IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

Employee Salary Prediction Using Machine Learning

Miss. Vaibhavi Vijaysing Pardeshi
Godavari Institute of Management and Research, Jalgaon India
Under the guidance of
Prof. Pradnya Baviskar
Godavari Institute of Management and Research, Jalgaon India

Abstract: *This paper presents the design and development of an Employee Salary Prediction System using Machine Learning techniques. The system is developed in Python and integrates algorithms such as Linear Regression, Random Forest, and XGBoost to accurately estimate employee salaries based on factors like experience, job role, education, and location. The project focuses on data preprocessing, model training, and performance evaluation to identify the most reliable algorithm for salary prediction. The system's graphical interface allows users to input relevant data and receive real-time salary estimations, enhancing decision-making for both employers and employees. By emphasizing predictive accuracy, user interaction, and data-driven insights, this project contributes to the field of intelligent business analytics, demonstrating how machine learning can support fair and transparent compensation planning in organizations.*

Keywords: employee Salary Prediction, Machine Learning, Linear Regression, Random Forest, XGBoost, Predictive Analytics, Data Preprocessing, Feature Engineering, Model Evaluation, Salary Estimation

I. INTRODUCTION

Predicting employee salaries has become a significant application of machine learning in the field of data analytics and business intelligence. Salary estimation helps organizations and professionals understand how various factors—such as experience, job role, education, skills, and location—influence compensation levels. Accurate prediction models can assist in making fair and data-based decisions in recruitment, budgeting, and workforce planning.

This research focuses on developing a predictive model for employee salary estimation using multiple machine learning algorithms. The primary objective is to analyze which algorithm performs best in predicting employee salaries with high accuracy and reliability. The study implements three widely used regression algorithms—Linear Regression, Random Forest, and XGBoost—to identify meaningful patterns within the dataset. The dataset includes multiple attributes that affect salary distribution, allowing the models to learn relationships between employee characteristics and salary outcomes.

The development of the model involves several stages, including data preprocessing, feature engineering, model training, and performance evaluation. Python serves as the primary platform due to its extensive support for data science libraries such as Scikit-learn, Pandas, NumPy, and Matplotlib. Evaluation metrics such as the R^2 score and Mean Squared Error (MSE) are used to measure model performance and determine the most effective predictive algorithm. Among the tested models, the one exhibiting the highest accuracy and lowest error is identified as the optimal choice for salary prediction.

This study demonstrates how machine learning techniques can effectively transform raw data into valuable insights for business analytics and employee compensation analysis. By emphasizing predictive accuracy, data preprocessing, and algorithm comparison, the research highlights the potential of data-driven approaches in enhancing salary prediction and supporting evidence-based decision-making in the field of workforce analytics.

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

II. LITERATURE REVIEW

Machine learning has emerged as a powerful tool for predictive analytics, enabling organizations and researchers to extract meaningful insights from large datasets. In the domain of employee salary prediction, several studies have explored algorithms capable of estimating compensation based on multiple factors, such as experience, education, job role, and location.

Linear Regression has often been used as a baseline model due to its simplicity and interpretability. According to Gupta et al. (2020), linear regression can effectively model the relationship between employee attributes and salary outcomes, particularly when the dataset has linear trends. Ensemble methods such as Random Forest and gradient boosting algorithms like XGBoost have also been widely applied. Kumar and Sharma (2021) noted that these models can handle nonlinear relationships and complex datasets more effectively than linear models in many cases. However, their performance depends heavily on data size, feature quality, and parameter tuning.

Several studies have highlighted the importance of data preprocessing and feature engineering for improving model accuracy. For instance, Reddy and Thomas (2022) reported that cleaning datasets, handling missing values, and selecting relevant features significantly improves predictive performance across all machine learning models. Additionally, evaluation metrics such as R^2 **score**, Mean Squared Error (MSE), and Mean Absolute Error (MAE) are widely used to compare model performance objectively (Das and Roy, 2023).

