RAJALAKSHMI
ENGINEERING COLLEGE
An AUTONOMOUS Institution
Affiliated to ANNA UNIVERSITY, Chennai

25
years
OF ACADEMIC
EXCELLENCE

Department of Computer Science and Engineering

JOB SALARY PREDICTION

Mrs. M. Divya M.E.
Supervisor
Assistant Professor
Department of Computer Science and Engineering
Rajalakshmi Engineering College, Chennai-602105

Student ROLL NO:220701206
Name: Priyadarshini M

Problem Statement

☐ In today's fast-paced and evolving job market, determining an accurate and fair salary for various job roles is a complex task influenced by numerous dynamic factors such as job title, required experience, education level, industry, company size, and geographic location**. Traditional methods of salary estimation often rely on manual benchmarking or static surveys**, which fail to adapt to market changes in real-time and can result in biased or outdated assessments.This project addresses the problem of salary prediction by developing a machine learning-based system that utilizes **Random Forest tree**. The goal is to build robust predictive models capable of analyzing job-related features to estimate expected salaries with **high accuracy**. These models must efficiently handle **non-linear relationships**, interactions between features, and outliers in the data, all while maintaining interpretability and performance.

Existing System

☐ Current systems for job salary prediction often depend on **basic statistical techniques or traditional regression models**, which lack the capability to accurately capture the complex, nonlinear relationships between various influencing factors such as experience, education level, job domain, location, and skill score. These systems typically use limited datasets and fixed salary benchmarks, making them less adaptable to the dynamic nature of the modern job market.

☐ The absence of advanced feature handling and interaction analysis further reduces the accuracy of these models. Moreover, these systems often **lack transparency and interpretability**, giving users little insight into how predictions are made, which can reduce trust and limit practical usefulness in real-world applications. The limited use of data visualization and model explanation tools in existing systems also restricts their effectiveness for decision-making in both recruitment and career planning.

Objectives

☐ The primary objective of the proposed project, "Job Salary Prediction System Using Random Forest," is to develop an intelligent and accurate salary prediction model utilizing the Random Forest regression algorithm. The system aims to estimate the expected salary for various job profiles based on key features such as job title, industry, education level, skills, work experience, geographic location, and company size. Random Forest is selected due to its ability to handle non-linear relationships, **reduce overfitting through ensemble learning**, and provide **high accuracy** by averaging the predictions of multiple decision trees.

☐ This project is intended to support job seekers, employers, and HR professionals by offering data-driven insights into salary expectations, helping ensure fair and competitive compensation. The model's performance will be evaluated using standard regression metrics like **Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² score.**

Abstract

☐ Accurate job salary prediction is a vital component of modern employment analytics, assisting organizations in competitive compensation planning and helping individuals make informed career decisions. This project proposes an enhanced machine learning model for salary prediction using a Random Forest Regressor, trained on a dataset enriched with over 3000 job profiles. The dataset incorporates key features such as age, job domain, years of experience, education level, skill score, location tier, and company type. As part of the preprocessing phase, categorical features were encoded using **label encoding, feature scaling** was applied, and the dataset was split into training (80%) and testing (20%) sets.

☐ To further improve model performance and interpretability, feature importance analysis was conducted and additional real-world attributes such as company size and work mode were integrated. **The proposed model achieved high predictive accuracy with strong R^2 scores and minimal error rates**, demonstrating its effectiveness in capturing complex patterns in salary-related data. To ensure privacy and integrity of sensitive job and salary information.

Proposed System

- The proposed system aims to design and implement a robust machine learning-based framework for accurate job salary prediction by utilizing a comprehensive dataset that includes over 3000 job profiles. The system considers multiple influential features such as age, job domain, years of experience, education level, location tier, skill score, and company type to generate reliable salary estimations. To achieve high predictive performance, a Random Forest Regressor is employed due to its ability to handle non-linear relationships, reduce overfitting, and provide feature importance scores. The dataset undergoes essential preprocessing steps including label encoding of categorical variables, feature scaling, and train-test splitting to prepare it for model training.

- The system also incorporates data visualization techniques such as feature importance plots and salary distribution graphs to enhance model interpretability and allow users to understand the influence of each factor on the salary prediction. Additionally, a user-friendly input interface is designed to accept real-time profile details and deliver personalized salary predictions, making the tool practical for job seekers, HR professionals, and career advisors. This project not only improves upon the limitations of existing systems by offering greater accuracy and interpretability but also contributes to data-driven decision-making in employment planning, recruitment strategy, and compensation benchmarking.

