Salary Classification & Prediction based on Job Field and Location using Ensemble Methods

1st Jocelyn Verna Siswanto
Data Science Program, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
jocelyn.siswanto@binus.ac.id

2nd Laurentia Alyssa Castilani
Data Science Program, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
laurentia.castilani@binus.ac.id

3rd Natasha Hartanti Winata
Data Science Program, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
natasha.winata@binus.ac.id

4th Nathania Christy Nugraha
Data Science Program, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
nathania.nugraha@binus.ac.id

5th Noviyanti T M Sagala
Statistics Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
noviyanti.sagala@binus.edu

***Abstract*— The economy is one of the determinants of how a person can live their life. In this current economic situation, inflation occurs everywhere, causing the prices of necessities to rise. In order to have a decent life, people must find a job with the highest possible salary to fulfill their needs. Various job industries have their salary range. Obtaining the information of salary level for the respective job is helpful for employers and employees to estimate the expected salary. This work aims to classify the salary level of jobs available in Indonesia and determine whether those salaries are decent enough. The learning methods are logistic regression, decision tree, k-nearest neighbor, support vector machine, voting classifier, bagging classifier, random forest, and boosting classifier. Random Forest achieved the best result with an accuracy rate of 72%. Based on the analysis result, factors such as job field, educational background, working experience, working hours, and job location influence salary.**

Keywords—salary, job field, location, machine learning, ensemble methods, predictive model

I. INTRODUCTION

Basic human needs, such as physiological, personal, or socio-economic, can be met financially [1]. Economic needs are necessities for every class in society—the need results from an income level at or below the poverty line [2]. Economic need is one of the determinants of how people can live their lives. Many necessities, such as food and clothing, depending on people's income through economic activities. Not to mention the many additional needs such as education, health, vehicles, and entertainment [3]. However, in this current economic situation where inflation happens in most countries all over the world, the prices of necessities tend to rise.

Indonesia is one of many countries in the world that are facing an inflation rate. According to Indonesia's Central Bureau of Statistics, BPS, the monthly inflation rate for September 2022 was 1.17%. It was the highest rate since December 2014. Rice, fuel, and inner-city transportation are some of the most critical factors contributing to September's inflation [4]. This situation leads to societal confusion on balancing primary, secondary, and tertiary needs. Therefore, this is one of the main reasons why people looking for a job tend to look for jobs with a high salary and seek the highest possible salary available to fulfill all their needs [5].

Several factors include what position or job title the person has, the field of work where that person works, the working experiences that the person had, what educational background the person has, and where the work location of the person influences the difference in salary range in every job [6]. This research aims to predict a person's salary class in Indonesia, whether it is low-income, medium-income, or high-income, based on the minimum wage of each province, the field where the person works, and the location of the work itself. The dataset used in this research is drawn from publicly available data on Kaggle because it is the only platform that provides data on salaries for various jobs in Indonesia and the factors that affect them. The results of this work may provide insights into whether the salary can fulfill someone's needs financially or not.

This research uses a machine learning algorithm to predict the salaried class accurately by creating predictive models. With the help of a machine learning algorithm that can analyze a large volume of data quickly and recognize patterns that are not visible to humans, it can generate highly accurate predictions. Moreover, machine learning can swiftly adapt to changes because when the dataset is updated, machine learning will also adapt the predictions [7]. The machine learning methods used are individual classifiers and ensemble methods. Individual classifier consists of logistic regression, k-nearest neighbor, decision tree, and support vector machine. Meanwhile, the ensemble method consists of a voting classifier, bagging classifier, random forest, and boosting classifier.

To the best of our knowledge, there has been no research that predicts a person's salary level using a machine learning classification technique by knowing the factors that influence salary. Thus, this research is a new idea in classifying salaries (into low, medium, and high class) to answer whether the salary is at the right level to determine the balance between one's income and expenses.

II. RELATED WORK

Salary grade prediction in a job is crucial for every company [8]. A system is needed to control the assets and profits that a company receives and to prevent the company's loss.

Research conducted by Gopal predicted the salary using a machine learning regression technique. The prediction used department, program, job training, candidate certifications, and CGPA variables [8]. Other research conducted by Lothe predicted salary based on the Degree, Major, Years of Experience, and Industry types using a machine learning regression technique [9].

The educational background influences most predictions, position in the company, work location, company type (start-up/multinational), work experiences, and working hours [10]. Money is a sensitive issue for some people. Therefore, a fair and decent salary is needed as a price to pay for someone's soft and hard skills. Predicting one's salary is needed to determine whether one's incomes and expenses are balanced. This research can also answer the people who wonder whether their job is worth their skill and can fulfill their needs.

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Ensemble methods comprise a set of classifiers, e.g., decision trees, and their predictions are aggregated to identify the most popular result. The most well-known ensemble methods are bagging, bootstrap aggregation, and boosting [11].

III. METHODOLOGY

A. Dataset Description

The collected data were from the Kaggle platform collecting information about workers in Indonesia from various backgrounds on March 13, 2022 [12]. Data consists of 15 attributes and 34,746 rows of data collected. Attribute details of the collected data can be seen in Table I.

TABLE I. DATASET DESCRIPTION

Attribute	**Description**
id	the id of the data (unique)
job_title	Name of the job
location	Company location
salary_currency	The currency used in salary
career_level	Career level such as manager, CEO, etc
experience_level	Experience required for applicants
education_level	Education required for applicants
employment_type	Full-time, part-time, or internship
job_function	Category of the job
job_benefits	Benefit given by the company
company_process_time	Average time the company will give response to the applicants
company_size	Number of employees in the company
company_industry	The company's sector of service
job_description	Description of the job
salary	Salary offered by the company per month (per March 2022)

B. The Proposed Methodology

The proposed methodology contains (1) Exploratory Data Analysis, (2) Data Cleaning, (3) Data Transformation, (4) Modeling, and (5) Result, as shown in Figure 1.

