2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON) | 979-8-3503-1824-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ECTIDAMTNCON60518.2024.10480058

A Novel Salary Prediction System Using Machine Learning Techniques

Abu Asaduzzaman, Md Raihan Uddin
Dept of Electrical and Computer Engineering
Wichita State University
Wichita, Kansas, USA
abu.asaduzzaman@wichita.edu

Yoel Woldeyes
School of Computing
Wichita State University
Wichita, Kansas, USA
yswoldeyes@shockers.wichita.edu

Fadi N. Sibai
Dept of Electrical and Computer Engineering
Gulf University for Science and Technology
Hawally, Kuwait
sibai.f@gust.edu.kw

***Abstract*—The purpose of this work is to build a salary prediction system using machine learning techniques. The experiments are done using the data from 1994 census database which has 32,561 records of employee data. The techniques used in determining whether an employee salary is less than or greater than $50,000 are: logistic regression, decision tree, Naive Bayes classifier, K-nearest neighbor, and support vector machine. We implement these algorithms using original train data and oversampled train data. The results of these models are analyzed and compared with respect to accuracy. According to the experimental results, decision tree model outperforms the other models with original train data.**

Keywords—salary prediction systems, model accuracy, machine learning, decision tree

I. INTRODUCTION

Salary prediction using machine learning is a crucial aspect for employees, employers, and students, as it serves various purposes [1-7]. In an organization, employers can determine the appropriate salary to offer a new employee based on their profile. Likewise, employees can predict their potential salary based on their skills and years of experience in the job market. This prediction can help individuals make informed decisions about their career paths and negotiate compensation packages effectively.

The primary objective of this project is to classify whether a person's salary will be less than or greater than $50,000. Several attributes are considered in predicting the salary, including age, work class, final weight, education, education num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country, and salary [8-10]. By analyzing these attributes, machine learning algorithms can identify patterns and relationships that contribute to salary levels, enabling accurate predictions to be made.

The salary prediction process involves several steps. The first step is data munging, which involves preprocessing the data to ensure its quality and consistency. This includes handling missing values, transforming variables, and preparing the data for further analysis. The second step is exploratory data analysis, where descriptive statistics are used to summarize the data, visualizations are created to gain insights, and correlations between variables are examined. Outliers, if present, are addressed to ensure the accuracy of the predictions.

In the subsequent step, the lasso regression method is employed to perform oversampling. This technique equalizes the number of instances in each salary class, balancing the dataset and mitigating any potential bias towards a particular class. By balancing the data, the machine learning model can effectively learn from both classes and make accurate predictions for each salary category.

Finally, the model is trained using various machine learning algorithms, considering the oversampled and original training data. Different algorithms, such as decision trees, logistic regression, or random forest, may be employed to identify the best-performing model for salary prediction. The model's performance is evaluated using appropriate metrics, such as accuracy, precision, recall, or F1 score, to assess its predictive capabilities.

By harnessing the power of machine learning, salary prediction models have revolutionized the way organizations and individuals approach compensation and career planning [11, 12]. These models leverage advanced algorithms and comprehensive analysis of relevant attributes to provide valuable insights into salary estimation. Machine learning techniques enable organizations to make data-driven decisions regarding fair compensation practices and efficient talent management. One of the key advantages of using machine learning for salary prediction is the ability to consider a wide range of attributes beyond traditional factors such as education and experience. These models can incorporate variables like job title, industry, geographic location, company size, and even individual performance metrics. By analyzing these diverse factors, machine learning models can capture intricate patterns and relationships that contribute to salary determination. Moreover, machine learning models continually learn and improve over time, adapting to changing market dynamics and evolving compensation trends.

The literature review, described in Section II, provides an overview of existing research and studies on salary prediction using machine learning techniques. Section III describes related classification methods. Section IV presents the proposed methodology. Section V discusses the experiments conducted and presents the results obtained, showcasing the performance and accuracy of the developed models. Finally, Section VI concludes the study by summarizing the findings of the research in the context of salary prediction using machine learning.