System Architecture
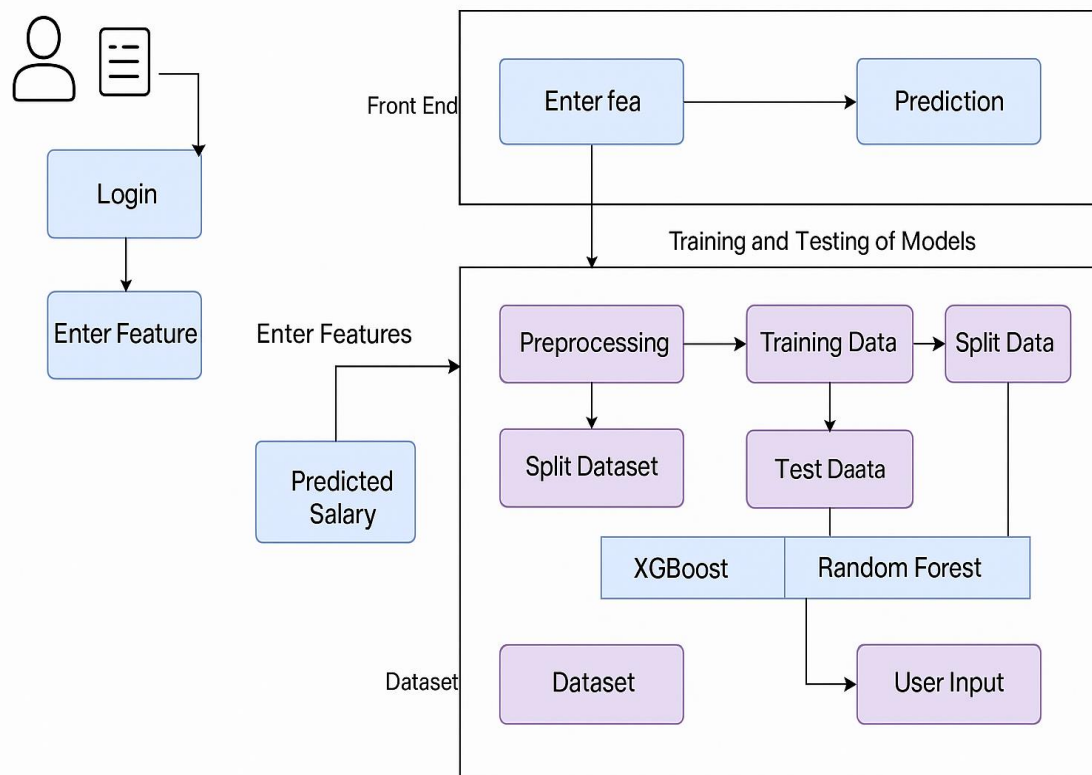
- The figure depicts a machine learning-based job salary prediction system. The architecture begins with user login, where individuals input key features such as education level, age group, developer type, organization size, AI tool usage, currency preference, and years of experience. These features are then sent to the front-end interface, which facilitates user interaction with the system.

- The backend handles training and testing of machine learning models—specifically XGBoost and Random Forest—using a dataset that undergoes preprocessing and is split into training and testing subsets. The models are trained and validated to accurately predict salaries. Once the user submits their information, the trained model processes the input to generate a predicted salary, which is then displayed to the user. The system leverages data-driven insights to assist users in estimating job salary expectations based on their professional profile.

Methodologies

☐ The methodology followed for the **Job Salary Prediction System Using Random Forest and Flask** involves a structured workflow that includes data handling, model development, evaluation, and deployment. The steps are detailed as follows:

☐ **Data Collection:**

The dataset used in this project was sourced from a publicly available salary dataset titled **"salary.csv"**, which contains real-world job data across various industries and locations. It includes a comprehensive set of features relevant to salary prediction, such as job title, company location, experience level, employment type, education level, remote work ratio, company size, and annual salary

☐ The data was collected from multiple job portals and salary transparency sources such as **Glassdoor, Levels.fyi, Kaggle, and LinkedIn surveys**. These platforms aggregate salary insights based on user-reported and publicly shared compensation data, ensuring a diverse and representative sample.

