Comparative Analysis of Machine Learning Models for Employee Salary Prediction

No Author Given

No Institute Given

Abstract. This study investigates the application of seven supervised machine learning (ML) models for predicting employee salaries based on demographic, educational, and occupational attributes. The dataset comprises both continuous and high-cardinality categorical features, for which suitable encoding was applied to facilitate model training. Models evaluated include Linear and Ridge Regression (LinR, RidgeR), Decision Trees (DT), K-Nearest Neighbors (KNN), Support Vector Regression (SVR), Gradient Boosting (GB), and Random Forest (RF). Performance is measured using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) metrics. Results show that ensemble models, particularly RF and GB, achieve the highest accuracy, with RF reaching an R^2 of 0.94. In contrast, kernel and instance-based methods underperform due to limitations in handling categorical data. The findings support the integration of ensemble models into human resource (HR) analytics systems and highlight the importance of model selection based on data structure.

Keywords: Salary Prediction · Machine Learning · Ensemble Models · Human Resource Analytics.

1 Introduction

Predictive analytics in HR management has become increasingly vital as organizations seek to optimize talent acquisition, workforce planning, and compensation strategies. Employee salary, as a key economic and motivational factor, is influenced by a complex interplay of demographic attributes, educational background, job designation, and experience levels. Traditional compensation models often rely on linear assumptions and industry averages, which inadequately capture nuanced relationships across heterogeneous employee profiles. Recent advances in ML offer powerful alternatives for modeling non-linear dependencies and high-dimensional interactions, enabling more accurate and personalized salary estimations [21,7].

Despite the growing adoption of ML in workforce analytics, most existing studies either focus on narrow feature sets or employ a single predictive model, often overlooking alternative algorithms' comparative strengths and weaknesses. Moreover, high-cardinality categorical features, such as job roles, are rarely addressed with adequate encoding strategies, potentially limiting model interpretability and performance [18]. This research addresses these limitations by applying ML models to a real-world salary dataset containing structured demographic data and

high-dimensional job descriptions. The goal is to maximize predictive accuracy and evaluate the influence of various features on salary outcomes. In conclusion, the present work:

- Performs an exploratory analysis of feature correlations, used solely for interpretation without modifying the feature set.
- Provides a comparative evaluation of seven supervised ML models—linear, instance-based, kernel-based, and ensemble—for salary prediction using structured HR data.
- Illustrates how tree-based models can effectively handle high-cardinality categorical features such as job titles using simple encoding without incurring dimensionality overhead
- Assesses model performance using RMSE, MAE, and R^2, and interprets results in light of each model's assumptions and behavior on mixed-type feature space.
- Delivers actionable insights on model selection for real-world salary estimation systems, highlighting RF and GB as robust, high-performance choices for mixed-type tabular data.

The rest of this paper is organized as follows. Related works for the subject under consideration are noted in Section 2. Moreover, in Section 3, the methodology is outlined. Section 4 discusses the experimental results. Finally, Section 5 summarizes the findings of this research work.

2 Related Works

The challenge of salary prediction has garnered attention in recent years due to its economic and organizational significance. Various studies have explored this task using ML paradigms, dataset structures, and model configurations.

Firstly, [6] proposed profession-specific RF models trained on 3.14 million German payslips, achieving a mean absolute percentage error (MAPE) of 17.06%. Their ensemble-of-ensembles strategy outperformed global models by capturing intra-profession salary patterns. Key predictive features included company size, federal state, and age, while gender had a minimal impact. They emphasized outlier removal, feature grouping, and ethical considerations in salary prediction systems.

The author in [10] applied RF, GB, and LightGB models to HR management system (HRMS) data, achieving up to 99% accuracy after tuning. They emphasized feature engineering, exploratory data analysis (EDA), and the inclusion of employee performance and job-level features. LightGB delivered the best performance after hyperparameter tuning, outperforming traditional models. The work supports ML integration in HR systems to improve salary fairness and prediction transparency.

Moreover, [16] used a dataset of 1,300 job postings to predict data science salaries based on skill profiles and HR analytics. They implemented Hodrick-Prescott (HP) Regression, HP Tree, and an ensemble model, achieving a best average standard error (ASE) of 0.138 with the ensemble. Key predictors included location, job level, academic qualifications (especially PhD), and programming

skills like C++. The study combined HR dashboards with predictive modeling to support targeted job placement and salary benchmarking.