979-8-3503-1824-1/24/$31.00 ©2024 IEEE

II. LIRERATURE REVIEW

In our work, we aim at determining whether a person's salary is less than or greater than a given amount (e.g., $50,000). It enables accurate predictions of salary, thereby motivating employees and helping students better plan their financial future and evaluate potential job opportunities. Several related papers have been reviewed in this section. This review highlights the effectiveness of different algorithms and methodologies in achieving accurate salary predictions.

Khongchai et al. [4] focuses on the analysis of a salary prediction system using data mining techniques. In this study, they compare the profiles of current students and graduates to provide individual-based pay forecasts. Additionally, they contrast the data mining methods that excel at the task. Therefore, they carry out an experiment employing various graduating student data records through nearly 10-fold cross-validation. According to the data findings, the Multilayer Perceptron algorithm provides the lowest accuracy, approximately 35.08%, while K-NN achieves the highest accuracy at an estimated value of roughly 80.3%. The statistics show that the system is beneficial in enhancing motivation for studying and helping individuals realize their career strategies, based on an evaluation of about 45 to 50 sampled users. Consequently, they discover the simplest system usage technique, resulting in straightforward and thorough prediction outcomes. Identifying the appropriate salary to represent a specific dataset on an x-y graph requires some trial and error. There are usually multiple suitable graph types that can be used, depending on the organization's decisions and how the data is presented. With just a few clicks, one can visualize the data as a bar graph, line graph, or circle graph using one of the many graph types available in modern spreadsheet programs like Excel. However, it's important to note that this forecast is accurate only up to a certain percentage. To improve precision, K-nearest regression can be utilized, and the most accurate prediction can be chosen from the results. One strength of the paper is its comprehensive comparison of data mining methods for salary prediction. By evaluating multiple algorithms, the researchers provide valuable insights into the strengths and weaknesses of each approach. This analysis helps researchers and practitioners in selecting appropriate methods for accurate salary predictions. However, the weakness of [4] is the limited sample size used for evaluation. The evaluation was based on a relatively small number of users, approximately 45 to 50, which may not fully represent the broader population. Expanding the sample size could enhance the generalizability of the findings.

The research by Zhang et al. [7] focuses on salary prediction for job applicants using Random Forest model. They incorporate additional features such as industry, location, and company size to improve predictive accuracy. The study shows that Random Forest model yield promising results and provide valuable insights for both job seekers and employers.

In the study by Chen et al. [8], investigates salary prediction in the information technology (IT) industry using a deep learning model based on a recurrent neural network. Their research demonstrated the effectiveness of deep learning techniques in capturing complex patterns in salary data and achieving accurate predictions. Chen et al. [9] conducts a study using Random Forest Regression, Support Vector Regression, and Decision Tree Regression to predict the salary levels of individuals based on their years of expertise. Their research makes significant contributions in several areas. Firstly, they conducted a comprehensive study to identify the fields that exert the greatest influence on job income and explored the relationships between these fields. This analysis sheds light on the factors that significantly impact salary levels. Secondly, they approach the salary prediction problem as a classification task, which led to improved accuracy by focusing on discrete salary ranges rather than continuous values.

The approaches presented in [8, 9] allow for more precise predictions and better aligns with the practical needs of salary estimation. Thirdly, they compared various classifiers to determine the model that achieves the highest accuracy in predicting salary ranges. By evaluating the annual wage and years of experience of employees, they thoroughly examined different data mining methods for the task. Their experiment, conducted using data from 1000 organizations, demonstrated that Random Forest Regression outperformed other methods in terms of accuracy. This emphasizes the effectiveness of Random Forest Regression as a robust tool for numerical prediction. Not only does it accurately fit data points, but it also enables the forecasting of discrete classes, providing valuable insights into salary estimations. The purpose of the paper is to forecast an employee's income, allowing them to receive the appropriate pay based on their credentials and skill set. Algorithms for Supervised Learning and other Classification techniques/methods can be used to achieve this. In this case, the linear regression approach is employed to forecast salary levels. The fundamental model achieves an accuracy of 96-98%. Additional parameters are then incorporated, and the model's performance is accurately assessed using Mean Squared Error (MSE). Three methods, namely Polynomial Transformation, Ridge Regression, and Random Forest, are utilized to reduce the MSE. By comparing the system's output/results with those of other algorithms using metrics such as the F1 score, accuracy, receiver operating characteristic (ROC) curve, etc., the accuracy of the model is calculated to be 76%.