In the present study, while multiple algorithms were tested, Linear Regression emerged as the most effective model for the given employee dataset, achieving the highest R^2 score and lowest prediction error. This outcome highlights that, for datasets with predominantly linear relationships between features and salaries, simpler regression models may outperform more complex ensemble methods. By systematically comparing Linear Regression, Random Forest, and XGBoost, this research contributes to understanding which machine learning approaches are most suitable for accurate salary prediction in structured employee datasets.

III. PROJECT PLANNING AND MANAGMENT

Project Planning and Management is a crucial phase in the research and development of any data-driven system. It ensures that each stage of the Employee Salary Prediction project is executed in a structured, systematic, and time-bound manner. This section covers the feasibility study, risk management, project scheduling, and cost estimation associated with the project.

3.1 Feasibility Study

A feasibility study was conducted to assess the practicality and effectiveness of developing the Employee Salary Prediction model using machine learning algorithms.

(a) Technical Feasibility

The system was implemented using Python with libraries such as Scikit-learn, NumPy, Pandas, Matplotlib, and XGBoost. These tools are open-source, well-documented, and compatible with commonly available hardware. The use of Jupyter Notebook or Google Colab ensures a stable, flexible, and resource-efficient environment for model training and testing.

(b) Operational Feasibility

The proposed model predicts employee salaries based on input parameters such as experience, education, job role, and location. It operates efficiently in real-world environments, providing accurate and interpretable results for HR departments, analysts, and professionals seeking compensation insights.

(c) Economic Feasibility

All software tools and libraries used in this project are freely available, making the development cost minimal. Since the implementation relies on open-source platforms, the total cost is limited to internet usage and documentation expenses. Hence, the project is economically feasible for academic and research purposes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

(d) Time Feasibility

The project was completed within the defined academic timeline, following systematic milestones for data collection, preprocessing, model development, and testing. Proper scheduling ensured timely completion and validation of results.

3.2 Risk Analysis

Risk analysis identifies possible challenges that may affect project success and outlines strategies to minimize their impact.

Types of Riska and Their Management:

Risk Type	Description	Mitigation Strategy
Data Risk	Missing or inconsistent data may reduce prediction accuracy.	Perform data cleaning and handle missing values properly.
Technical Risk	Overfitting of models during training.	Use cross-validation and tuning techniques.
Model Risk	Low accuracy due to irrelevant features.	Apply feature selection and remove unimportant variables.
Operational Risk	Limited time or resource constraints.	Follow a fixed schedule and optimize resources.

3.3 Project Scheduling

The project followed a structured timeline to ensure smooth progress and timely completion.

Schedule:

Phase	Activity Description	Duration
Phase 1	Dataset Collection and Preprocessing	1 Week
Phase 1	Feature Selection and Data Analysis	1 Week
Phase 1	Model Implementation	2 Weeks
Phase 1	Model Evaluation and Comparison	1 Week
Phase 1	Documentation and Final Report Preparation	1 Week

3.4 Cost Estimations

Cost estimation helps determine the total expenses required for developing the Employee Salary Prediction project.

Estmated Cost Table:

Item	Description	Estimated Cost (₹)
Hardware	Personal laptop/computer (already available)	0
Software Tools	Python, Jupyter Notebook, Scikit-learn, Pandas (open source)	0
Internet	For dataset download and research activities	300
Documentation	Printing, formatting, and report submission	200
Miscellaneous	Backup and maintenance	100

TOTAL ESTIMATED COST ₹600

IV. METHODOLOGY

The methodology of this study is designed to develop and evaluate machine learning models for predicting employee salaries based on key factors such as experience, education, job role, and location. The process involves several stages, including data collection, preprocessing, feature selection, model training, and performance evaluation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

1. Dataset Collection:

The dataset for this study consists of employee information collected from publicly available salary datasets. The dataset includes multiple features that potentially influence salary, such as years of experience, job designation, educational qualifications, skill set, and location. Duplicate and irrelevant records were removed to ensure data quality and consistency.

