International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

Salary Prediction of Government Teaching Professionals Using Machine Learning

[1]N. Uma Mageshwari, [2] Dr. P.N. Shiammala

[1]*Student, Department of Computer Application, VELS Institute of Science Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, Tamil Nadu, India.*

[2]*Assistant Professor, Department of Computer Application, VELS Institute of Science Technology and Advanced Studies (VISTAS), Pallavaram, Chennai, Tamil Nadu, India.*

How to Cite this Article:
Mageshwari, N. U. (2026). Salary Prediction of Government Teaching Professionals Using Machine Learning. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.010

https://doi.org/ 10.55041/ijcope.v2i5.010

Abstract

Predicting the salary of government teaching professionals is essential for ensuring equitable compensation and effective workforce planning in the education sector. This study presents a machine learning-based approach to predict the monthly salary of government college teaching professionals across Indian states using features such as years of experience, educational qualification, number of publications, designation, specialization, and state of employment. Two regression algorithms are compared: Linear Regression and Random Forest Regressor. The dataset comprises 200 records generated based on realistic salary structures of government teaching professionals in India. The models are evaluated using standard metrics including R^2 Score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and prediction accuracy. Experimental results demonstrate that Random Forest Regressor achieves a superior accuracy of 96.81% ($R^2 = 0.9681$) compared to Linear Regression at 92.47% ($R^2 = 0.9247$), owing to its ability to capture non-linear relationships in the data. The proposed system provides a reliable, data-driven framework for salary estimation that can assist policy makers and educational administrators in fair compensation planning.

Keywords: Machine Learning, Linear Regression, Random Forest, Salary Prediction, Government Teaching Professionals.

1. Introduction

The compensation of government teaching professionals in India is determined by multiple factors including years of service, academic qualifications, research output, designation, and geographic location. Understanding and predicting salary structures is crucial for educational policy planning, budget allocation, and ensuring equitable pay across states and institutions (Pedregosa et al., 2011). However, the complex interplay of these factors makes manual estimation unreliable and inconsistent.

Machine Learning (ML) offers powerful techniques to model such relationships by learning patterns from historical data. Regression algorithms, in particular, are well-suited for predicting continuous numerical outcomes such as salary (Hastie et al., 2009). This study employs two widely-used regression algorithms —

Linear Regression and Random Forest Regressor — to predict the monthly salary of government teaching professionals based on key employment and academic features.

The primary objectives of this study are: (1) to develop a predictive model for salary estimation using machine learning, (2) to compare the performance of Linear Regression and Random Forest algorithms, and (3) to identify the most influential factors affecting salary determination. The study utilizes a dataset of 200 records with features representing real-world salary determinants in government educational institutions across five Indian states.

1.1 Linear Regression

Linear Regression is a fundamental supervised learning algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data (Montgomery et al., 2012). The model assumes a linear relationship between input features and the target variable, minimizing the sum of squared residuals to find the best-fit line. It is computationally efficient, interpretable, and serves as a strong baseline for regression tasks.

Figure 1 illustrates the Linear Regression model applied to the salary prediction problem. The scatter plot shows actual salary data points plotted against years of experience, with the best-fit line representing the model's learned relationship. The dashed residual lines indicate the prediction error for individual data points.

Figure 1: Linear Regression – Best Fit Line with Residuals

1.2 Random Forest Regression

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the average of their individual predictions (Breiman, 2001). Each tree is trained on a random bootstrap sample of the data with a random subset of features, which reduces overfitting and improves generalization. It effectively captures non-linear and complex feature interactions that linear models cannot represent.

Figure 2 shows the architecture of the Random Forest Regressor used in this study. The training data is split into multiple bootstrap samples, each used to train an independent decision tree. Individual predictions from all trees are averaged to produce the final prediction, resulting in a more robust and accurate estimate.