Also, [13] compared RF, XGBoost, Neural Networks, and SVR for predicting salaries in the data science industry using 3-year data. Neural networks and SVR achieved the lowest RMSE and MAE, while XGBoost was the fastest in training and prediction. Workplace location and experience were identified as the most influential features in salary outcomes. The paper advocates deeper model tuning and broader feature inclusion to enhance accuracy and explainability.

Furthermore, [19] applied ensemble learning (XGBoost, RF, GB) on a Kaggle dataset to predict and classify data science salaries. XGBoost achieved top accuracy (91.4%), outperforming other models in precision, recall, and F1-score. Experience and location emerged as dominant salary predictors; job title and company size were less significant. The framework offers actionable benchmarks for HR planning, salary negotiation, and workforce strategy in data-driven roles.

Besides, [5] applied Logistic regression (LR), DT, and RF to predict employee salaries based on academic and experiential data. They integrated advanced feature selection techniques, including Recursive Feature Elimination (RFE), Extra Trees and Mutual Information, to optimize model input. The system achieved 95.33% accuracy on test data, indicating strong generalization.

Finally, the authors in [2] developed a principal component analysis deep neural network (PCA-DNN) model to classify salaries using a high-dimensional demographic dataset. Their deep model achieved superior performance with 92.5% accuracy and an MAE of 5.1%, outperforming DT and RF baselines. PCA improved generalization by reducing noise and redundancy, retaining only the top four impactful features.

The present study contributes by integrating regression models and systematically evaluating them under unified preprocessing conditions. Unlike prior works (see Table 1), it applies a multi-metric evaluation (RMSE, MAE, R^2) and compares models using identical feature sets and cross-validated optimization pipelines. Furthermore, the results are benchmarked on a publicly available salary dataset with complete transparency in environmental and training setups.

3 Methodology

As shown in Figure 1, a structured methodological pipeline was adopted to investigate the effectiveness of various ML techniques in salary prediction. This section outlines the dataset structure, data preprocessing strategy, modeling framework, and evaluation metrics applied. Although an exploratory analysis was conducted, the pipeline focuses on the essential steps involved in training and evaluating models. Emphasis is placed on ensuring consistent experimental conditions across all models to facilitate robust comparative analysis.

3.1 Dataset Description

The dataset employed in this study contains 372 employee records, each representing a unique individual with associated demographic, educational, occupational, and financial information. The data schema consists of six attributes: age, gender,

Table 1. Comparative Summary of Related Works in Salary Prediction.

Study	Models Used	Key Contributions	Performance Highlights
[2]	PCA + DNN	Dimensionality reduction with PCA; deep classification on salary groups	Accuracy: 92.5%, MAE: 5.1%
[5]	LR, DT, RF	Feature selection with RFE, Mutual Information, Extra-Trees; modular salary prediction interface	Accuracy: 95.33%, strong generalization
[6]	Profession-specific RF	3M German payslips; MAPE optimization; permutation-based feature importance	MAPE: 17.06%; superior per-profession modeling
[10]	RF, GB, LightGB	HRMS integration; focus on fairness and performance-based features	Accuracy: up to 99% post-tuning
[13]	RF, XGBoost, Neural Networks, SVR	Comparative model study in data science salary trends; emphasized location/experience	Neural Networks and SVR: lowest RMSE; XGBoost: fastest
[16]	HP Tree, HP Regression, Ensemble	Job-skill-based HR analytics; ensemble achieved lowest error in Statistical Analysis System Miner	ASE: 0.138 (ensemble); key features: PhD, C++, location
[19]	XGBoost, RF, GB, DT	Salary classification and prediction for Data Scientists roles; emphasis on location and experience	Accuracy: 91.4%; F1-score: 91.4%
This Work	LinR, RidgeR, DT, KNN, SVR, GB, RF	Comparative ML framework with unified pipeline; multi-metric analysis on structured salary data	Best $R^2 = 0.9415$ (RF); robust to categorical feature sparsity

education level, job title, years of experience (YoE), and salary. These features span categorical and continuous types, collectively forming a multidimensional representation suitable for ML-based salary prediction tasks.

