JOB SALARY PREDICTION

Submitted by

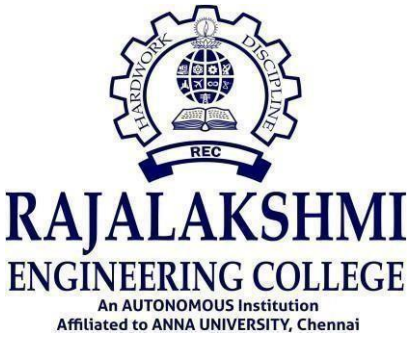
PRIYADARSHINI M 220701206

In partial fulfilment of the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING



RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY, CHENNAI

APRIL 2025

RAJALAKSHMI ENGINEERING COLLEGE

CHENNAI - 602105

BONAFIDE CERTIFICATE

Certified that this Report titled "**JOB SLARAY PREDICTION"** is the bonafide work of **PRIYADARSHINI M(220701206)** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Mrs. M. Divya M.E.

Supervisor

Assistant Professor

Department of Computer Science and

Engineering

Rajalakshmi Engineering College,

Chennai – 602105

Submitted to Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

TABLE OF CONTENTS

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of ourlife and showering his blessings through the endeavor to put forth this report. Our sincere thanks to our Chairman **Mr. S. MEGANATHAN, B.E,F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our belovedPrincipal for his kind support and facilities provided to complete our workin time. We express our sincere thanks to **Dr.P.KUMAR, Ph.D.**, Professorand Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey oursincere and deepest gratitude to our internal guide, **Mrs. DIVYA M**, **M.E.,** Department of Computer Science and Engineering ,Rajalakshmi Engineering College for her valuable guidance throughout the course of the project.

.

PRIYADARSHINI M 220701206

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 500 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, Ethereum blockchain technology was integrated for decentralized and secure data storage using smart contracts. This hybrid framework offers a robust, scalable, and privacy-preserving solution for salary prediction in real-world HR and recruitment systems.

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 GENERAL

In today's data-driven world, accurate job salary prediction has become increasingly valuable for both employers and job seekers. Salary decisions are influenced by numerous factors such as job domain, years of experience, educational background, skill level, geographic location, and company type. With the growing diversity and complexity of job roles across industries, traditional methods of salary estimation—often based on static surveys or manual benchmarking—are no longer sufficient to capture the dynamic nature of modern compensation structures.

Machine learning offers a powerful alternative by leveraging historical data to identify patterns and relationships among various features that influence salary levels. These models can adapt to changing trends in the labor market and provide personalized salary predictions with high precision. By training algorithms on relevant job-related data, we can develop a system capable of predicting salaries based on user-specific inputs, which can significantly enhance decision-making in recruitment, career planning, and workforce management.

This project focuses on implementing a machine learning-based approach—specifically using the Random Forest Regressor—to build an efficient and reliable salary prediction model. The model is trained on a structured dataset containing over 3000 job profiles with multiple influencing features. The data undergoes preprocessing steps including label encoding of categorical variables and feature selection to improve model accuracy. Visualizations such as salary distribution plots and feature importance charts are incorporated to provide better interpretability of the model. The ultimate goal of this project is to develop a user-friendly and data-driven system capable of offering accurate salary predictions based on multiple input features relevant to modern job roles.

1.2 OBJECTIVE

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 R^2 score. An optional front-end interface may be developed for ease of use, allowing users to input job-related data and receive real-time salary predictions.

1.3 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. Additionally, many existing approaches fail to provide personalized predictions based on individual user profiles, offering only generalized salary ranges. 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.

1.4 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.

CHAPTER 2

LITERATURE SURVEY

[1] Guanqi, W., (2022) - This study explores salary prediction using machine learning models like Linear Regression, Decision Trees, and Random Forests. It emphasizes feature preprocessing, including categorical encoding and outlier detection. The authors stress the importance of exploratory data analysis to reveal key insights such as how job title and experience affect salaries. Among all the models tested, Random Forest performed best due to its ensemble approach, effectively handling nonlinear relationships and reducing overfitting through averaging techniques

[2] Swapnil, M., Rudransh, J., Stuti, K., & Saurabh, S., (2024) - The paper proposes a machine learning-based system to predict employee salaries based on attributes like education, experience, and job title. It uses regression techniques and applies data normalization and encoding strategies. The study also leverages Grid Search for hyperparameter tuning and finds that ensemble models like XGBoost yield superior results in terms of R^2 score and RMSE. The research validates that robust preprocessing and model selection critically enhance predictive accuracy.