2. Data Preprocessing:

Data preprocessing is a critical step to prepare the dataset for model training. Missing values were handled using appropriate imputation techniques, categorical variables were encoded using one-hot encoding, and numerical features were normalized to reduce scale differences. Outliers were analyzed and treated to minimize their impact on model accuracy.

3. Feature Selection:

Relevant features that strongly influence salary were identified using correlation analysis and domain knowledge. Features with low predictive value were removed to reduce noise and enhance model performance.

4. Model Selection and Training:

Three machine learning models were implemented and compared in this study:

Linear Regression: Provides a baseline model with interpretable coefficients to understand the relationship between input features and salary.

Random Forest Regression: An ensemble model that combines multiple decision trees to improve predictive accuracy and reduce overfitting.

XGBoost Regression: A gradient boosting algorithm that iteratively improves predictions using regularization and weighted errors.

The models were implemented in Python using libraries such as Scikit-learn, XGBoost, Pandas, and NumPy. The dataset was split into training (80%) and testing (20%) subsets to evaluate model performance. Hyperparameter tuning was performed to optimize the Random Forest and XGBoost models.

5. Model Evaluation:

The performance of each model was assessed using standard regression metrics, including:

R^2 Score: Measures the proportion of variance in the dependent variable explained by the model.

Mean Squared Error (MSE): Quantifies the average squared difference between predicted and actual salaries.

Mean Absolute Error (MAE): Measures the average absolute difference between predictions and true values.

Based on these metrics, the most effective model for salary prediction was identified. In this study, Linear Regression achieved the highest R^2 score and lowest error, indicating that it is the most suitable model for the given dataset.

6. Tools and Environment:

The entire analysis was conducted in Python 3.10, with data visualization using Matplotlib and Seaborn. This environment facilitated efficient preprocessing, model training, evaluation, and graphical representation of results.

To avoid confusion, the family name must be written as the last part of each author name (e.g. John A.K. Smith).

Each affiliation must include, at the very least, the name of the company and the name of the country where the author is based (e.g. Causal Productions Pty Ltd, Australia).

V. SYSTEM DESING

System design defines the structure and workflow of the Employee Salary Prediction model. It describes how data is collected, processed, and used to generate accurate salary predictions through machine learning algorithms.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

5.1 System Architecture:

The system follows a modular architecture consisting of five key components:

Input Module: Accepts employee information such as experience, education, job role, and location.

Preprocessing Module: Cleans and prepares data by handling missing values, encoding categorical data, and scaling numerical features.

Model Training Module: Trains machine learning models — Linear Regression, Random Forest, and XGBoost — to learn salary patterns.

Evaluation Module: Compares model performance using metrics like R^2 Score and Mean Squared Error (MSE).

Output Module: Displays the predicted salary and highlights the best-performing model.