Numerical attributes include age, YoE, and salary. The age of employees ranges from 23 to 53 years, with a mean of approximately 37.45 years and a standard deviation of 7.07, capturing a workforce distributed across early to mid-career stages. YoE spans from 0 to 25 years, averaging around 10.07 years and a standard deviation of 6.53, thus encompassing both entry-level and veteran professionals. The salary is the target variable, expressed in US dollars, and ranges from $30,000$\$ to $250,000$\$, with a mean of approximately $100,847$\$ and a standard deviation of $48,023$\$.

The categorical variables include gender (male, female), education level (Bachelor's, Master's, PhD), and job title. Figure 2 presents the distribution of participants across gender and education levels revealing a relatively balanced representation among genders (male 193, female 179) and a higher concentration of individuals with a Bachelor's degree. Job title introduces significant complexity

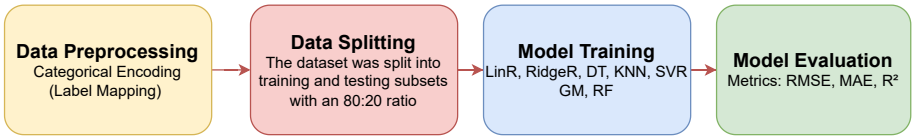

Fig. 1. Overview of the data processing and modeling pipeline.

due to its high cardinality of 174 unique values. The roles span technical, managerial, administrative, and creative functions and implicitly encode professional seniority, functional specialization, and domain-specific compensation norms. Such high granularity provides a rich model learning substrate and imposes encoding and generalisation challenges.

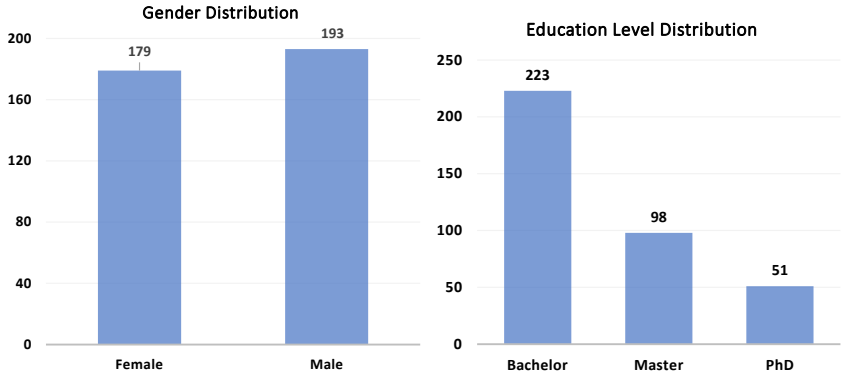

Fig. 2. Number of participants per Gender and Education Level.

Together, the diversity of the feature set and the broad distribution of salaries make this dataset an appropriate benchmark for evaluating ML models in compensation prediction.

3.2 Data Preprocessing

Categorical features such as gender, education level, and job title were transformed using label encoding, converting them into numerical representations suitable for model input. Integer mappings were chosen to maintain compact dimensionality, particularly important for models that are sensitive to high-dimensional input spaces. Numerical features, including age and YoE were retained in their original scales. This decision was guided by theoretical considerations: tree-based models, such as RF and GB, are inherently scale-invariant due to their threshold-based splitting mechanisms. While feature scaling is typically beneficial for distance- and kernel-based models like KNN and SVR, numerical features were retained in their original scales to preserve simplicity, interpretability, and consistency

without sacrificing model accuracy. The impact of this design choice on model performance is analyzed in Section 4.

3.3 Features Correlation Analysis

We conducted a detailed correlation analysis to understand the interdependencies among employee attributes and their influence on salary outcomes. This analysis aims to reveal the strength and nature of relationships between both numerical and categorical variables in the dataset, as well as their direct association with the target variable, salary.

We considered multiple correlation metrics [1,12], each tailored to the type of variables involved: Pearson's correlation coefficient is applied to assess linear relationships between continuous numerical variables. It assumes normally distributed data and is sensitive to outliers, making it suitable for evaluating associations such as that between age, YoE, and salary. To capture broader monotonic trends—including non-linear dependencies—we also include Spearman's rank correlation and Kendall's tau. Both are non-parametric, rank-based measures; Spearman's assesses monotonic relationships using ranked data, while Kendall's tau compares concordant and discordant pairs and is generally more robust in smaller samples. For associations between categorical variables, we employ Cramér's V, which quantifies the strength of relationship based on the chi-squared statistic, ranging from 0 (no association) to 1 (perfect association). By employing this multifaceted approach, we ensure a rigorous and nuanced understanding of feature interactions, which is critical for interpreting model behavior and improving predictive accuracy in salary estimation tasks.