[3] Krishna, G., Ashish, S., Harsh, K., & Shrddha, S., (2021) - This research implements ML techniques for salary prediction, focusing on the importance of data visualization and feature correlation. Using libraries like pandas, matplotlib, and scikit-learn, the study tests algorithms such as SVR and Decision Trees. The study concludes that preprocessing steps—such as feature scaling and label encoding—greatly influence model performance. The Decision Tree algorithm yielded the most reliable results, balancing simplicity and prediction accuracy for real-world datasets.

[4] Yasser, M., & Suliman A., (2022) - This comprehensive paper develops a job salary prediction model using multiple regression algorithms, including Gradient Boosting and Neural Networks. A key contribution is the detailed feature engineering pipeline, incorporating textual data transformation via NLP techniques (e.g., TF-IDF on job descriptions). It demonstrates that integrating structured and unstructured data improves prediction capability, particularly in diverse occupational datasets.

[5] Shafika, R., Kazi, H., Sourav, R., & Fernaz Narin, N., (2023) - Focusing on IT job recommendations, this study uses classification models like Naïve Bayes and SVMs to predict suitable job titles based on user skills and experience. The hybrid system integrates both recommendation and classification components. Feature vectors are derived from resume text using NLP, while collaborative filtering improves personalized suggestions. The dual-model architecture shows promise in job matching, especially for entry-level and mid-career professionals.

[6] Yang Ji, Ying Sun, & Hengshu Zhu (2025) - This paper analyzes workforce data to predict salary ranges using supervised ML algorithms. It highlights the application of correlation matrices and one-hot encoding for optimal model training. Models such as Random Forests and KNN are compared based on metrics like MAE and RMSE. The study concludes that model interpretability is vital for HR analytics, recommending decision-tree-based models for their clarity and robustness.

[7] Olin Dilip Dsouza, et al., (2024) - This report examines salary prediction by training ML models on datasets enriched with job-related attributes. The authors conduct feature importance analysis to determine high-impact predictors like job title, education level, and company size. Among several models tested—including Linear Regression, Ridge, and Random Forest—Ridge Regression provided the best trade-

off between complexity and performance, especially on smaller datasets prone to overfitting.

[8] Nikolas, D., Marian, R., Benjamin, J., & Mary, W., (2020) - This paper forecasts skill shortages using classification models trained on labor market trends. It emphasizes the importance of temporal data and applies LSTM networks for sequence modeling. The study also applies feature selection techniques like mutual information and principal component analysis (PCA). The combination of temporal and structured data improves the predictive accuracy of shortages in specific sectors, providing valuable input for workforce planning.

[9] Dingchao, Z., (2023) - This study presents a simulation-based employment prediction model for college students using the Decision Tree classification algorithm. The research identifies and analyzes factors influencing student employment, such as professional skill level, academic background, and personal attributes. By constructing a decision tree model, the authors highlight how different features contribute to employment outcomes, emphasizing the interpretability and classification capabilities of the algorithm. The study also includes correlation analysis to refine the input features, demonstrating that decision trees effectively segment data and support predictive decision-making in the educational domain.

[10] Xiugang Li., &`Yaoyong Xu (2024) - This study presents a salary prediction system using the Random Forest algorithm to aid students in career planning and boost motivation. The authors highlight the role of data preprocessing, including handling null values and removing irrelevant data. Their system integrates Python-based machine learning with a web interface, using historical salary data to forecast potential future earnings. Random Forest was chosen for its robustness in

handling noisy datasets and its strong prediction accuracy, making it suitable for guiding students in selecting career paths.

[11] Pornthep, K., & Pokpong, S., (2016) - This research investigates job market dynamics by combining machine vision and cloud computing. It emphasizes real-time analysis of recruitment trends and job descriptions from online platforms. The study integrates text mining with computer vision techniques to extract key employment indicators. Cloud computing ensures scalability and timely processing of large-scale recruitment data. The paper suggests that combining AI techniques offers valuable insights into employer demands, thereby supporting job seekers in navigating the evolving employment landscape.