Several studies have explored the application of Random Forest in salary prediction and related fields. In a study conducted by Li et al. [13], the researchers used Random Forest to predict employee salaries based on various features such as education, experience, and job title. Their results demonstrated that Random Forest outperformed other algorithms in terms of accuracy and prediction performance.

Wang et al. [14] conducts a study on salary prediction for job seekers using ensemble learning methods. The researchers compared the performance of various ensemble models, including random forests, bagging, and boosting, to identify the most effective approach. Their findings revealed that ensemble learning models consistently outperformed individual algorithms in salary prediction tasks. Salary prediction can provide the income range for a specific time period, which is useful when making decisions regarding credit, careers, and human resource management. A prominent technique for salary prediction is the decision tree,

which summarizes the experience of training data. However, the decision tree's ability to forecast salaries is hindered by the high dimension and large variance in data splitting. To address this issue, the Random Forest technique is applied. Prior to building a random forest model to reduce the variance of a single decision tree, the data should be preprocessed by handling missing data, categorical data, and removing irrelevant data.

Yousaf et al. [15], develops a hybrid model combining gradient boosting and support vector regression for salary prediction in the healthcare sector. The study showed that the hybrid model achieved better accuracy compared to individual algorithms and traditional regression methods.

III. Theoretical Background on Classification Techniques

In this section theoretical idea on the working of classification techniques used in this project is explained.

A. Logistic Regression

Logistic Regression is a widely used classification algorithm in machine learning that extends the concept of linear regression to predict binary outcomes. It is particularly useful in the context of salary prediction, where the goal is often to determine whether an individual's salary will fall above or below a certain threshold. In the field of salary prediction, logistic regression can be applied to predict the likelihood of an individual earning a salary above a specific value based on various independent variables such as education, years of experience, job title, and industry. The binary outcome in this case could be represented as 1 if the predicted salary is above the threshold or 0 if it is below. The logistic regression model is trained to converge and predict with 120 iterations. Computational complexity for each iteration is obtained by using Equation (1).

$$C = O(M) + O(N * M) + O(N) + O(N * M) + O(M) \quad (1)$$

Where, N represents the number of data samples and M is the number of features. And, $O(M)$ is the computation for initializing and updating weights and bias, $O(N * M)$ represents the computation for forward pass and backpropagation, and $O(N)$ is the computation for loss function.

One of the key advantages of logistic regression is its interpretability. The model estimates the probability of the binary outcome by applying the logistic function to a linear combination of the independent variables. This allows for a clear understanding of the relationship between the predictors and the likelihood of achieving a specific salary level. Coefficients associated with each independent variable can indicate the direction and strength of their influence on the salary prediction outcome.

B. Naive Bayes Classifier

Naïve Bayes classifier, a popular machine learning algorithm for classification tasks, utilizes the principles of the Bayes Theorem and conditional probability to make predictions. In this algorithm, the presence or absence of one feature does not affect the presence or absence of other features in the dataset. In the context of salary prediction, Naïve Bayes classifier can be leveraged to estimate the probability of an individual falling into a specific salary range based on a set of features such as education level, years of experience, job title, and industry. By calculating the conditional probabilities of each feature given a particular salary range, the classifier can determine the likelihood of an individual belonging to that salary category. The computational complexity for GNB is calculated using the formula in Equation (2).

$$Total\_Comp = O(N * M * K) \quad (2)$$

Where, N and M remain the same and K represents the number of classes.