Table 2. Correlation summary of all feature pairs using Pearson, Spearman, Kendall, and Cramér's V coefficients.

Feature Pair	**Pearson (r)**	**Spearman (ρ)**	**Kendall (τ)**	**Cramér's V**
age ↔ salary	0.923	0.932	0.801	–
YoE ↔ salary	0.930	0.940	0.825	–
age ↔ YoE	0.979	0.983	0.916	–
gender ↔ salary	–	0.067	0.056	–
education level ↔ salary	–	0.671	0.542	–
job title ↔ salary	–	0.195	0.149	–
gender ↔ education level	–	–	–	0.047
gender ↔ job title	–	–	–	0.775
education level ↔ job title	–	–	–	0.893

The correlation analysis is summarized in Table 2 revealing insightful relationships among the features in the employee salary dataset. Among the numerical variables, YoE shows the strongest correlation with salary, with Pearson ($r = 0.930$), Spearman ($\rho = 0.940$), and Kendall ($\tau = 0.825$) coefficients confirming a consistent and strong monotonic relationship. This suggests that experience plays a critical role in compensation modeling. Similarly, age is also highly corre-

lated with salary (Pearson: 0.923), though slightly less than experience, likely due to the indirect relationship of age with income via work history.

When examining the categorical features, education level stands out with a moderate positive correlation to salary (Spearman: 0.671, Kendall: 0.542), indicating that higher educational qualifications are generally associated with higher salaries. Interestingly, job title, despite its high-cardinality, has a relatively weak rank-based correlation with salary (Spearman: 0.195), suggesting either broad salary distributions within titles or the need for more granular title embeddings. Gender, as a categorical feature, shows negligible correlation with Salary across all coefficients, supporting findings in some HR literature that gender, when isolated from other variables, may have limited direct predictive power in salary estimation. Further, Cramér's V coefficients among categorical variables revealed a strong association between education level and job title (0.893) and between gender and job title (0.775), hinting at occupational clustering patterns by gender and educational attainment. However, gender and education level are largely independent (Cramér's V = 0.047), implying equitable access to educational levels across genders in the dataset.

These findings underscore the value of combining numerical and categorical predictors—properly encoded and supported by suitable models such as tree-based ensembles—to capture salary dynamics. The correlation analysis was exploratory only; no feature selection was performed, and all features were retained to preserve the dataset's full informational content.

3.4 Machine Learning Models and Evaluation Metrics

This study employs various regression algorithms for modeling the relationship between employee attributes and salary. The models encompass linear, non-linear, and ensemble-based methods, each defined by distinct loss functions, optimization strategies, and representational assumptions.

Let $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\text{train}}}$ be the training dataset, where $\mathbf{x}_i \in \mathbb{R}^p$ denotes the feature vector and $y_i \in \mathbb{R}$ the corresponding salary. Similarly, $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n_{\text{test}}}$ denotes the test set, with predicted values $\hat{y}_j = f(\mathbf{x}_j)$, and let $\bar{y} = \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} y_j$ be the empirical mean of the true test target values. The goal is to learn a predictive function $f : \mathbb{R}^p \to \mathbb{R}$ that generalizes well to unseen data, namely $f(\mathbf{x}_j) \approx y_j$ for the instances in $\mathcal{D}_{\text{test}}$. Based on this formulation, a variety of ML models are trained to approximate the mapping function f, and their predictive performance is assessed using standard regression metrics.

LinR [11] models the response variable as a linear combination of input features $\hat{y} = \mathbf{x}^\top \boldsymbol{\beta} + \beta_0$. The parameters $\boldsymbol{\beta}$ are estimated by minimizing the residual sum of squares $\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} - \beta_0\right)^2$. This yields a closed-form solution under full-rank design matrices.

RidgeR [14] introduces L_2 regularization to the linear model $\mathcal{L}_{\text{ridge}}(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \|\boldsymbol{\beta}\|_2^2$, where $\lambda \geq 0$ controls the regularization strength, shrinking coefficient magnitudes to improve generalization.