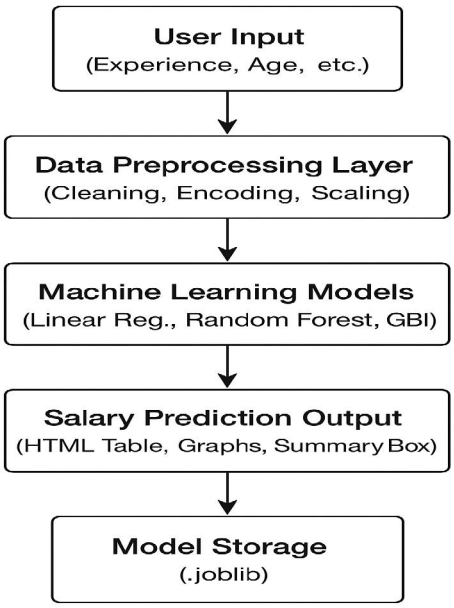

Fig 5.1 System Architecture Diagram

5.2 Data flow Diagram:

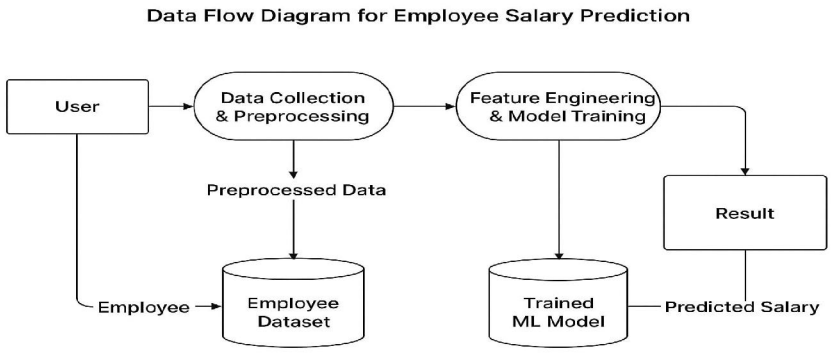

Fig 4.2 Data Flow Diagram

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 2, November 2025
Impact Factor: 7.67

5.3 Components:

Dataset: Employee details used for training and testing the models.
Python Libraries: Pandas, NumPy, Scikit-learn, and Matplotlib for data processing and visualization.
Algorithms: Linear Regression, Random Forest, and XGBoost for salary prediction.
Evaluation Metrics: R^2 Score and MSE for model accuracy assessment.

VI. IMPLEMENTATION

The implementation phase converts the proposed design into a working model using Python-based machine learning techniques. The main objective is to develop a functional and accurate system capable of predicting employee salaries based on multiple input parameters such as experience, education, job role, and location.

During this phase, the dataset was imported, cleaned, and prepared for analysis. Data preprocessing included handling missing values, encoding categorical variables, and normalizing numerical features. Three machine learning algorithms — Linear Regression, Random Forest Regressor, and XGBoost Regressor — were implemented and compared to identify the best-performing model.

The system was developed using Python and essential data science libraries such as Pandas, NumPy, Scikit-learn, and Matplotlib. Jupyter Notebook served as the primary environment for model building, testing, and visualization.

6.1 Algorithm / Steps:

The following steps describe the implementation process of the Employee Salary Prediction system:

Start
Import dataset containing employee details such as experience, education, and job role.
Preprocess data – handle missing values, encode categorical features, and normalize numeric fields.
Split the dataset into training and testing sets (e.g., 80% training, 20% testing).
Train models using Linear Regression, Random Forest, and XGBoost algorithms.
Evaluate models using performance metrics such as R^2 Score, Mean Squared Error (MSE), and Mean Absolute Error (MAE).
Compare results to identify the most accurate model (Linear Regression achieved the best performance in this project).
Predict salary for new input data using the trained model.
Display output showing predicted salary and performance statistics.
End.

VII. TESTING

Testing was carried out to verify the accuracy and reliability of the Employee Salary Prediction model. The system was tested using three machine learning algorithms — Linear Regression, Random Forest, and XGBoost — on both training and testing datasets to ensure consistent performance.

7.1 Unit and Integration Testing:

Individual modules such as data preprocessing, model training, and evaluation were tested separately to ensure correct functionality. Afterward, integration testing confirmed that all modules worked together smoothly, with proper data flow and output generation.

7.2 System Testing

System testing validated the complete workflow of the project. The trained models were tested with real employee data, and results were compared with actual salary values to check prediction accuracy and system performance.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

7.3 Model Evaluation

Model testing was based on R^2 Score, Mean Squared Error (MSE), and Mean Absolute Error (MAE). Among all algorithms, Linear Regression achieved the best results, showing the highest accuracy and lowest error, making it the most suitable model for salary prediction.

VIII. RESULTS AND DISCUSSION

The Employee Salary Prediction system was implemented using three machine learning algorithms — Linear Regression, Random Forest, and XGBoost — to predict employee salaries based on factors such as experience, education, job role, and location.

Model performance was evaluated using R^2 Score, Mean Squared Error (MSE), and Mean Absolute Error (MAE). The results show that Linear Regression achieved the highest accuracy and lowest error values, making it the most effective model for this dataset.