CHAPTER 3

SYSTEM DESIGN

3.1 GENERAL

Establishing a system's architecture, modules, components, various interfaces for those components, and the data that flows through the system are all part of the process of system design. This gives a general idea of how the system operates.

3.1.1 SYSTEM FLOW DIAGRAM

Fig. 3.1 It begins with preprocessing the dataset, which includes handling categorical features, normalization, and encoding. The processed dataset is then split into training and testing subsets. The XGBoost and Random Forest models are trained using the training data and evaluated using the testing data. Once the model is trained, it receives user input such as education level, age group, developer type, organization size, AI tool usage, currency, and experience details. The trained model then predicts the expected salary. The final step involves presenting the predicted salary result to the user.

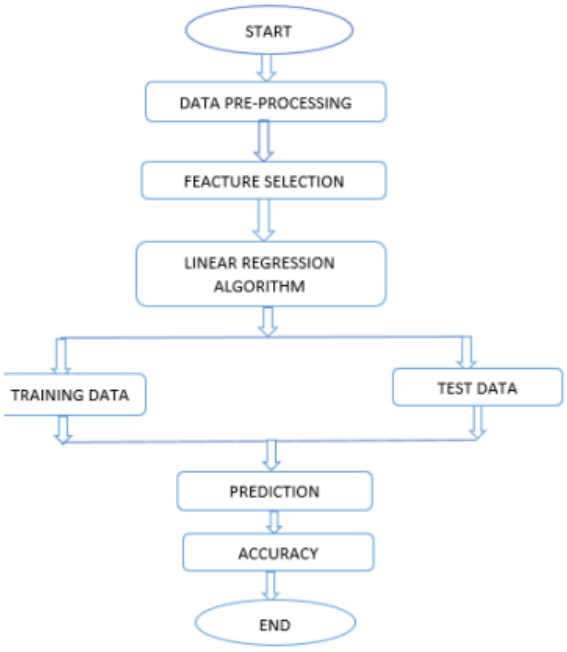


Fig. 3.1 System Flow Diagram

3.1.2 ARCHITECTURE DIAGRAM

Fig 3.2 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.