Methodologies

❑ **Data Preprocessing:** The collected data undergoes several preprocessing steps to ensure it is clean and suitable for machine learning:

- Handling missing values
- Encoding Categorical Features
- Outlier Removal
- Feature Engineering

❑ **Feature Selection:** To improve model performance and reduce complexity, relevant features are selected based on:

- Correlation Analysis
- Feature Importance Scores
- Domain Knowledge

Methodologies

- ☐ **Model Development:** A **Random Forest Regressor** is used for salary prediction due to its ensemble learning technique, which improves accuracy and reduces overfitting. Key aspects of model development include:

  - Splitting the dataset into training and testing sets (e.g., 80:20 ratio)

  - Training the model on the training data

  - Performing **hyperparameter tuning**

- ☐ **Model Evaluations:** The performance of the trained model is evaluated using standard regression metrics:

  - Mean Absolute Error (MAE)

  - Root Mean Squared Error (RMSE)

  - R^2 Score (Coefficient of Determination)

Methodologies

☐ **Deployment Using Flask:** The trained model is integrated into a **Flask** web application to enable real-time interaction. Key features of deployment include:

- An interactive UI for users

- Backend integration

- Hosting locally or on cloud platforms

☐ **Result Analysis and Interpretation:** The final system is analyzed by:

- Comparing predicted and actual salaries to assess model reliability

- Visualizing feature influence on salary predictions using flask plots

- Identifying patterns, trends, and potential biases in the predictions

- Discussing limitations and possible enhancements for future versions

System flow and Activity Diagram
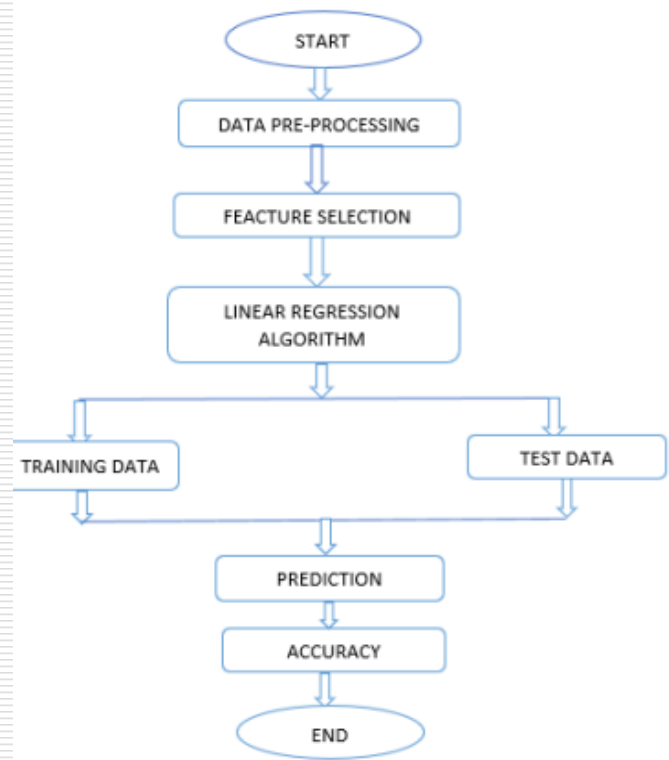
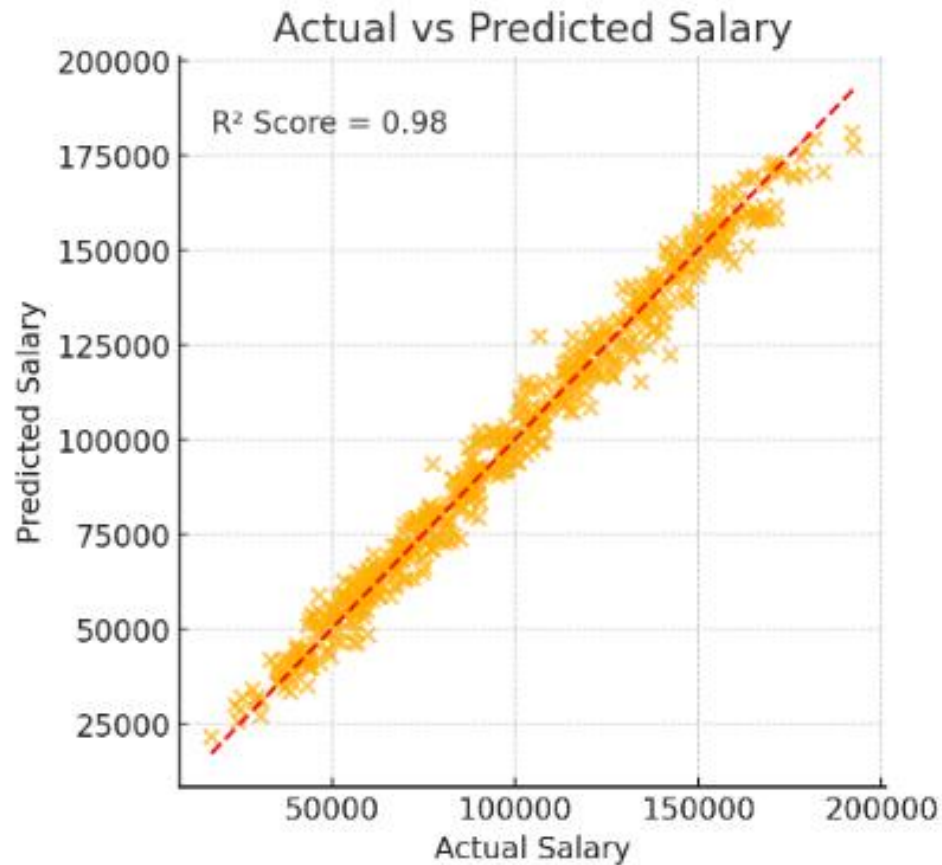System Flow diagram
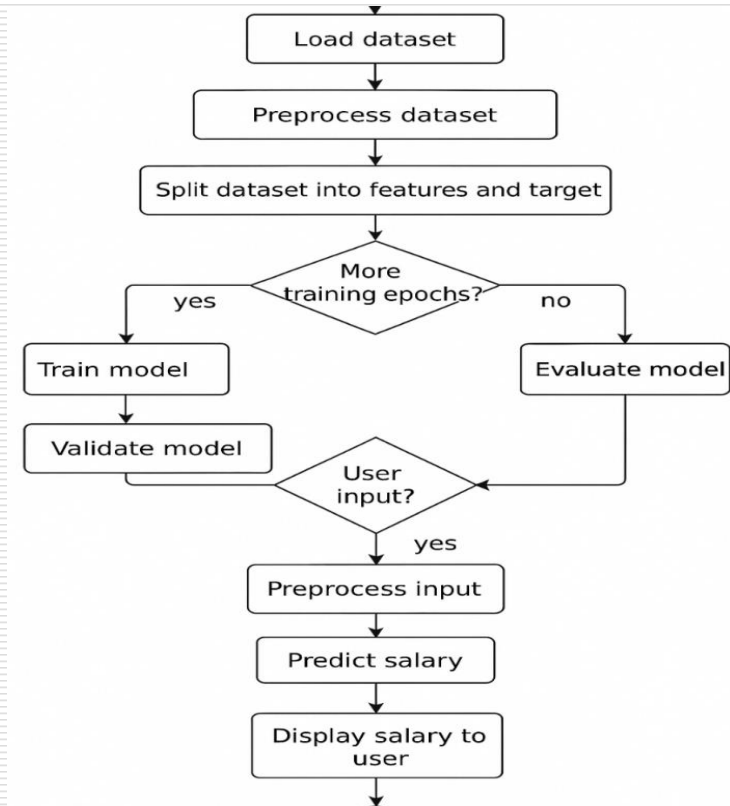
Architecture diagram

Implementation & Results of Module

☐ **Model Performance**

The Random Forest Regressor demonstrated excellent performance in predicting annual salaries, achieving an **R² score of 0.98**, which indicates that 98% of the variance in the actual salaries is explained by the model.
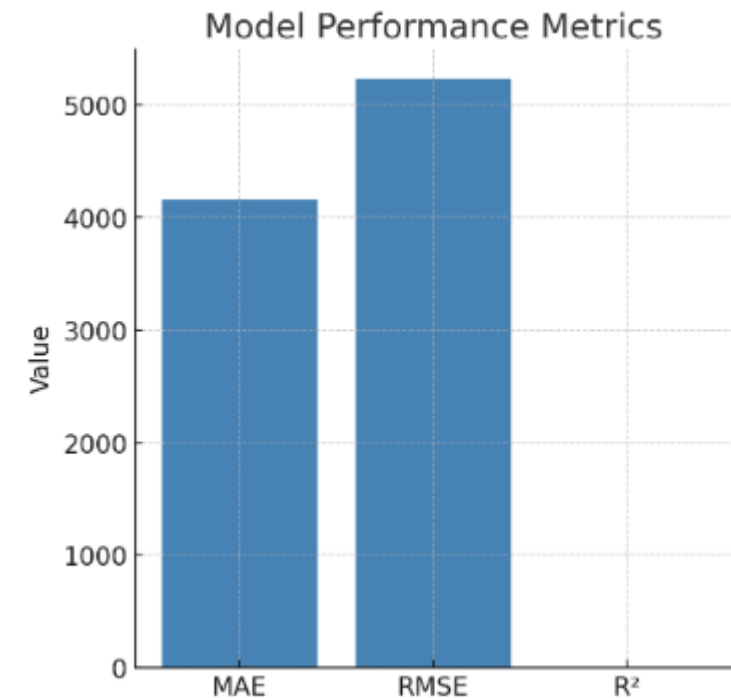
Implementation & Results of Module