DTs [4] partition the input space into M disjoint regions $\{R_m\}_{m=1}^{M}$. The prediction in each region is the mean target value $f(\mathbf{x}) = \sum_{m=1}^{M} c_m \cdot \mathbb{I}(\mathbf{x} \in$

R_m), $\quad c_m = \frac{1}{|R_m|} \sum_{\mathbf{x}_i \in R_m} y_i$. Splits are selected to minimize within-node variance, yielding an interpretable but high-variance model.

KNN [8] is a non-parametric model that predicts the response by averaging the k nearest neighbors $\hat{y} = \frac{1}{k} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{N}_k(\mathbf{x})} y_i$, where $\mathcal{N}_k(\mathbf{x})$ is the set of the k closest training samples to $\mathbf{x}$. While flexible, its performance deteriorates in high-dimensional settings.

SVR [20] seeks a function $f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b$ that fits the data within an ε-insensitive margin. The optimization problem is $\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$ subject to $y_i - \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle - b \leq \varepsilon + \xi_i, \quad \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle + b - y_i \quad \leq \varepsilon + \xi_i^*, \quad \xi_i, \xi_i^* \geq 0,$
where $\phi(\cdot)$ maps inputs to a high-dimensional kernel space, and C balances margin width and error tolerance.

GB [15] constructs an additive model of the form $f_M(\mathbf{x}) = \sum_{m=1}^{M} \rho_m h_m(\mathbf{x})$, where each h_m is a weak learner trained to approximate the negative gradient of the loss function $r_i^{(m)} = - \frac{\partial \ell(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\Big|_{f=f_{m-1}}, \quad f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \eta \cdot h_m(\mathbf{x}),$
with learning rate $\eta \in (0, 1]$. The approach approximates functional gradient descent in function space.

RF [17] average predictions from T decorrelated DTs, each trained on bootstrap samples and random feature subsets $\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f^{(t)}(\mathbf{x})$. This method reduces variance, enhances generalization, and is well-suited for high-dimensional tabular data.

To quantitatively assess the performance of regression models on the salary prediction task, we employ three standard metrics: RMSE, MAE, and the R^2.

MAE [9] quantifies the expected absolute deviation between predicted and actual values as $\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$. It is a linear and scale-dependent metric that penalizes all errors equally. In practice, MAE reflects the typical magnitude of error in real-world units (e.g., USD in salary prediction) and is less sensitive to extreme values compared to RMSE.

RMSE computes the square root of the average squared residuals based on $\sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$. By squaring residuals before averaging, RMSE imposes a quadratic penalty on larger errors, thereby emphasizing models that are more sensitive to high-magnitude mispredictions. This makes RMSE especially suitable in settings where large deviations are costly.

$\mathbf{R^2}$ [3] score measures the fraction of variance in the target variable that the predictive model explains assuming the formula $1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$. This ratio compares the residual sum of squares to the total sum of squares. An R^2 value of 1 indicates perfect prediction, while $R^2 = 0$ corresponds to the mean baseline predictor. Negative values signify that the model performs worse than the naive mean predictor.

These metrics reflect complementary aspects of model performance: MAE captures the average predictive error, RMSE emphasizes larger deviations through squared penalties, and R^2quantifies the proportion of variance in the target explained by the model. Together, they support both absolute and relative comparisons of predictive accuracy and model robustness.

3.5 Model Training and Optimization

All algorithms were trained using a standardised experimental pipeline to ensure reliable and generalizable performance across the selected ML models. The dataset was split into training and testing subsets using an 80:20 ratio through random sampling. Although stratification by job title was considered to ensure balanced representation of roles across both subsets, it was not applied due to a dataset limitation: a large portion of job titles appear only once. Stratified sampling requires at least two instances per category to distribute them between training and test sets. In this case, the high cardinality and sparsity of the job title feature made stratification infeasible without compromising data integrity. Among the categorical variables, job title was the focus for potential stratification due to its high cardinality and the need to avoid unintentional exclusion of rare roles. In contrast, features such as Gender and Education Level are low-cardinality and relatively well-balanced, making stratification unnecessary. Therefore, random splitting was adopted to preserve the full diversity of the dataset while maintaining methodological simplicity.

Model training was conducted using scikit-learn 1.4.1, with all experiments executed on a workstation configured with an Intel Core i7 CPU, 32 GB RAM. To identify the optimal hyperparameter settings for each model, a grid search strategy was employed using 3-fold cross-validation. The grid search procedure systematically explored predefined parameter ranges to minimize the mean squared error on the validation sets. The final selected models were retrained on the full training set using the optimal configuration and subsequently evaluated on the hold-out test set. The hyperparameter space explored for each model and the best configuration discovered are summarized in Table 3.