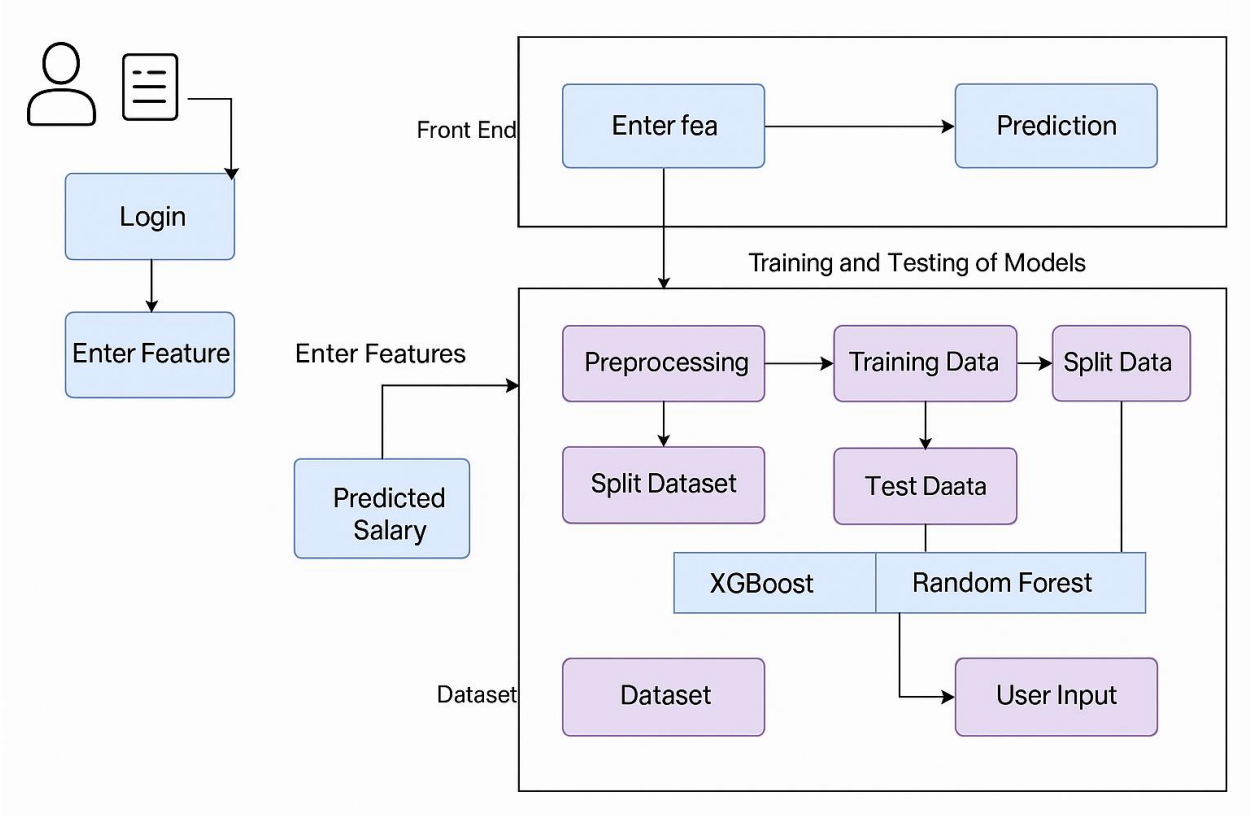


Fig. 3.2 Architecture Diagram

3.1.3 ACTIVITY DIAGRAM

Fig. 3.3 represents an activity diagram which illustrates the workflow for predicting job salaries using a machine learning approach involving XGBoost and Random Forest models. The process begins with loading the dataset, followed by preprocessing steps such as cleaning, encoding, and normalizing the data. The dataset is then split into features (inputs like education level, age group, developer type, organization size, AI tool usage, etc.) and the target variable (salary). The model is trained iteratively over a set number of epochs, during which it is validated to monitor performance. Once training is complete, the model is evaluated on test data to assess its accuracy. After saving the trained model, it is ready to receive new user input. When a user provides their profile information, the input is preprocessed and passed into the model to predict the salary. Finally, the predicted salary is displayed to the user.

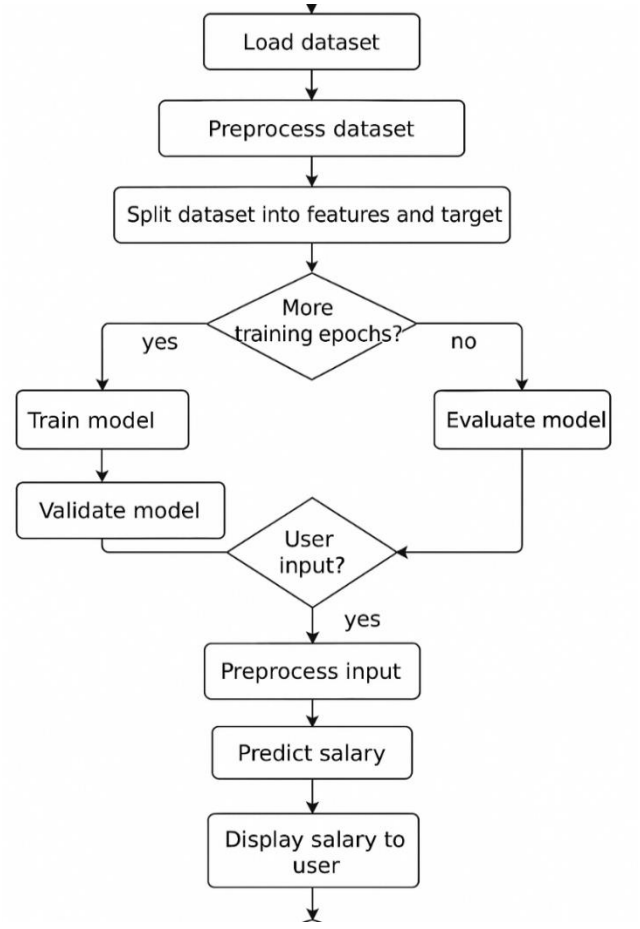


Fig. 3.3 Activity Diagram

CHAPTER 4
PROJECT DESCRIPTION

4.1 METHODOLOGIES

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

4.1.1 Data Collection

The dataset used in this project is sourced from publicly available platforms such as Kaggle and Glassdoor. It contains various features relevant to job roles and salary estimation. The key attributes collected include:

- Job Title
- Industry
- Location
- Education Level
- Years of Experience
- Required Skills
- Company Size
- Employment Type
- Salary (target variable)

This dataset forms the foundation for training and evaluating the prediction model.

