IMPROVING SALARY OFFER PROCESSES WITH CLASSIFICATION BASED MACHINE LEARNING MODELS

Rukiye Kaya[1,a], Mehtap Saatçi[1], Mehmet Gökhan Bakal[2]
[1]Abdullah Gül University, Industrial Engineering Department,
[2]Abdullah Gül University, Computer Engineering Department,
Türkiye
[a]Corresponding Author: rukiye.kaya@agu.edu.tr

2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP) | 979-8-3315-3149-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IDAP64064.2024.10710706

Abstract*— In job applications, salary is major motivational factor for employees and making accurate salary prediction is crucial for both employers and employees. Utilizing advanced technologies can significantly enhance the accuracy and efficiency of salary prediction process. In this study, we explore Machine Learning (ML) methods to enhance salary prediction process. We evaluated seven classification models for predicting salary categories, with the Artificial Neural Network (ANN) model achieving the highest accuracy at 58.2% on the test dataset, followed by the K-Nearest Neighbors (KNN) model with an accuracy of 56.8%. Additionally, we employed ensemble models to further enhance prediction accuracy. Among these, the Majority Voting Classifier using Hard Voting achieved the highest accuracy at 59.3%, demonstrating the potential of ensemble techniques in refining salary predictions. The developed salary prediction tool estimates the most appropriate salary category for each candidate and help mitigate potential biases in manual salary assessments, hence enables a more objective and consistent compensation system. **CRITICAL: Do Not Use Symbols, Special Characters, or Math in Paper Title or Abstract, and do not cite other papers in the abstract.

Keywords — Salary Prediction, Machine Learning, Artificial Neural Network (ANN), K-Nearest Neighbors (KNN)

I. INTRODUCTION

In the employment process, accurate determination of salary is critical for both employees and employers, here the primary goal is to ensure equitable salary offerings. Human Resources (HR) departments face significant challenges in determining appropriate salary for a job position, while considering the candidate's skills. This issue is critical as it directly impacts both the employer's ability to attract and retain talent and the employee's satisfaction and motivation. The difficulty arises from the need to balance internal equity, external market conditions, and the unique qualifications and competencies of each candidate.

In this study, we offer an effective decision support system leverages advanced artificial intelligence techniques to estimate appropriate salary categories for job candidates to enhance strategic decision-making of HR departments, ultimately benefiting both the organization and its employees. We utilize machine learning (ML) techniques to increase the efficiency and accuracy of the salary offer process. The proposed salary category estimation with classification-based machine learning models, focuses on predicting appropriate salary categories for job candidates based on various attributes and historical data. By analysing historical salary data and candidate attributes, the system provides data-driven salary recommendations, helping to mitigate wage inequality among employees. This AI-driven approach not only supports fair compensation practices but also aids in the development of standardized salary offer processes. Together, these integrated systems aim to enhance human resources operations by providing robust, AI-driven tools for salary estimation. This approach ensures a more equitable and efficient salary prediction process, ultimately contributing to better hiring decisions and improved employee satisfaction. By employing multiple classification algorithms, the system aims to provide accurate salary recommendations, thereby reducing wage inequality and streamlining the salary offer process.

The rest of the paper is constructed as follows. The literature review presents general background knowledge about the relevant topics, while the methodology section presents proposed systems in detail along with corresponding subsections. The results & evaluations section demonstrates experimental results and their evaluations. Ultimately, the conclusion section concludes the study with an overall summary.

II. LITERATURE REVIEW

Zhen Quan and Raheem [1] presents studies on human resources visualization analytics and salary forecast modelling. In the study, salary estimation is considered as a classification problem. HP Regression and HP Tree Models, which use automatic parameter optimizations and complex structures, were used to achieve high prediction accuracy. At the same time, the Ensemble model of these two models was also implemented. The proposed Ensemble model, which combines Optimized HP Regression and HP Tree, resulted in a model with 56% classification accuracy. Similarly, Umar Zambuk et al. [2] has focused on evaluating the effectiveness of various classification algorithms for predicting employee salary categories. This study addresses classification challenges as small sample sizes, high dimensionality, and noise by proposing an advanced approach that incorporates Principal Component Analysis (PCA) and a Deep Neural Network (DNN) model. The proposed method,

which involves selecting relevant features and utilizing a DNN for classification, demonstrates superior performance compared to traditional machine learning methods like Decision Trees (DT) and Random Forests (RF). DNN model achieves a Mean Absolute Error (MAE) of 94.9%, outperforming DT and RF, which have MAE scores of 89.6% and 76.4%, respectively.