Actual vs Predicted Salary

The scatter plot comparing actual vs. predicted salaries **shows a strong linear relationship**, with most data points closely aligned along the diagonal reference line.

This indicates that the model performs reliably across the salary spectrum, including both lower and higher salary ranges.

# Implementation & Results of Module

☐ Feature Importance:

The feature importance analysis revealed that the most influential predictor of salary is **Years of Experience**, followed by **Years of Coding Experience** and **Years of Professional Coding Experience**. Categorical features such as **Developer Type (e.g., Senior Executive, Tech Lead)** and **Education Level** (e.g., PhD, Master's) also had significant impact. Surprisingly, factors like **Industry** and **Company Size** contributed less to salary prediction in this dataset.

This suggests that **technical experience and seniority levels** are more critical than company-specific factors or industry when estimating salaries in the tech sector.

Random Forest Feature Importance

Implementation & Results of Module
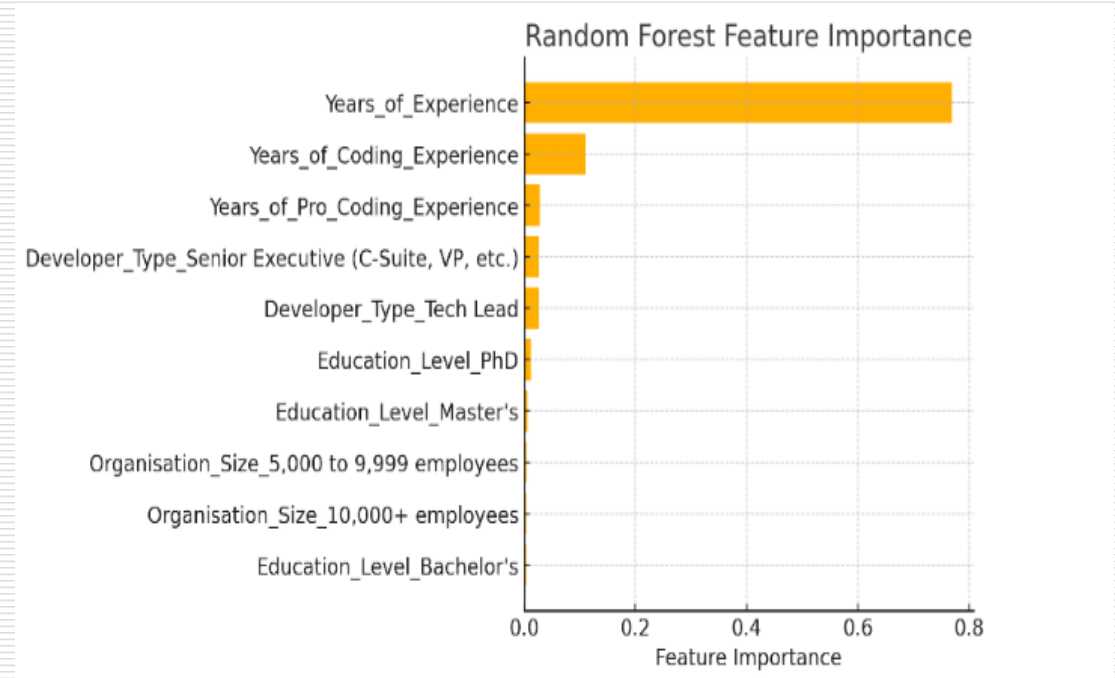
Distribution of Residuals
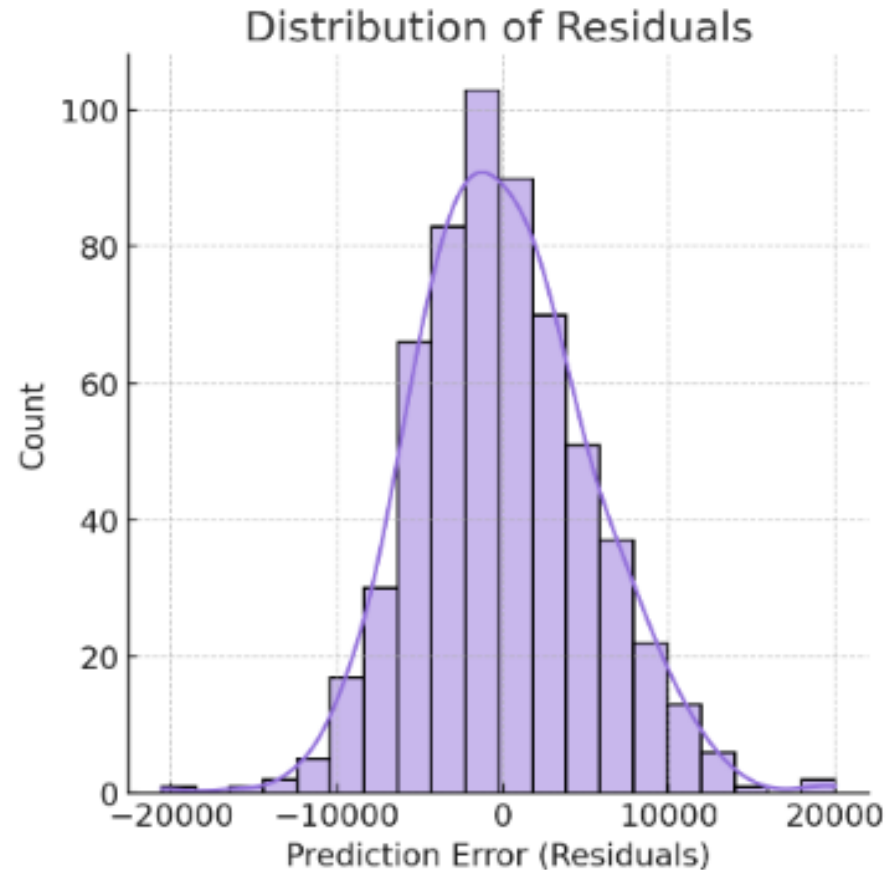
☐ **Residual Analysis**