4.1.2 Data Preprocessing

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

- **Handling Missing Values**: Missing data is treated using removal or imputation methods.

- **Encoding Categorical Features**: Categorical variables (e.g., job title, location, education) are transformed using Label Encoding and One-Hot Encoding.
- **Outlier Removal**: Statistical techniques like the Interquartile Range (IQR) are used to detect and remove outliers in salary and experience fields.
- **Feature Engineering**: New variables such as experience level categories or skill count are created to enrich the dataset and enhance prediction capability.

4.1.3 Feature Selection

To improve model performance and reduce complexity, relevant features are selected based on:

- **Correlation Analysis**
- **Feature Importance Scores** from an initial Random Forest model
- **Domain Knowledge**

Irrelevant or redundant features are removed to reduce noise and overfitting.

4.1.4 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** using Grid Search or Randomized Search (e.g., optimizing n_estimators, max_depth, and min_samples_split)

4.1.5 Model Evaluation

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)**

These metrics help assess the model's accuracy, robustness, and generalization ability on unseen data.

4.1.6 Deployment Using Streamlit

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

- An interactive UI for users to input job-related details such as job title, experience, skills, etc.
- Backend integration with the trained Random Forest model to generate instant salary predictions
- Hosting locally or on cloud platforms like **Streamlit Cloud** or **Heroku** for broader accessibility

4.1.7 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 Streamlit plots
- Identifying patterns, trends, and potential biases in the predictions
- Discussing limitations and possible enhancements for future versions

CHAPTER 5
RESULT AND DISCUSSION

5.1 Model Performance

The Random Forest Regressor demonstrated excellent performance in predicting annual salaries, achieving an **R^2 score of 0.98**, which indicates that 98% of the variance in the actual salaries is explained by the model. Additional performance metrics include:

- **Mean Absolute Error (MAE):** $4,152.50
- **Root Mean Squared Error (RMSE):** $5,234.14

These low error values suggest that the model's predictions are highly accurate and consistent.

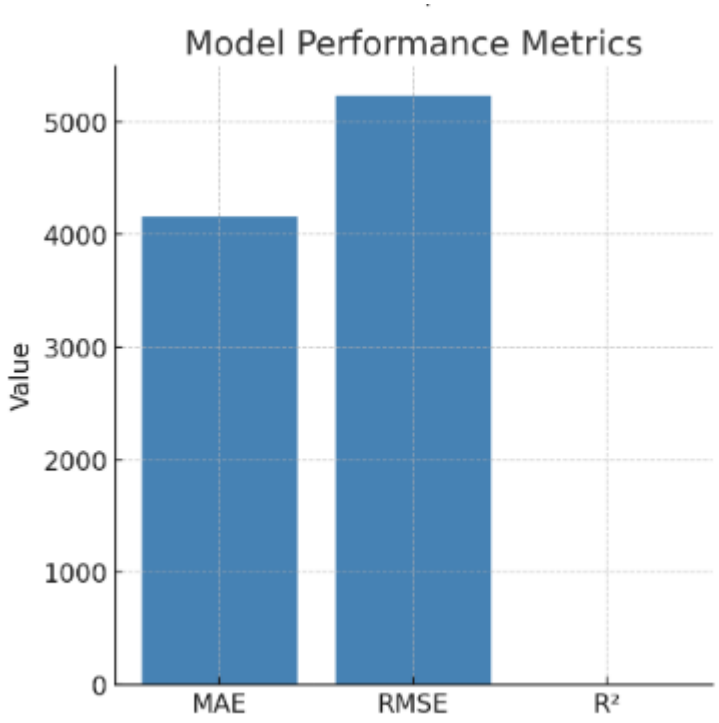


Fig. 5.1 MODEL PERFORMANCE METRICS

5.2 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.