Wang et al. [3] examines the factors that influence the starting salaries of finance and economics university graduates, predicting whether their starting salaries are high. Human capital, social capital and labor market were used as explanatory variables in the research. The effects of these factors on the dependent variable were investigated with the logistic regression model. It also used five machine learning methods – SVM, Naive Bayes, CART, Random Forest and XGBoost – to predict whether graduates would earn a high starting salary. The main factors affecting the starting salary of graduates have been identified as human capital accumulation and social capital. It has been stated that integrated models perform better in predicting whether the starting salary will be high or not. The integrated model, XGBoost, was concluded as the best predictive model. The use of 2020 data in the study is a limitation and the importance of studies on the employment of female university students is emphasized.

Matbouli et al. [4] utilize various supervised ML algorithms to create salary prediction models using survey data from the Saudi Arabian labor market. Bayesian Gaussian Process Regression notably improved accuracy, with R^2 rising from 0.50 to 0.98 and reducing errors significantly compared to multiple linear regression. Artificial Neural Networks were superior in predicting salaries across major occupational groups, enhancing R^2 from 0.62 to 0.94 and decreasing errors by about 60%. This framework effectively estimates annual salaries across different economic activities, organizational sizes, and job categories. Sukumar et al. [5] also explore the complex process of forecasting salaries for job postings by combining advanced web scraping techniques with machine learning algorithms. It involves dataset of job postings from various online platforms through web scraping and comprehensive surveys. The resulting model, developed using Python and integrated into the Flask web framework. Asaduzzaman et al. [6] focuses on creating a salary prediction system using machine learning techniques. It utilizes a dataset from the 1994 census, comprising 32,561 employee records. The purpose of the study is to determine whether the person's salary is above or below $50,000. For this purpose, a total of 5 machine learning models were used, including Logistic regression, Decision tree, Naive Bayes Classifier, K-Nearest Neighbor, and Support Vector Machine. The decision tree model was found to exhibit a higher accuracy rate than other models with original training data (82%). It has been reported that this study can inform broader workforce planning and human resource management strategies. Similarly, Kablaoui and Salman [7] utilizes three supervised ML techniques—linear regression, random forest, and neural networks to analyze a dataset of over 20,000 salary records from the USA. The results demonstrate that neural networks achieve the highest accuracy at 83.2%, while linear regression provides the fastest training time at 0.363 seconds. The study highlights the superior accuracy of neural networks compared to other models and the efficiency of linear regression in model training.

III. Classification-Based Machine Learning Models

In this study, the considered machine learning problem is a classification problem. The classification-based machine learning models are presented.

A. Decision Tree Classifier

The decision tree method is a popular machine learning algorithm used for both classification and regression tasks [8]. Information Gain, Gain Ratio and Gini Index metrics are used for decision tree algorithm feature selection [9]. The advantages of decision tree methods include simplifying complex relationships, being easy to understand and interpret, and being able to perform operations without the need for data transformation. On the other hand, the model also has disadvantages such as small data sets may be subject to overfitting or underfitting and inappropriate variables are selected [8].

B. Random Forest Classifier

Random Forest is a popular method used in machine learning for classification and regression tasks and belongs to the family of decision tree-based algorithms. This method uses the CART algorithm and is characterised by good robustness to noise and outstanding performance in classification ability [10]. The advantages of this method include that it can estimate the generalisation error, effectively solve the unbalanced classification problem, and process data efficiently. Its disadvantage is that it easily leads to instability and overfitting.

C. Multinomial Logistic Regression

Multinomial logistic regression is a regression technique used to solve multiclass classification problems. It allows predicting more than one class, as opposed to two classes [11]. In the working principle of the model, to predict a class label, the probability of the independent variables being in a certain category is calculated and the category with the highest probability is selected. Advantages of this method include that model results are easy to interpret and allow for appropriate probability measurements.