One of the key advantages of the Naïve Bayes classifier is its simplicity and efficiency. The algorithm requires a relatively small amount of training data to estimate the probabilities accurately, making it computationally efficient. Additionally, Naïve Bayes models can handle both categorical and numerical features, making it applicable to a wide range of salary prediction scenarios.

C. K- Nearest Neighbors

K-NN is a versatile machine learning algorithm that can be utilized for both regression and classification tasks. In the context of salary prediction, K-NN can be employed as a regression algorithm to estimate an individual's salary based on proximity to similar data points. The K-NN algorithm operates on the principle of proximity, assuming that points with similar characteristics tend to be close to each other in the feature space. In the case of salary prediction, K-NN considers the features of individuals such as education level, years of experience, job title, and industry, and identifies the K nearest neighbors in the training dataset based on the similarity of these features. The algorithm then predicts the salary of the target individual by taking in to account the salaries of its nearest neighbors. Total computational complexity for KNN is computed using Equation (3) for each leaf [16-17].

$$Total = O(N * M * K) \quad (3)$$

Where, $N, M, and\ K$ remain the same.

One of the advantages of K-NN in salary prediction is its simplicity and ease of implementation. It does not require any explicit assumption about the underlying data distribution, making it a non-parametric method. Additionally, K-NN can handle both numerical and categorical features, making it applicable to various types of salary prediction datasets.

D. Decision Tree Regression

Decision Tree Regression is a powerful machine learning technique used for salary prediction and other regression tasks. This algorithm creates a model in a tree-like structure, where the data is divided into smaller subsets based on specific features or attributes. Each subset is further split into branches, forming decision nodes and leaf nodes in the tree. The decision nodes in the tree represent the features or attributes that are used to make decisions about how the data should be split. These decisions are based on specific conditions or thresholds related to the features. The leaf nodes, on the other hand, represent the final predicted salary values or outcomes. The simplified computational complexity for decision tree is given below in Equation (4) [18].

$$Total = O(N * M * \log(N)) \quad (4)$$

Where, $N\ and\ M$ remain the same.

One of the key advantages of Decision Tree Regression is its interpretability. The tree structure allows for clear visualization

and understanding of the decision-making process. Decision nodes provide insights into the important features that contribute to the salary prediction, while leaf nodes provide the final predicted salary values.

E. Support Vector Machines

Support Vector Machines (SVM) is a versatile machine learning algorithm that can be used for both regression and classification tasks, including salary prediction. SVM works by creating a hyperplane in an n-dimensional space that effectively separates and categorizes different classes or predicts continuous values. In the context of salary prediction, SVM regression (SVR) can be employed to estimate salary levels based on various input features. SVR aims to find a hyperplane that best fits the data points, considering the maximum number of support vectors. Support vectors are the data points located closest to the hyperplane, which play a crucial role in defining the optimal hyperplane and determining the salary prediction. The computational complexity of SVM is represented by the formula in Equation (5) [19-20].

$$Total = O(N^2 * M) \quad (5)$$

Where, $N\ and\ M$ remain the same.

In the context of salary prediction, SVM regression (SVR) can be employed to estimate salary levels based on various input features. SVR aims to find a hyperplane that best fits the data points, considering the maximum number of support vectors. Support vectors are the data points located closest to the hyperplane, which play a crucial role in defining the optimal hyperplane and determining the salary prediction.

The key advantage of SVM is its ability to handle high-dimensional data efficiently. It can capture complex relationships and patterns in the data by transforming the input features into a higher-dimensional space using kernel functions. This flexibility allows SVM to handle non-linear relationships between input features and salary, providing more accurate predictions.

IV. METHODOLOGY

In this study, we used a total of 9 stages, shown in Figure 1, which include data collection, importing libraries, data wrangling, exploratory data analysis, oversampling method, feature selection, variables, model training, model comparison and analysis.

A. Data Collection

Dataset has been obtained from Kaggle. The Data present in the dataset is tracked back to 1994 Census Database which was retrieved by Barry Becker.