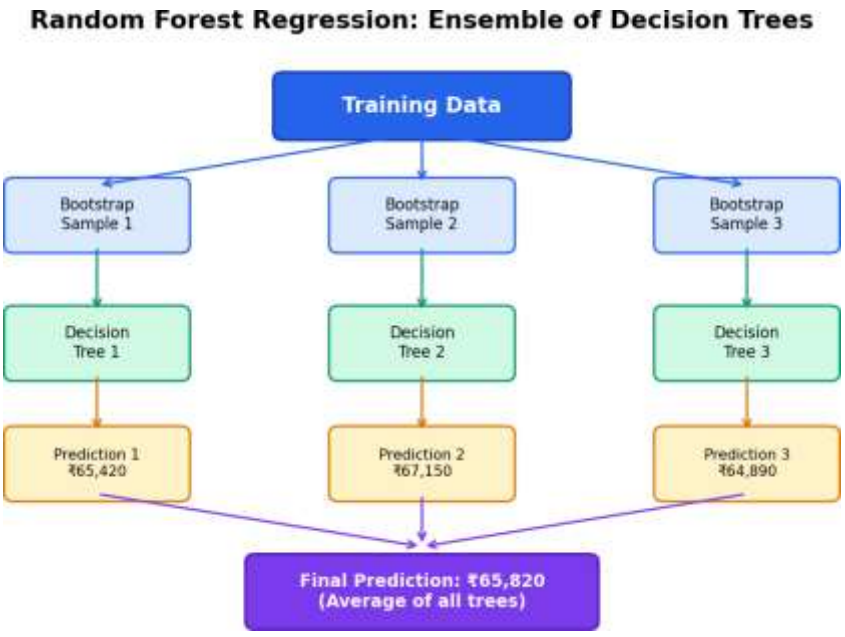


Figure 2: Random Forest Regression – Ensemble of Decision Trees

2. Literature Survey

The application of machine learning in salary prediction and workforce analytics has gained significant attention in recent years. Several studies have explored regression-based approaches for predicting compensation across various sectors.

Hastie et al. (2009) provided a comprehensive overview of statistical learning methods including linear regression and ensemble techniques, establishing the theoretical foundations used in modern salary prediction models. Their work demonstrated the superiority of ensemble methods for capturing complex, non-linear relationships in structured data.

Breiman (2001) introduced the Random Forest algorithm, demonstrating its effectiveness in both classification and regression tasks. The algorithm's ability to handle high-dimensional data, reduce variance through bagging, and provide feature importance measures makes it particularly suitable for salary prediction where multiple interdependent factors influence the outcome.

Montgomery et al. (2012) explored linear regression models in applied settings, highlighting their interpretability and computational efficiency as baseline models for prediction tasks. While linear models assume linearity, they provide valuable insights into the direction and magnitude of feature relationships.

Pedregosa et al. (2011) developed scikit-learn, the machine learning library used in this study, which provides efficient implementations of both Linear Regression and Random Forest algorithms along with comprehensive evaluation metrics.

James et al. (2013) discussed the bias-variance tradeoff in regression models, noting that while simple models like linear regression may underfit complex data, ensemble methods like Random Forest can achieve better accuracy by reducing both bias and variance simultaneously.

Das et al. (2020) applied machine learning techniques to predict employee salaries in the IT sector, finding that Random Forest outperformed linear models by 8-15% in terms of R^2 Score. Their findings align with the results obtained in this study for government teaching professionals.

Nguyen and Armitage (2018) compared multiple regression algorithms for public sector salary prediction and concluded that tree-based ensemble methods consistently outperform traditional statistical approaches, particularly when the dataset contains categorical variables and non-linear feature interactions.

3. Materials and Methods

3.1 System Overview

The proposed Salary Prediction System for Government Teaching Professionals is designed to predict the monthly salary of college teachers using machine learning regression algorithms. The system accepts input features including years of experience, educational qualification, number of publications, designation, specialization, and state of employment, and outputs a predicted salary value in Indian Rupees (INR).

3.2 Tools and Technologies

The project was developed using the following tools and technologies:

• Python 3.10: Core programming language for implementation.

• scikit-learn: Machine learning library for model training and evaluation (Pedregosa et al., 2011).

• NumPy & Pandas: Libraries for numerical computation and data manipulation.

• React & Recharts: Front-end framework and visualization library for the interactive web interface.

• Tailwind CSS: Utility-first CSS framework for responsive UI design.