D. K-Nearest Neighbors Algorithm

K-Nearest Neighbors (K-NN) algorithm is a simple yet powerful supervised machine learning algorithm used for classification and regression tasks. There are two important concepts for this algorithm. The first of these is the distance parameter. KNN determines the class with the maximum number of points sharing the least distance to the data point that needs to be classified [12]. The second important factor is the determination of the number of neighbours (k). The appropriate choice of k has a significant impact on the diagnostic performance of the KNN algorithm [13]. The advantages of this algorithm include ease of interpretation, short calculation time and high predictive power [12]. Its disadvantages include the difficulty of determining the appropriate k number, the need for a lot of storage space, and the difficulty of determining the distance algorithms correctly.

E. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning algorithm used for classification and regression tasks. Its main

purpose is to find the hyperplane that best separates different classes in the feature space. The advantages of this method include its good theoretical foundations and high generalisation performance [14]. Its disadvantages are that it has some limitations regarding parameter selection, algorithmic complexity, multi-class datasets and unbalanced datasets.

F. XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a powerful and efficient machine learning algorithm used in classification and regression problems. It belongs to the category of community learning that powers algorithms. The advantages of this method are that the model works quickly and handles large amounts of data effectively [15].

G. Artificial Neural Network

Artificial Neural Network (ANN) is a computational model inspired by the structure and functioning of the neural networks of the human brain. ANN consists of multiple layers of simple processing elements called neurons [16].

To provide a more comprehensive understanding of the employed machine learning classifiers, we investigate their architectural details and discuss their relevance to the salary prediction task. Decision Tree classifiers utilize a hierarchical tree-like structure to partition the data based on feature values, which culminates in leaf nodes representing predicted salary categories. Random Forest classifiers combine multiple decision trees through bootstrapping and random feature selection, which enhances prediction accuracy and robustness. Multinomial Logistic Regression employs the softmax function to estimate probabilities for each salary category based on a linear combination of independent variables. K-Nearest Neighbors classifies new data points based on the majority class among its k-nearest neighbors in the feature space, influenced by the distance metric and the choice of k. Support Vector Machines aim to find the optimal hyperplane that maximizes the margin between different salary categories, utilizing kernel functions to handle non-linearly separable data. XGBoost, a gradient-boosting algorithm, sequentially adds weak learners (decision trees) to improve prediction accuracy, incorporating regularization techniques to prevent overfitting. Artificial Neural Networks consist of interconnected layers of neurons that learn through the adjustment of weights and biases during training. Our study employed a specific ANN model to capture complex relationships implicitly encoded within the data. These diverse architectures offer distinct approaches to learning patterns and making predictions, which ultimately contribute to a comprehensive evaluation of their suitability for the salary prediction task.

IV. METHODOLOGY

A. Dataset

In this study, the dataset used includes employee information received by a private company. A data set consisting of 1820 people working in the company was used for Salary Category Estimation with Classification-Based Machine Learning Models. This data set includes employees' experience, education, competency information and salary information. The dataset focuses on predicting salary categories of new candidates based on employee data. The predicted data in this dataset is the person's salary category. Salary categories consist of 7 categories in total. These categories were pre-processed by the company and delivered to us. Therefore, it is unknown which category represents which salary range. It is known that as the number of categories increases, the salary range also increases. The categories included in this data are given below.

TABLO I. CATEGORY DISTRIBUTION FOR "SALARY GIVEN CATEGORY"