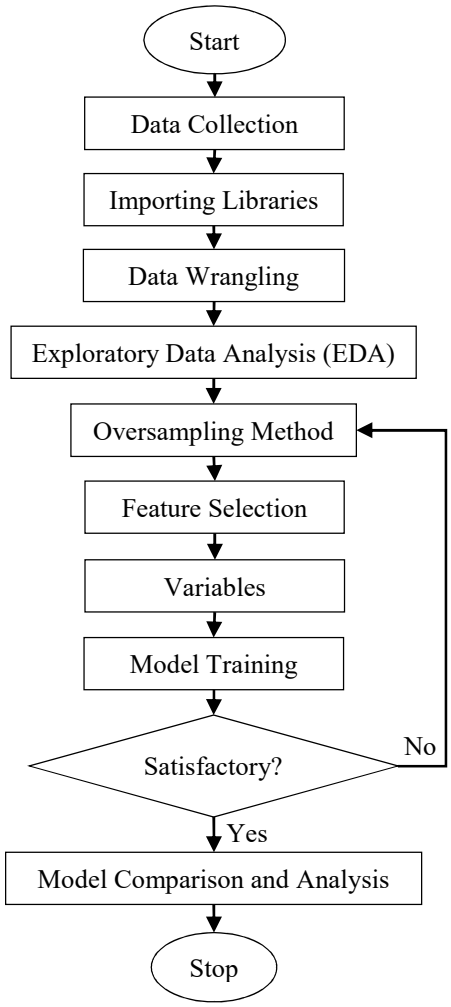

Fig. 1: Flowchart of the methodology.

B. Importing Libraries and Loading Dataset

We imported the necessary libraries like NumPy, Pandas, Matplotlib, Seaborn and Seikit-learn to perform the next steps. For loading the dataset, we use Pandas library.

C. Data Wrangling

Data Wrangling is also called as Data Munging. In this process, the raw data in the dataset is transformed so that it is accessible. In this step, the columns are renamed.

D. Exploratory Data Analysis (EDA)

EDA is used to analyze the dataset, look for patterns and outline the main characteristics by using visualization methods. We created dummy variables for the columns work class, education, marital status, occupation, relationship, race, and sex.

E. Oversampling Method

The data is unbalanced in the salary column. By using oversampling method, the data is balanced. In this way, the model can be trained without any bias.

F. Feature Selection

By using Lasso Method, the best four features (i.e., age, capital gain, capital loss, and hours per week for the model) are obtained out of all 15 features.

G. Dependent and Independent Variables

The outcomes obtained after the feature selection are taken as independent variables. Salary column is taken as dependent variable.

H. Model Training

For this work, 10 different models are created, and these models are trained with original train data and oversampled train data.

I. Model Comparison and Analysis

In order to compare the models, Accuracy, False Negative, and False Positive values are compared for the 10 models implemented.

V. Experiment and Results

In this section, we divide the content into three parts. The first part is preparing the data, the second part is applying the models and the third part is comparing the results obtained after using different algorithms.

A. Data Preparation

- In this step, we import the necessary libraries like NumPy, pandas, SciPy, sklearn, etc., and load the dataset. In this dataset, we have 32561 rows and 15 columns.
- Then we rename columns, remove any noise present in the data, replace data values in the Salary column with 0 and 1, fill categorical Null values with mode imputation. (<=50K value to '0' and >50K value to '1') and fill the categorical Null values with mode imputation.
- Next, we perform Exploratory Data Analysis (EDA) on the dataset and create dummy variables for necessary columns. After this, the number of columns will be increased from 15 to 65.
- As we have unbalanced data in the Salary column, we used oversampling method to get balanced data to train our models.
- For Feature Selection, we implement Lasso Regression and after that, we take the outcome variables as independent variables and Salary as the dependent variable.
- After performing all the above steps our model will be ready to train with Logistic Regression, Naïve Bayes Classifier, k-nearest neighbors, Decision Tree and Support Vector Machine algorithms.