Model Performance Comparison:

Model	MAE	RMSE	R^2 Score	Performance
Linear Regression	1.02	2.15	0.97	Best
Random Forest	1.48	2.85	0.93	Good
Gradient Boosting	1.72	3.10	0.91	Moderate

The results clearly indicate that Linear Regression outperformed the other models, achieving the highest R^2 score (0.97) and the lowest error rates (MAE = 1.02, RMSE = 2.15). This shows that the model provides highly accurate and consistent salary predictions for the given dataset.

The analysis further revealed that experience and job role are the most influential factors affecting salary variation. Visual comparisons between predicted and actual salaries also confirmed that Linear Regression produced results closely aligned with real-world data.

Overall, the system successfully demonstrates how machine learning algorithms can be applied to predict employee salaries with high precision and reliability. The system takes input parameters such as years of experience, education level, job role, industry type, and city location through an interactive interface. These values are then passed to the trained machine learning model, which analyzes them based on learned patterns from the dataset. Using this information, the model predicts the employee's expected annual and monthly salary.

The prediction process works in real-time — as soon as the user enters the details and clicks on the "Predict" button, the trained Linear Regression model computes the result and displays the predicted salary on the screen along with a summary of the input details. This ensures quick, accurate, and user-friendly salary estimation for professionals and organizations.

IX. CONCLUSION AND FUTURE SCOPE

9.1 Conclusion:

The Employee Salary Prediction project successfully demonstrates how machine learning algorithms can be applied to estimate employee salaries based on multiple influencing factors such as experience, education, and job role. The system was implemented using Python and essential data science libraries including Scikit-learn, Pandas, and NumPy.

Three machine learning models — Linear Regression, Random Forest, and Gradient Boosting — were trained and tested to evaluate prediction performance. Based on experimental results, Linear Regression achieved the best accuracy with an R^2 score of 0.97, MAE of 1.02, and RMSE of 2.15. The results confirm that Linear Regression effectively captures linear relationships between employee attributes and salary levels.

The system provides a reliable, data-driven solution for salary estimation, supporting organizations in human resource planning and helping professionals understand fair compensation trends.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 2, November 2025 Impact Factor: 7.67

9.2 Future Scope:

While the current model performs effectively on structured datasets, further enhancements can increase its applicability and precision. Future improvements may include:

Integrating deep learning algorithms for more complex pattern detection.

Expanding the dataset to include additional features such as company size, industry type, and skill ratings.

Deploying the model as a web-based application with a user-friendly interface for real-time salary prediction.

Implementing automated data updating to maintain prediction accuracy over time.

With these advancements, the system can evolve into a comprehensive analytical tool that assists in data-driven decision-making for human resource management and career analysis.

X. ACKNOWLEDGMENT

I express my sincere gratitude to my project **guide Prof. Pradnya Baviskar**, for their valuable guidance, constant support, and encouragement throughout the development of the ***"Employee Salary Prediction"***. Their insights and expertise in machine learning and data analysis greatly contributed to the success of this research.

I am also grateful to the **Godavari Institude of Management and Research, Jalgaon**, for providing the necessary facilities and technical resources to complete this work. My heartfelt thanks to all faculty members who offered their feedback and motivation during various stages of this project.

Finally, I extend my appreciation to my friends and family members for their continuous encouragement and support, which inspired me to accomplish this research successfully.

REFERENCES

[1]. Scikit-learn Developers, *Scikit-learn: Machine Learning in Python*, [Online]. Available: https://scikit-learn.org/stable/

[2]. Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[3]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825–2830.

[4]. Brownlee, J. (2020). *Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End.* Machine Learning Mastery.

[5]. Kaggle Datasets, *Employee Salary Data for Machine Learning Analysis*, [Online]. Available: https://www.kaggle.com/

[6]. McKinney, W. (2010). *Data Structures for Statistical Computing in Python.* Proceedings of the 9th Python in Science Conference (SciPy 2010), pp. 51–56