Category	Frequency
0	34
1	179
2	363
3	922
4	245
5	68
6	9

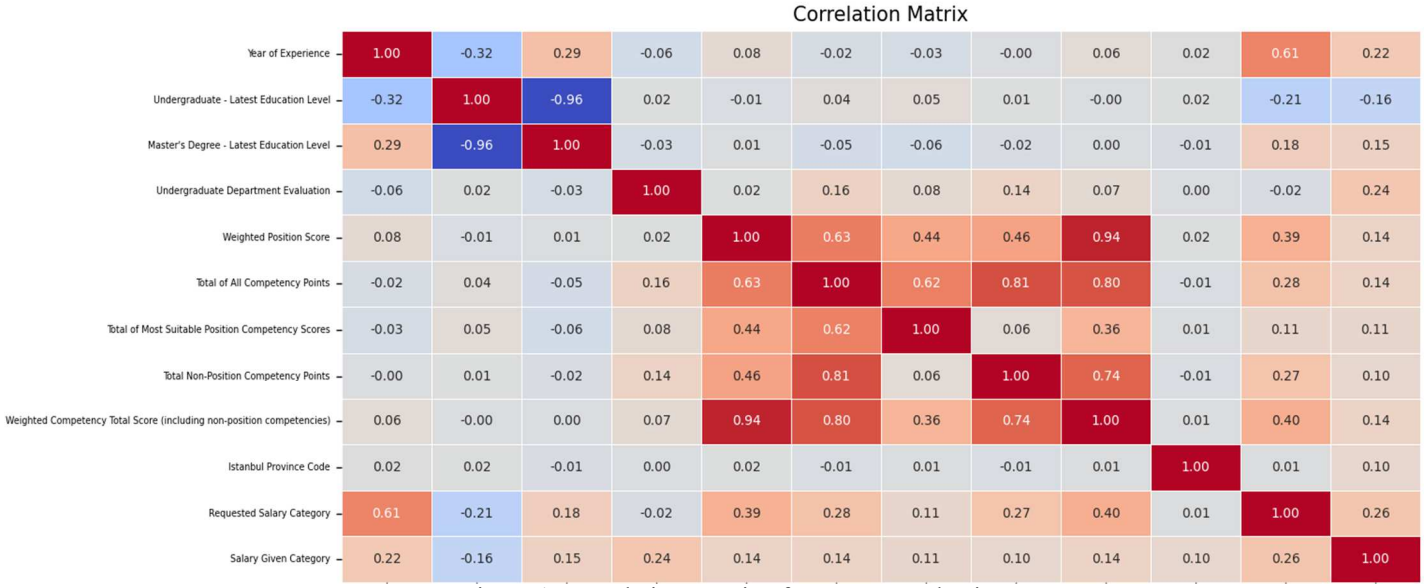

Figure 1. Correlation Matrix after Feature Selection

B. Feature Selection

The feature selection process was conducted using correlation analysis to determine the independent variables that were most strongly correlated with the dependent variable “Salary Category”. The correlation between each independent variable and the dependent variable was examined, and variables with a correlation coefficient of 0.1 or greater or -0.1 or less were considered significant. This analysis resulted in the selection of 11 independent variables to be included in the machine learning models. The final independent variables in our data and their correlations with the dependent variable are shown in the figure below. In addition, in order to increase the accuracy of feature selection, two more methods were applied for the selection of the best features: K-Best f_classif and K-Best mutual_info_classif. With these two methods, the relationships between the features in the data and the target variable were evaluated. The features selected in common by both methods include variables such as “Master’s Degree - Latest Education Level”, “Undergraduate Department Evaluation”, “Year of Experience” and “Weighted Position Score”. This shows that the features have a consistently high impact on the predictive performance of the model. Therefore, verifying the features selected within the correlation analysis with these methods can be considered as an important strategy to increase the accuracy and generalization capacity of the model. There are also features that are different within the two methods. These are the candidate's GPA and HR Interview score, but they were not included in the study because they are not common to both methods.

"Year of Experience" data represents the candidate's total years of experience. This data takes a discrete integer value between 0 and 21. "Undergraduate - Latest Education Level" data takes the value 1 or 0 depending on whether the person’s last education status is undergraduate or not. "Master’s Degree - Latest Education Level" data takes the value 1 or 0 depending on whether the person’s last education status is master’s degree or not. The "Undergraduate Department Evaluation" data includes values 1,2,3,4 and 5. If the person's department is in engineering and space sciences, this value is in category 5. Other relevant departments take other relevant values.

For the "Weighted Position Score" data, it was thought that the position the employee works in could be related to the person’s salary. Therefore, it was thought that a connection could be established between the competency scores of all employees and the position they work in, and that its consistency could be checked by creating an artificial column. The position column was produced using data previously obtained from the company regarding which competencies were sought for which position. Using this data, it was decided to create a new column where estimated position assignments would be made according to the competency scores entered by the candidates and then a score would be assigned according to how competent they were for this position. To create this, it was first determined for which position the person had the most competence. Then, a score column was created by dividing the sum of the competencies required for this position by the number of competencies required for that position. After it was seen that this score column had a satisfactory value in terms of correlation, it was decided to include it in the model based on this logic.

The data "Total of All Competency Points" was collected without discrimination between all points entered for all qualifications of employees and assigned as a value to this column.

"Total of Most Suitable Position Competency Scores" was created by adding up the employee's total qualification points for the qualifications required for that position.

"Total Non-Position Competency Points" were obtained by subtracting the qualifications required for the position to which

the employee is assigned and adding the scores of all remaining qualifications without distinguishing between them.