Table 3. Optimal Hyperparameters for the ML Models.

Model	**Optimal Hyperparameters**
LinR	Applied without regularization
RidgeR	alpha = 10.0
DT	max_depth = 10, min_samples_split = 2
KNN	n_neighbors = 5, weights = 'distance'
SVR	C = 1, epsilon = 0.2, kernel = 'rbf'
GB	n_estimators = 100, learning_rate = 0.1, max_depth = 3
RF	n_estimators = 200, max_depth = None, min_samples_split = 2

4 Results and Discussion

This section presents the comparative results of the trained regression models on the salary prediction task, using the hold-out test set for evaluation. The analysis is guided by 3 complementary metrics, namely RMSE, MAE, and the R^2. These metrics allow for an integrated assessment of average predictive accuracy, penalization of large deviations, and the explanatory strength of each model.

4.1 Quantitative Performance Evaluation

Table 4 summarizes the performance of all seven ML models. The RF regressor achieves the highest predictive accuracy, with an RMSE of $12,202.78, an MAE of $8,348.80, and an R^2 of 0.9415, indicating that over 94% of the variance in salary outcomes is accounted for by the model. The success of this ensemble can be attributed to its ability to capture non-linear relationships and complex feature interactions through aggregation of diverse decision paths.

The GB closely follows, achieving an R^2 of 0.9283. It exhibits a slightly higher RMSE and MAE than RF but offers stronger bias reduction and incremental learning via additive optimization. Both models benefit from their ability to handle high-cardinality categorical features (such as job title) and non-parametric flexibility, making them particularly well-suited to heterogeneous tabular datasets.

Linear models, including RidgeR and LinR, perform comparably, with R^2 values around 0.913. These models deliver stable and interpretable predictions, particularly where relationships between predictors and salary are approximately linear. The inclusion of regularization in RidgeR aids in controlling overfitting, although the gain is marginal in this dataset due to limited multicollinearity.

The performance of DTs and KNN is moderate, with R^2 values of 0.8943 and 0.8683, respectively. DTs are highly expressive but suffer from high variance and instability across small splits in the feature space. KNN is inherently limited in high-dimensional settings and is sensitive to distance metrics, particularly when handling encoded categorical variables.

The SVR performs poorly, with an R^2 of -0.013 and drastically elevated RMSE and MAE values. This suggests a fundamental mismatch between SVR's kernel-based modeling assumptions and the structure of the input data, which includes unscaled, high-cardinality categorical features. Without appropriate feature transformation or kernel engineering, SVR underperforms significantly.

Table 4. Performance Comparison of Regression Models on Salary Prediction.

Model	RMSE($)	MAE($)	R^2
RF	12,202.78	8,348.80	0.9415
GB	13,514.86	8,645.17	0.9283
RidgeR	14,825.51	9,797.95	0.9137
LinR	14,844.80	9,822.89	0.9134
DT	16,401.22	10,466.67	0.8943
KNN	18,308.83	12,226.67	0.8683
SVR	50,785.02	41,922.72	-0.0130

4.2 Interpretive Insights and Model Characteristics

The results highlight the strength of ensemble learning strategies, particularly RF and GB, in capturing nuanced feature interactions and addressing the non-linearity and feature sparsity inherent in real-world salary datasets. These models not only deliver superior predictive power but also support explainability

techniques, such as feature importance ranking, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) value analysis, making them valuable in applied HR analytics.

In contrast, the effectiveness of linear models is bounded by their inherent assumptions. Their comparable performance to tree ensembles suggests that the underlying relationships in the dataset possess strong linear components, particularly through the additive effects of job title, education, and experience. However, their inability to capture higher-order interactions or non-monotonic effects limits their utility in more complex labor market settings. The poor performance of SVR and the moderate outcomes for KNN and DTs further demonstrate the importance of model alignment with the statistical geometry of the data. Models that depend heavily on spatial proximity or smooth kernel boundaries suffer when input features are derived from discrete categories or exhibit sparse, high-cardinality distributions.