B. Data Training

Here, we implement the five algorithms (namely, Logistic Regression, Naïve Bayes Classifier, K-NN, Decision Tree, and SVM) with the oversampled train data and the original train data. The results obtained are shown below in Tables I and II. For the oversampled train data, SVM offers the best precision value (0.99), Decision Tree offers the best recall value (0.86), and Naïve Bayes Classifier and SVM offer the best F1 score (0.88). However, for the original train data, SVM offers perfect precision value (1.00) and Decision Tree offers the best F1 score (0.89).

TABLE I. Results obtained from Oversampled Train Data

Model	Results using Oversampled Data			
	Precision	***Recall***	***F1-score***	***Support***
Logistic Regression	0.93	0.81	0.87	9769
Naïve Bayes Classifier	0.95	0.82	0.88	9769
K-NN	0.68	0.84	0.75	9769
Decision Tree	0.81	0.86	0.83	9769
SVM	0.99	0.80	0.88	9769

TABLE II. Results obtained from Original Train Data

Model	Results using Original Data			
	Precision	***Recall***	***F1-score***	***Support***
Logistic Regression	0.97	0.81	0.88	9769
Naïve Bayes Classifier	0.95	0.82	0.88	9769
K-NN	0.88	0.82	0.85	9769
Decision Tree	0.99	0.81	0.89	9769
SVM	1.00	0.79	0.88	9769

Table III presents the results of the salary prediction models in terms of False Negative and False Positive values.

TABLE III. Results from Using Original and Oversampled Train Data

Algorithm	Model Comparison			
	False Negative		***False Positive***	
	Original	***Oversampled***	***Original***	***Oversampled***
Logistic Regression	1700	1587	233	505
Naïve Bayes Classifier	1587	1586	376	406
k-nearest neighbors	1463	995	851	2365
Decision Tree	1679	1015	74	1414
Support Vector Machines	1962	1850	8	93

C. Model Comparison

After implementing all the algorithms, we compare the results obtained for an algorithm with another algorithm. Table IV presents the results of the salary prediction models in terms of overall accuracy. According to the simulation results, Logistic Regression (original data), Naïve Bayes Classifier (oversampled data), SVM (both datasets), and K-NN all offer the same accuracy (0.80).

TABLE IV. RESULTS USING ORIGINAL AND OVERSAMPLED DATA

Algorithm	Model Comparison	
	Accuracy || *F1 Score*	
	Original	***Oversampled***
Logistic Regression	0.80 || 0.88	0.79 || 0.87
Naïve Bayes Classifier	0.79 || 0.88	0.80 || 0.88
k-nearest neighbors	0.76 || 0.85	0.66 || 0.75
Decision Tree	0.82 || 0.89	0.75 || 0.83
Support Vector Machines	0.80 || 0.88	0.80 || 0.88

From the simulation results, it is observed that the accuracy is highest for the Decision Tree with original train data (0.82) (as shown in Table IV). Therefore, the Decision Tree is the best model to predict whether the salary is greater than or less than $50,000 dollars.

VI. CONCLUSION

The application of machine learning techniques for salary prediction provides valuable insights and benefits to individuals, employers, and students. Five different machine learning models, Logistic Regression, Naïve Bayes Classifier, K-NN, Decision Tree, and SVM are employed in this study. Among the models evaluated, the Decision Tree with the original train data demonstrates superior performance in determining whether an employee's salary is less than or greater than $50,000. This model exhibits a high accuracy rate of 82%, surpassing the accuracy of other models, which were equal to or less than 80%. The Decision Tree model with the original train data minimizes both false positives (74 cases) and false negatives (1679 cases). This reduction in misclassifications enhances the reliability and precision of salary predictions. Consequently, the model becomes a valuable tool for employees, employers, and students seeking to make informed decisions about salaries.

The insights gained from the machine learning models can inform broader workforce planning and human resource management strategies. Organizations can leverage accurate salary predictions to allocate resources effectively, identify skill gaps, and develop targeted training programs.