The "Weighted Competency Total Score (including non-position competencies)" data was created by taking the values of two different columns in certain proportions. The previously calculated Weighted Position Score column was multiplied by 0.75 to get 75%, as it is important for the candidate's final score, and the less important Non-Position Qualification Scores column is multiplied by 0.25 to get 25%, and the value in this column was created by adding these two values.

"Istanbul Province Code" is the data whether the person lives in Istanbul or not and takes the value 0 and 1. "Requested Salary Category" is a data containing the categorical form of the salary amount requested by the person at the time of application. "Salary Given Category" data contains the salary category given to the person.

C. *Salary Category Estimation with Classification-Based ML Models*

We developed a salary category estimation system using classification-based machine learning models. The goal is to create a data-driven approach that can automate assigning salary classes to new hires, potentially improving efficiency and fairness within the company's salary structure. By leveraging ML algorithms, the system aims to analyse historical data related to employee attributes and their corresponding salary categories to learn patterns and predict appropriate salary categories for new candidates. The proposed work investigates and evaluates the performance of various classification algorithms, aiming to identify the most suitable model for accurately predicting salary categories based on available employee data. The salary categories given in the dataset are shown in Table 1. It shows the distribution of instances across the seven salary categories. Category 3 is the most frequent one, representing 922 employees, while Category 6 is the least frequent with only 9 employees. This uneven distribution highlights the potential challenge of class imbalance when training the machine learning models.

V. EXPERIMENT AND RESULTS

In the experiments, 80% of the total data was used as train and 20% as test. There are 7 categories in total in our target variable. Cross-validation was used to improve performance while applying the models. Variables obtained within the correlation analysis were used. Additionally, K-Best Feature was tested and it was observed that the accuracy of the model increased as the number of variables increased. Therefore, all independent variables (a total of 11) obtained as a result of the correlation were used for the models. The results obtained with machine learning methods are explained in detail below. Accuracy values are taken as weighted. The table below shows the accuracy rates.

As presented in the table, Artificial Neural Network (ANN) is the model that gives the highest accuracy with an accuracy rate of 0.582. The high accuracy rate shown by this model can generally be due to its capacity to learn complex relationships and data patterns. The highest accuracy and F1 score were obtained in this model, indicating that the model effectively learned the features in the dataset. The true positive detection rate of this model is the highest compared to other models. This model is followed by the K-Nearest Neighbors Algorithm with an accuracy rate of 0.568. This algorithm makes predictions using the interactions of close neighbors and can perform well in small datasets. It can be said that this model shows high performance in the small dataset used in the project. The model has a medium positive class prediction performance with a Recall rate of 0.568. These two models are followed by the Random Forest Classifier with an accuracy rate of 0.563. The accuracy of the Random Forest Classifier depends on its ability to reduce noise in the dataset and learn the interaction of features. For these reasons, it can be said that the model has a good level of performance. Other models exhibited poor performance depending on the specific classification difficulty and features in the dataset. The performance differences in the models included in the study are due to the basic features of the algorithms used. Support Vector Machines (SVM) and XGBoost Classifiers are in the fourth and fifth places with 0.541 and 0.546 accuracy rates, respectively. Although Support Vector Machines (SVM) and XGBoost have better learning capacity for the distinctions in the dataset, correct hyperparameter settings and data preprocessing may affect the accuracy of these models. Decision Tree and Multinomial Logistic Regression were determined as the lowest performing models with 0.497 and 0.519 accuracy rates, respectively. Decision Trees can reduce the accuracy in the model when they tend to overfit. Multiple Logistic Regression is effective in modeling linear relationships, but may have limited or poor performance in multi-category problems.

The table below presents a comparison of the accuracy rates for machine learning models in predicting salary categories. The observed accuracy rates, which range between 0.50 and 0.60, can be attributed to the substantial number of salary categories and the limited data available for categories with extremely low or high values. Given these constraints, the obtained accuracy rates are deemed adequate for the current problem. It is anticipated that with an increase in data volume, these models will be re-evaluated, potentially leading to improved accuracy rates.