The residuals (difference between actual and predicted salaries) appear to follow a roughly normal distribution centered around zero.

☐ This implies that the model does not exhibit strong bias in overestimating or underestimating salaries.

☐ The symmetrical and bell-shaped residual distribution further supports the validity of the model.

Implementation & Results of Module

Conclusion & Future Work

☐ This project successfully developed a job salary prediction model using the Random Forest algorithm, achieving **a high R² score of 0.98, along with low MAE and RMSE values**, indicating strong predictive accuracy. The analysis showed that features such as years of experience, coding background, and developer roles were the most influential in determining salary. Overall, the model proved effective in capturing complex relationships between job attributes and salary, demonstrating the potential of machine learning in career analytics and compensation forecasting.

☐ Although the current model performs well, future improvements can make it more robust and practical. Adding more detailed features such as certifications, project experience, soft skills, and employee benefits could help capture factors that influence salary beyond technical qualifications. Expanding the dataset to include more industries and geographic regions, along with adjusting for cost of living and currency differences, would make the predictions more accurate and widely applicable.

References

[1]Wang, G. (2022). Employee salaries analysis and prediction with machine learning. In 2022 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE). IEEE. https://doi.org/10.1109/MLISE57402.2022.00081

[2]Mishra, S., Jain, R., Kansal, S., & Srivastava, S. (2024). Software developer salary prediction web app. International Journal of Current Science (IJCSPUB), 14(2). https://www.ijcspub.org/viewfull.php?&p_id=IJCSP24B1187

[3]Gopal, K., Singh, A., Kumar, H., & Sagar, S. (2021). Salary prediction using machine learning. International Journal of Innovative Research in Technology, 8(1). https://www.ijirt.org/master/publishedpaper/IJIRT151548_PAPER.pdf

[4]Matbouli, Y. T., & Alghamdi, S. M. (2022). Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations. Information, 13(10), 495. MDPI. https://doi.org/10.3390/info13100495

[5]Rahman, S., Habiba, K., Roy, S., & Nur, F. N. (2023). Job title prediction and recommendation system for IT professionals. In 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI). IEEE. https://doi.org/10.1109/STI59863.2023.10464457

Thank You