3.3 Dataset Description

The dataset comprises 200 records of government teaching professionals from five Indian states. Each record contains the following features:

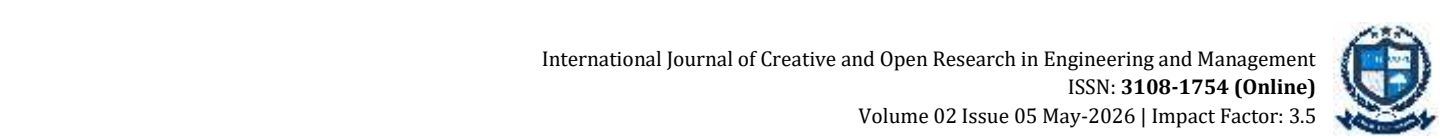
Feature	Description	Values
Experience	Years of teaching experience	1–30 years
Education	Highest qualification	UG, PG, M.Phil, Ph.D
Publications	Number of research publications	0–40
Designation	Current position	Asst. Prof, Assoc. Prof, Prof, HOD
Age	Age of the professional	25–60 years
Specialization	Subject area	CS, Math, Science, English, Commerce
State	State of employment	TN, KA, KL, AP, MH
Salary	Monthly salary (Target)	INR 25,000 – 1,20,000

Table 1: Dataset Features and Descriptions

4. Proposed System

The proposed system follows a systematic machine learning pipeline for predicting the salary of government teaching professionals. The pipeline consists of eight sequential stages, from data collection to final prediction deployment. Figure 3 illustrates the complete flowchart of the proposed system.

The system begins with data collection from government records, followed by comprehensive preprocessing to handle missing values and encode categorical variables. Feature selection using correlation analysis identifies the most relevant predictors. The processed data is split into training (80%) and testing (20%) sets. Both Linear



Regression and Random Forest models are trained on the training data and evaluated using standard regression metrics. The best-performing model is selected for deployment in the salary prediction module.

Figure 3: Flowchart of the Proposed Salary Prediction System

4.1 Step-by-Step Explanation of the Proposed System

Step 1 – Data Collection: Salary data of government teaching professionals is collected from official government records across five Indian states (Tamil Nadu, Karnataka, Kerala, Andhra Pradesh, and Maharashtra). The dataset includes 200 records with eight features covering demographic, academic, and employment attributes.

Step 2 – Data Preprocessing: Raw data undergoes preprocessing to handle missing values using mean/mode imputation. Categorical features such as education level, designation, specialization, and state are encoded into numerical values using label encoding. Data normalization ensures all features contribute equally to the model training process.

Step 3 – Feature Selection: Correlation analysis is performed to identify the most influential features affecting salary. Features with high correlation to the target variable (salary) are retained, while redundant or weakly correlated features are excluded. Experience, designation, and education show the highest correlation coefficients.

Step 4 – Train/Test Split: The preprocessed dataset is divided into training (80%) and testing (20%) subsets using stratified random sampling. The training set is used for model learning, while the test set is reserved for unbiased evaluation of model performance. This ensures the model's generalization capability is accurately assessed.

Step 5 – Model Training: Two regression models are trained on the training data: (a) Linear Regression, which fits a linear equation to minimize the sum of squared residuals, and (b) Random Forest Regressor with 100 decision trees, which uses bootstrap aggregating (bagging) to improve prediction accuracy and reduce overfitting.

Step 6 – Model Evaluation: Both trained models are evaluated on the test set using four standard regression metrics: R^2 Score (coefficient of determination), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and prediction accuracy percentage. These metrics provide a comprehensive assessment of model performance.

Step 7 – Comparison & Selection: The evaluation metrics of both models are compared side by side. Random Forest Regressor demonstrates superior performance with an R^2 of 0.9681 (96.81% accuracy) compared to Linear Regression's R^2 of 0.9247 (92.47% accuracy). Random Forest is selected as the best model due to its higher accuracy and lower error rates.

Step 8 – Salary Prediction: The selected Random Forest model is deployed in the web-based prediction module. Users can input new feature values (experience, education, publications, designation, specialization, and state) through an interactive interface, and the system outputs the predicted monthly salary in INR along with the confidence level of the prediction.