TABLO II. MODEL RESULTS

	Classification-Based Models						
	Decision Tree Classifier	***Random Forest Classifier***	***Multinomial Logistic Regression***	***K-Nearest Neighbors Algorithm***	***Support Vector Machine (SVM)***	***XGBoost Classifier***	***Artifical Neural Network (ANN)***
Accuracy	0.497	**0.563**	0.519	**0.568**	0.541	0.546	**0.582**
Precision	0.471	**0.512**	0.424	**0.540**	0.606	0.494	**0.518**
Recall	0.497	**0.563**	0.519	**0.568**	0.541	0.546	**0.582**
F1-Score	0.481	**0.519**	0.413	**0.548**	0.405	0.505	**0.524**

We also utilized the ensemble models and accuracy rates are given in below table.

TABLO III. ENSEMBLE MODEL RESULTS

	Ensemble Models				
	Majority Voting Classifier (Hard Voting)	***Majority Voting Classifier (Soft Voting)***	***Majority Voting Classifier (Weighted Voting)***	***Bagging Ensemble Model***	***Stacking Ensemble Model***
Accuracy	**0.593**	0.577	0.574	**0.580**	0.571
Precision	**0.547**	0.526	0.517	**0.541**	0.485
Recall	**0.593**	0.577	0.574	**0.580**	0.571
F1-Score	**0.553**	0.534	0.530	**0.538**	0.499

In order to further increase the prediction accuracy of salary categories, ensemble models were also used. The accuracy rates for ensemble models are presented in the table above. When these models are examined, the Majority Vote Classifier using the Hard Voting method reached the highest accuracy rate of 0.593. Since the final decision is made according to the majority's prediction in this model, the model showed high performance. It can be said that the high accuracy rate in the model is obtained as a result of reducing the error rate by combining different models. The Hard Voting model is followed by the Bagging Ensemble Model with an accuracy rate of 0.580. In this method, accuracy can be increased by combining more than one model. Soft Voting and Weighted Voting using Majority Voting Classifiers are in the third and fourth place with accuracy rates of 0.577 and 0.574, respectively. Soft Voting was able to reach a balanced result by taking into account the probability distribution of the output of each model. Weighted Voting allows adjusting the balance by making certain models more effective. The Stacking Ensemble Model ranks fifth with an accuracy rate of 0.571. While Hard Voting Classifier and Bagging Ensemble Model methods stand out with their high accuracy rates, other methods have lower but still acceptable performance.

Despite the application of ensemble methods, which typically enhance model performance by combining the strengths of multiple individual models, the accuracy rates remained within the range of approximately 0.57 to 0.59. This modest improvement in accuracy can be attributed to several factors. The target variable comprises seven distinct salary categories, with certain categories having significantly fewer instances than others, leading to imbalanced data. This imbalance presents a challenge for both individual and ensemble models, as they may struggle to accurately predict categories with sparse data. Additionally, predicting salary categories is inherently complex due to the myriad of factors influencing salary, many of which may not be fully captured in the available data. The limited explanatory power of the independent variables, despite their selection through correlation analysis, further complicates the task of achieving high accuracy. Moreover, while ensemble methods like Majority Voting, Bagging, and Stacking are designed to leverage the strengths of multiple models, their effectiveness relies on the diversity and complementary nature of the base models. In this case, the base models may share similar weaknesses, particularly in handling the imbalanced and complex nature of the salary prediction task, leading to only marginal gains in performance. In conclusion, although the ensemble models provided a slight improvement in accuracy, the overall accuracy rates reflect the challenges inherent in the data and the prediction task. Future efforts to increase data volume, especially for underrepresented salary categories, and to incorporate more informative features may help enhance the predictive performance of both individual and ensemble models.

Finally, an interface was created using Python code for the salary category prediction system with machine learning. In order for the model to learn the data, the user must first upload the data to the system in the appropriate format (.csv) using the "Open" button. The user is then expected to enter the information of the candidate whose salary category they want to estimate into the system. Undergraduate - Latest Education Level, Master's Degree - Latest Education Level and Province Code fields expect a value of 0 or 1 from the user. The Undergraduate Department Evaluation section takes a value from 1 to 5, depending on the candidate's department. Other parts can take numbers as integers or decimals without any limit. The figure below shows the data entered into the system. Then, when the "Predict Salary Class" button is pressed, the predicted salary category for the candidate appears on the screen. In the example below, the salary category for the candidate whose data is entered is estimated as 4.

