Prediction using Ensemble
Learning

---

1. Introduction

In today's competitive job market, predicting salary
based on various factors like skills, experience,
education, and job role is a significant challenge for both
job seekers and recruiters. Accurate salary prediction
helps candidates negotiate better pay and assists
companies in offering competitive packages. This
project aims to develop a machine learning model to
predict employee salary using ensemble learning
techniques, which combine multiple learning algorithms
to improve predictive performance.

---

2. Project Overview

The project focuses on building an effective salary
prediction system leveraging ensemble learning

methods such as Random Forest, Gradient Boosting, and AdaBoost. The system takes multiple input features including educational qualification, years of experience, job location, skill sets, and other relevant parameters to predict the expected salary. By comparing different ensemble models, the project aims to select the best-performing algorithm that provides higher accuracy and reliability.

---

3. Objectives

- To analyze factors influencing salary in various job sectors.

- To apply ensemble learning techniques for salary prediction.

- To evaluate and compare model performances using metrics such as RMSE, MAE, and R^2 score.

- To develop a user-friendly interface for inputting data and displaying predicted salaries.

---

4. Modules

The project is divided into the following modules:

1. **Data Collection and Preprocessing:**
   Collection of datasets from reliable sources and cleaning to handle missing values, categorical encoding, and normalization.

2. **Exploratory Data Analysis (EDA):**
   Visualization and statistical analysis to identify trends, correlations, and outliers in the dataset.

3. **Feature Engineering:**
   Creation and selection of relevant features to improve model accuracy, including encoding categorical variables and scaling.

4. **Model Building:**
   Implementation of ensemble algorithms such as Random Forest, Gradient Boosting, AdaBoost, and possibly stacking methods.

5. **Model Training and Validation:**
   Splitting data into training and testing sets, hyperparameter tuning, and cross-validation.

6. **Performance Evaluation:**
   Comparing models based on evaluation metrics like RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R^2 (coefficient of determination).

7. **User Interface Development:**
   Designing a simple GUI or web-based interface for users to input parameters and get predicted salaries.

8. **Deployment and Testing:**
   Final testing of the system on new data and deployment considerations.

---

5. Technologies Used

- **Programming Language:** Python

- **Libraries:**

  - Scikit-learn (for machine learning algorithms)

  - Pandas and NumPy (for data manipulation)

  - Matplotlib and Seaborn (for data visualization)

- **Development Environment:** Jupyter Notebook or any Python IDE

- **User Interface:** Tkinter (for GUI) or Flask/Django (for web-based interface)

- **Others:** Git (for version control), Excel/CSV (data storage)

---

6. Expected Outcome

The project is expected to deliver a robust salary prediction model with higher accuracy using ensemble learning methods. The user interface will allow users to input various parameters and get an instant salary estimate. This tool can be extended and customized for different industries or job profiles.

---

7. Conclusion

This project combines data science and machine learning to address a real-world problem of salary prediction. By utilizing ensemble learning techniques, it aims to provide a more accurate and reliable model compared to traditional single-algorithm methods. The practical implementation will demonstrate how machine learning can be leveraged in human resource and recruitment processes.