ACKNOWLEDGMENT

We thank Nomu Sai Badugu, Gnana Deepika Pathuri, Pranavi Reddy Lakkadi, Siril Raj Singa, and Priyanka Turaka (all are master's students) for assistance with the machine learning code and the initial manuscript.

REFERENCES

[1] Jerrim, J., "Do college students make better predictions of their future income than young adults in the labor force?" Education Economics, 23:2, 162-179, 2013.

[2] Hamlen, K.R. and Hamlen, W.A. "Faculty salary as a predictor of student outgoing salaries from MBA programs," Journal of Education for Business, 91:1, 38-44, 2015.

[3] Navyashree, M., Navyashree, M.K., Neetu, M., Pooja, G.R., and Arun, B. "Salary Prediction in It Job Market," International Journal of Computer Sciences and Engineering, 2019.

[4] Khongchai, P. and Songmuang, P. "Implement of Salary Prediction System to Improve Student Motivation using Data Mining Technique," International Conference on Knowledge, Information and Creativity Support Systems, Yogyakarta, Indonesia, 2016.

[5] Magel, R. and Hoffman, M. "Predicting Salaries of Major League Baseball Players," International Journal of Sports Science, 5.2, 51-58, 2015.

[6] Lee, Y-J. and Sabharwal, M. "Education–Job Match, Salary, and Job Satisfaction Across the Public, Non-Profit, and For-Profit Sectors: Survey of recent college graduates," Public Management Review, 18:1, 40-64, 2014.

[7] Zhang, Z., Chen, Y., Chen, Z., Zhang, H., and Jiang, H. "Employee Salary Prediction Based on Random Forest Algorithm," Proceedings of the ACM International Conference on Computational Intelligence and Data Science, pp. 18-22, 2020.

[8] Chen, Y., Liu, S., and Liu, W. "A deep learning model for salary prediction based on recurrent neural networks," IEEE conference, 2019.

[9] Chen, J., Mao, S., and Yuan, Q. "Salary prediction using random forest with fundamental features," Proc. SPIE 12167, third International Conference on Electronics and Communication; Network and Computer Technology, 2021.

[10] Das, S., Barik, R., and Mukherjee, A., "Salary Prediction Using Regression Techniques," Industry Interactive Innovations in Science, Engineering and Technology, 2020.

[11] Gopal1, K., Singh, A., Kumar, H., and Sagar, S. "Salary Prediction Using Machine Learning," 1, 2, 3-B. Tech CSE, 4-Professor Galgotias University, Greater Noida IJRIT.

[12] Lothe, D.M., Tiwari, P., Patil, N., Patil, S., and Patil, V. "Salary Prediction Using Machine Learning," International Journal of Advance Scientific Research and Engineering Trends, 2021.

[13] Li, X., Yu, Y., and Xiong, C. "Salary prediction model based on random forest algorithm," International Journal of Intelligent Systems and Applications, 2018.

[14] Wang, Q., Feng, Z., and Lv, L. "Salary prediction for job seekers: A study using ensemble learning methods," PLoS ONE, 2017.

[15] Yousaf, M., Saba, T., Niazi, M.A., Tariq, A., Iqbal, W., and Akhtar, N. "Hybrid support vector regression model for salary prediction in the healthcare industry," Computational and Mathematical Methods in Medicine, 2020.

[16] Goldberger, J., Roweis, S., Hinton, G., and Salakhutdinov, R., "Neighbourhood Components Analysis," Advances in Neural Information Processing Systems, Vol. 17, pp. 513-520, 2005.

[17] "Nearest Neighbors," 2023. https://scikit-learn.org/stable/modules/neighbors.html.

[18] "Complexity, Decision Tree," 2023. https://scikit-learn.org/stable/modules/tree.html#complexity.

[19] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J., "LIBLINEAR: A Library for Large Linear Classification," Journal of Machine Learning Research, 9(2008), 1871-1874, 2008.

[20] "Complexity, Support Vector Machines," 2023. https://scikit-learn.org/stable/modules/svm.html#complexity