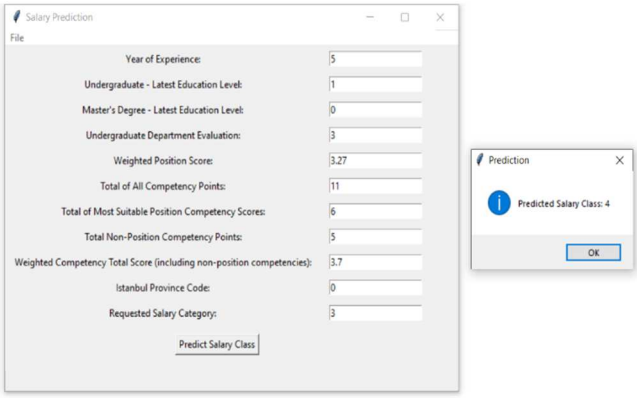

Figure 2. Display of the Predicted Category in the Interface

VI. CONCLUSION

The developed data-based salary class prediction system with machine learning models provides effective and fair salary offer process for the company. A cluster-based forecasting system was developed to determine the candidate's salary category. Within this system, the person's data was entered and the salary category suitable for the candidate was automatically estimated. Since the K-Nearest Neighbor method gave the best results, this model was used to create the interface. Individual

and ensemble classification models were tested and accuracy rates in range of 50% - 60% were obtained. The pre-process of the data was realized by the company and due to the scarcity of some class data accuracy couldn't be improved. As a future research, the tests can be conducted with larger data set by utilizing other machine learning and regression models.

REFERENCES

[1] T. Zhen Quan and M. Raheem, "Human Resource Analytics on Data Science Employment Based on Specialized Skill Sets with Salary Prediction," 2023.

[2] F. Umar Zambuk et al., "Salary Prediction Model using Principal Component Analysis and DeepNeural Network Algorithm," 2023. [Online]. Available: https://www.researchgate.net/publication/378207843

[3] P. Wang, W. Liao, Z. Zhao, and F. Miu, "Prediction of Factors Influencing the Starting Salary of College Graduates Based on Machine Learning," Wirel Commun Mob Comput, vol. 2022, 2022, doi: 10.1155/2022/7845545.

[4] Y. T. Matbouli and S. M. Alghamdi, "Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations," Information (Switzerland), vol. 13, no. 10, Oct. 2022, doi: 10.3390/info13100495.

[5] J. G. Sukumar, M. S. Ram Reddy, N. Sambangi, S. Abhishek, and T. Anjali, "Enhancing salary projections: a supervised machine learning approach with flask deployment," in Proceedings of the 5th International Conference on Inventive Research in Computing Applications, ICIRCA 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 693–700. doi: 10.1109/ICIRCA57980.2023.10220707.

[6] A. Asaduzzaman, M. R. Uddin, Y. Woldeyes, and F. N. Sibai, "A Novel Salary Prediction System Using Machine Learning Techniques," in Proceedings - 2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering, ECTI DAMT and NCON 2024, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 38–43. doi: 10.1109/ECTIDAMTNCON60518.2024.10480058.

[7] R. Kablaoui and A. Salman, "Machine Learning Models for Salary Prediction Dataset using Python," in 2022 International Conference on Electrical and Computing Technologies and Applications, ICECTA 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 143–147. doi: 10.1109/ICECTA57148.2022.9990316.

[8] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," Shanghai Arch Psychiatry, vol. 27, no. 2, pp. 130–135, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.

[9] H. H. Patel and P. Prajapati, "Study and Analysis of Decision Tree Based Classification Algorithms," International Journal of Computer Sciences and Engineering, vol. 6, no. 10, pp. 74–78, Oct. 2018, doi: 10.26438/ijcse/v6i10.7478.

[10] L. Yingchun and Y. Liu, "Random forest algorithm in big data environment," 2014. [Online]. Available: www.cmnt.lv

[11] A. M. EL-HABIL, "An Application on Multinomial Logistic Regression Model," Pakistan Journal of Statistics and Operation Research, pp. 271–291, 2012.

[12] Taunk Kashvi, Sanjukta De, Srishti Verma, and Swetapadna Aleena, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019.

[13] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," Ann Transl Med, vol. 4, no. 11, Jun. 2016, doi: 10.21037/atm.2016.03.37.

[14] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[15] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[16] A. D. Dongare, R. R. Kharde, and A. D. Kachare, "Introduction to Artificial Neural Network," 2008.