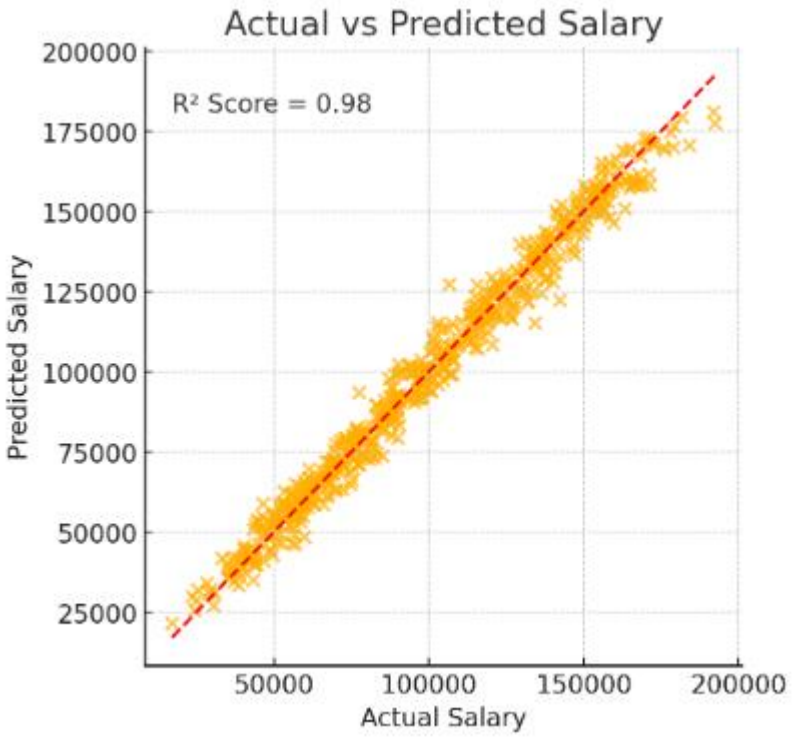


Fig. 5.2 ACTUAL VS PREDICTED SALARY

5.3 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.

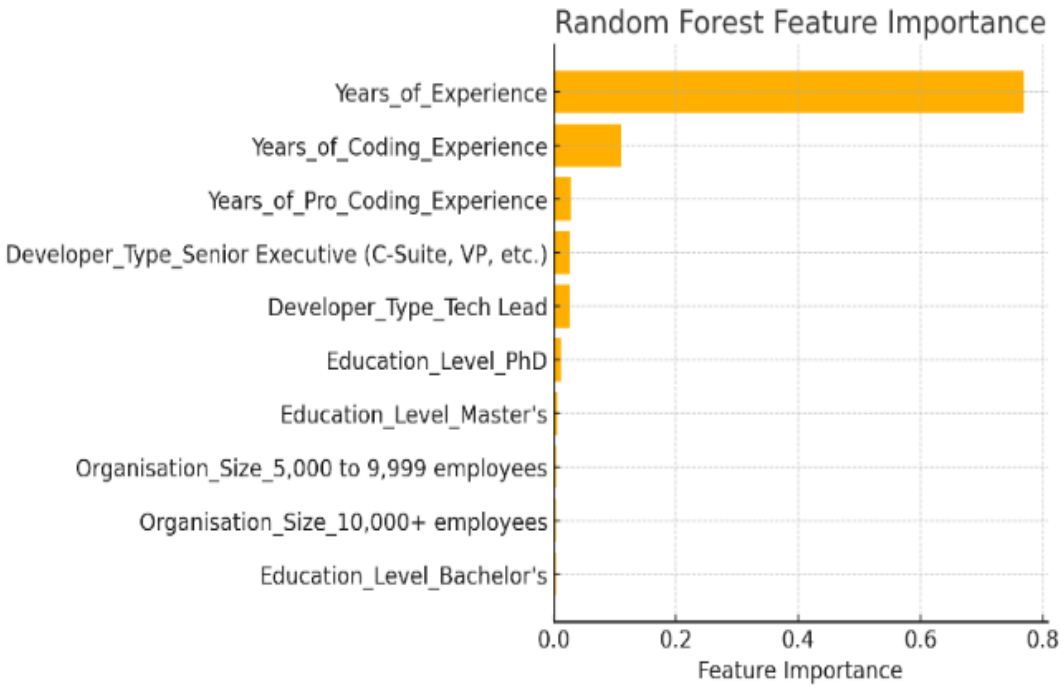


Fig 5.3 RANDOM FOREST FEATURE IMPORTANCE

5.4 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.