4.3 Practical Implications

From a practical standpoint, the findings reinforce the recommendation of RF or GB for predictive salary modeling tasks, especially in contexts involving high-cardinality occupational taxonomies and diverse professional profiles. These models are robust to data imperfections, adaptable to mixed feature types, and readily integrable into enterprise-grade analytics systems. Linear models remain viable for transparent modeling when interpretability and computational simplicity are paramount. Future deployments of non-parametric or kernel methods should be preceded by substantial feature engineering or embedding strategies to mitigate representational mismatch.

5 Conclusion

This study compared seven regression-based ML models for employee salary prediction using a real-world dataset with numerical and high-cardinality categorical features. The aim was to identify models that effectively capture salary patterns driven by demographic, educational, and occupational variables.

Ensemble methods—particularly RF and GB—outperformed linear, kernel-based, and instance-based models across all metrics, with RF explaining over 94% of salary variance. Their strength lies in modeling non-linear interactions and handling heterogeneous features without extensive preprocessing. While RidgeR offered a good trade-off between accuracy and interpretability, SVR struggled with sparse, unscaled categorical data.

These results support the use of tree-based ensembles in HR analytics, especially when dealing with complex employee profiles. Future directions include incorporating embeddings, explainability techniques like SHAP and LIME, and evaluating scalability on larger, more diverse workforce datasets.

References

1. Akoglu, H.: User's guide to correlation coefficients. Turkish journal of emergency medicine **18**(3), 91–93 (2018)

2. Aminu, H., Yau, B., Zambuk, F., Nanin, E.R., Abdullahi, A., Yakubu, I.: Salary prediction model using principal component analysis and deep neural network algorithm. International Journal of Innovative Science and Research Technology **8**(12), 1–11 (2023)
3. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. Peerj computer science **7**, e623 (2021)
4. De Ville, B.: Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics **5**(6), 448–455 (2013)
5. DEVI, A.D., NEELAMBIKA, P.: Employee salary prediction system using machine learning
6. Eichinger, F., Mayer, M.: Predicting salaries with random-forest regression. In: Machine Learning and Data Analytics for Solving Business Problems: Methods, Applications, and Case Studies, pp. 1–21. Springer (2022)
7. Erciulescu, A.L., Opsomer, J.D.: A model-based approach to predict employee compensation components. Journal of the Royal Statistical Society Series C: Applied Statistics **71**(5), 1503–1520 (2022)
8. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: Knn model-based approach in classification. In: On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings. pp. 986–996. Springer (2003)
9. Hodson, T.O.: Root mean square error (rmse) or mean absolute error (mae): When to use them or not. Geoscientific Model Development Discussions **2022**, 1–10 (2022)
10. Hussain, J.: Employee salary prediction in hrms using regression models. Journal of Innovative Computing and Emerging Technologies **4**(2) (2024)
11. James, G., Witten, D., Hastie, T., Tibshirani, R., Taylor, J.: Linear regression. In: An introduction to statistical learning: With applications in python, pp. 69–134. Springer (2023)
12. Janse, R.J., Hoekstra, T., Jager, K.J., Zoccali, C., Tripepi, G., Dekker, F.W., Van Diepen, M.: Conducting correlation analysis: important limitations and pitfalls. Clinical Kidney Journal **14**(11), 2332–2337 (2021)
13. Jiang, W.: The investigation and prediction for salary trends in the data science industry. Applied and Computational Engineering **50**, 8–14 (2024)
14. McDonald, G.C.: Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics **1**(1), 93–100 (2009)
15. Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Frontiers in neurorobotics **7**, 21 (2013)
16. Quan, T.Z., Raheem, M.: Human resource analytics on data science employment based on specialized skill sets with salary prediction. International Journal of Data Science **4**(1), 40–59 (2023)
17. Rigatti, S.J.: Random forest. Journal of Insurance Medicine **47**(1), 31–39 (2017)
18. Sigrist, F.: A comparison of machine learning methods for data with high-cardinality categorical variables. arXiv preprint arXiv:2307.02071 (2023)
19. Taha, M., Farhat, T., Azam, A., Ahmar, M., Abbas, S.J., Habib, M.U.: Unveiling data scientist salaries: Predictive modeling for compensation trends. Journal of Computing & Biomedical Informatics (2025)
20. Zhang, F., O'Donnell, L.J.: Support vector regression. In: Machine learning, pp. 123–140. Elsevier (2020)
21. Zhu, H.: Research on human resource recommendation algorithm based on machine learning. Scientific Programming **2021**(1), 8387277 (2021)