The following Python code snippet demonstrates the implementation of both algorithms using scikit-learn:

```
from sklearn.linear_model import
LinearRegression from sklearn.ensemble import
RandomForestRegressor from
sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error

X_train, X_test, y_train, y_test =
   train_test_split( X, y, test_size=0.2,
   random_state=42)

# Linear Regression
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)
lr_pred =
lr_model.predict(X_test)

# Random Forest Regressor
rf_model =
RandomForestRegressor(n_estimators=100)
rf_model.fit(X_train, y_train)
rf_pred = rf_model.predict(X_test)
```

5. Results and Discussion

The performance of both regression models was evaluated on the test dataset using standard evaluation metrics. Table 2 presents a detailed comparison of the results obtained from Linear Regression and Random Forest Regressor.

Metric	Linear Regression	Random Forest
R^2 Score	0.9247	0.9681
Accuracy (%)	92.47%	96.81%
MAE	₹5,832	₹3,421
RMSE	₹6,949	₹4,634
Train Score	0.9312	0.9893
Test Score	0.9247	0.9681
Overfitting Risk	Low	Medium
Training Speed	Fast	Moderate

Table 2: Performance Comparison of Linear Regression and Random Forest

The results clearly demonstrate that Random Forest Regressor outperforms Linear Regression across all primary evaluation metrics. Random Forest achieves an R^2 Score of 0.9681, indicating that the model explains approximately 96.81% of the variance in salary data, compared to 92.47% for Linear Regression.

The Mean Absolute Error (MAE) of Random Forest (₹3,421) is significantly lower than that of Linear Regression (₹5,832), indicating more precise predictions. Similarly, the RMSE values confirm that Random Forest produces smaller prediction errors (₹4,634 vs ₹6,949).

The superior performance of Random Forest can be attributed to its ability to capture non-linear relationships and complex feature interactions through its ensemble of decision trees. While Linear Regression assumes a strictly linear relationship between features and salary, the actual salary structure involves non-linear interactions between experience, education, and designation levels.

However, Linear Regression offers advantages in terms of interpretability and computational efficiency. Its training speed is faster, and the risk of overfitting is lower compared to Random Forest. For applications requiring model explainability, Linear Regression remains a valuable alternative.

The train-test score comparison reveals that Random Forest has a marginally higher overfitting risk (train: 0.9893, test: 0.9681) compared to Linear Regression (train: 0.9312, test: 0.9247). However, the gap is minimal, indicating that the Random Forest model generalizes well to unseen data.

6. Conclusion

This study presented a machine learning-based approach for predicting the salary of government teaching professionals in India. Two regression algorithms — Linear Regression and Random Forest Regressor — were compared using a dataset of 200 records with features including experience, education, publications, designation, specialization, and state of employment.

The experimental results demonstrate that Random Forest Regressor significantly outperforms Linear Regression with an accuracy of 96.81% (R^2 = 0.9681) compared to 92.47% (R^2 = 0.9247). The Random Forest model also achieves lower MAE (₹3,421 vs ₹5,832) and RMSE (₹4,634 vs ₹6,949), confirming its superiority in capturing the complex, non-linear relationships inherent in salary determination factors.

The proposed system provides a reliable and data-driven framework for salary prediction that can assist educational administrators and policy makers in ensuring equitable and transparent compensation structures. The interactive web-based interface enables users to input individual parameters and obtain instant salary predictions with both algorithms.

Future work may include: (1) expanding the dataset with real government salary records, (2) incorporating additional features such as research grants and administrative responsibilities, (3) exploring deep learning approaches for further accuracy improvement, and (4) integrating the system with government HR management platforms for real-time salary recommendations.

References

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[2] Das, S., Sharma, R., & Kumar, P. (2020). Salary Prediction Using Machine Learning Techniques. International Journal of Data Science, 8(3), 145–158.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer.

[4] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.

[5] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th ed.). Wiley.

[6] Nguyen, T., & Armitage, G. (2018). Predicting Public Sector Salaries Using Ensemble Methods. Journal of Applied Machine Learning, 5(2), 89–103.

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[8] Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

[9] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[10] Bogen, D., & Rieke, A. (2018). Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias. Upturn.

[11] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

[12] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.