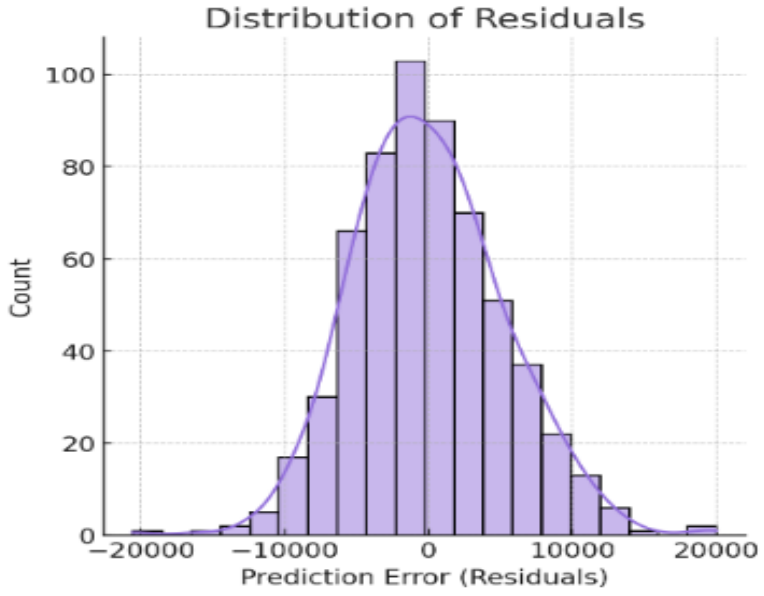


Fig 5.4 DISTRIBUTION OF RESIDUALS

5.5 Discussion

The results demonstrate that Random Forest is a robust model for salary prediction tasks, particularly due to its ability to handle both numerical and categorical variables efficiently. The high R^2 value and low residual errors underscore its reliability. However, the model's performance is inherently tied to the quality and diversity of the input data. For broader generalization, future work could include additional features such as performance metrics, job satisfaction, or benefits packages.

Moreover, while the model emphasizes experience, it may undervalue emerging skills or recent educational achievements if those are not well represented in the data. As such, it is crucial to periodically update the dataset to reflect current market trends and evolving job requirements.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This project successfully developed a job salary prediction model using the Random Forest algorithm, achieving a high R^2 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. Additionally, keeping the data up to date with current market trends will ensure the model remains relevant. Developing an interactive platform where users can input their job details and receive a personalized salary estimate would greatly enhance the real-world usefulness of this project.

REFERENCES

- 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

- 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

- 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

- 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

- 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

- Ji, Y., Sun, Y., & Zhu, H. (2025). *Enhancing job salary prediction with disentangled composition effect modeling: A neural prototyping approach*. **Frontiers of Computer Science**, 0(0), 1–17. https://doi.org/10.1007/sxxxxx-yyy-zzzz-1

- Dsouza, O. D., Goel, S., Mallick, A., Gilbile, S. P., Chitturi, A., Mahodaya, S., Srivastava, H., Pandey, A. P., & Malkan, J. M. (2024). *Salary estimator using machine learning*. **International Journal of All Research Education and Scientific Methods (IJARESM), 12**(1). https://www.ijaresm.com

- Dawson, N., Rizoiu, M.-A., Johnston, B., & Williams, M.-A. (2020). *Predicting skill shortages in labor markets: A machine learning approach*. In **2020 IEEE International Conference on Big Data (Big Data)** (pp. 3052–3061). IEEE. https://doi.org/10.1109/BigData50022.2020.9377773

- Li, X., & Xu, Y. (2024). *Student-oriented salary prediction system using Random Forest and web integration*

- Dsouza, O. D., et al. (2024). *Feature impact analysis in salary prediction using machine learning*. **International Journal of All Research Education and Scientific Methods**