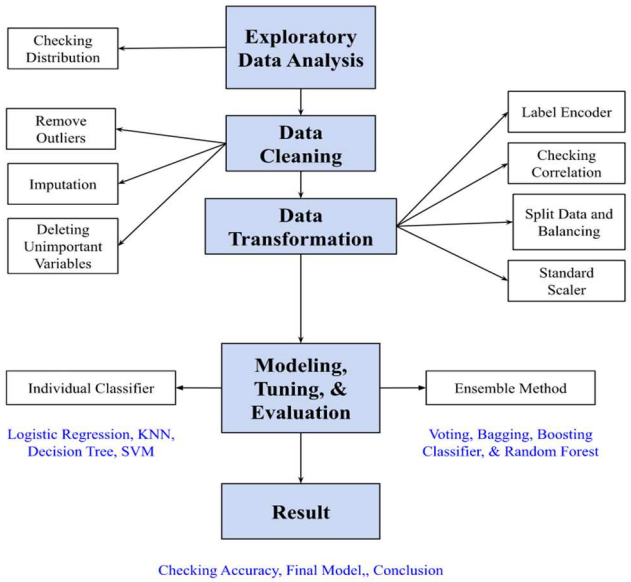

Fig.1 The Proposed methodology

1) *Exploratory Data Analysis:* The research started by importing the data using python programming. The process includes retrieving insights using the libraries and functions that are available and applicable to help.
2) *Data cleaning:* Data cleaning consists of removing outliers, imputation, encoding labels, and checking the data distribution to simplify the analysis process.
3) *Data transformation:* Data transformations cover deleting unimportant variables, scaling, checking correlations, and splitting and balancing the data. The process includes filtering out the variables with invalid values, deleting outliers, performing imputations on some missing variables, and removing the variables that are not highly correlated/ do not have a significant impact on this analysis. These actions are taken out of consideration with the help of the result produced by the data exploration. After processing the data through exploratory data analysis & data preparation, the data are divided into three parts: the training set, validation set, and testing set. These three datasets will be used in modeling and model evaluation.
4) *Modeling, Tuning, & Evaluation:* The models were created and trained using the training set. Then, the models are implemented into the testing set. Two algorithm methods were used, namely individual classifiers and ensemble methods. Ensemble methods are a combination of individual classifiers. In the testing phase, the models are evaluated, whether they were overfitting or underfitting. The model results will be compared by considering the accuracy, precision, recall, and F1-score level from the macro average. Based on these values, insight will be obtained into which attributes affect the classification process.

5) *Result:* After evaluating the model, the best model will be applied to the testing set

C. SMOTE (Synthetic Minority Oversampling Technique)

One of the popular oversampling rebalancing techniques is SMOTE. Traditional oversampling techniques replicate minority data from a minority data population [13]. The amount of data increases, but there is no new information or variation in the machine learning model. Based on this problem, Nitesh Chawla introduced a new technique to create synthetic data for oversampling purposes in 2002 [14]. SMOTE generates synthetic data by selecting random data from the minority class and applying the K-Nearest Neighbor (KNN) algorithm. Then, the synthetic data is created between the randomly chosen KNN and the random data. This process is repeated until the minority class has the same proportions as the majority class [13].

D. Machine Learning Methods

In this work, the performance of the machine learning methods was investigated, enhanced, and compared to obtain the best possible salary level in Indonesia.

1) ***Logistic Regression*** predicts a target variable by analyzing the correlation between independent variables in the dataset. It computes a weighted sum of the independent variables and generates the logistic of coefficients in the linear combinations [15].
2) ***K Nearest Neighbors*** chooses K random data as the initial centroid, then counts the distance between each centroid with all the available data. The data will be clustered based on the nearest centroid. Then each cluster will be averaged, and the data with the closest value to the average value will be the new centroid [16].
3) ***Decision trees*** recursively evaluate different features. At each node, it uses a feature that best splits the data. The decision tree uses a divide-and-conquer strategy when identifying optimal split points in the tree [17].
4) ***Support vector machine*** takes the data points and outputs a hyperplane (which has the most significant margin between it and the closest data point of each point) that is considered the best at separating [15].
5) ***Voting classifier*** trains multiple classifiers and predicts the output based on the class that achieved the majority of votes [18].
6) ***Bagging classifier*** fits each base classifier to a random subset of the original data set and aggregates the individual predictions (by matching or averaging) to form a final prediction.
7) ***Random forest*** uses different subsets of features randomly so that each tree will differ. Therefore, the variance of the trees that have been created is averaged, and the model will have a lower bias rate.
8) ***Boosting Classifier*** combines several weak classifiers into robust classifiers by training predictors sequentially and correcting mistakes that happened in the previous sequence [15].

IV. RESULT AND DISCUSSION

There are 15 attributes and 34,746 observations in the dataset. However, around 25,000 records still need values. Also, six variables are considered unimportant for building the model. Missing values in Salary Variable and less critical variables in the data are removed in the Data Cleaning stage, whereas missing values in other variables are imputed with mode. After that, in the Data Transformation stage, grouping was done on 'location' attributes so that it only consists of 34 provinces. Then, the salary attribute is grouped into three classes, namely high income, medium income, and low income, based on each province's UMP (minimum wage). The high-income class is a salary that is more than two times the UMP, the medium class has a value more than the UMP up to 2 times the UMP, and the low class is less than or equal to the UMP. This new attribute is named Salary Level.

After the Data Preparation stage (Data Cleaning and Transformation), 9343 rows remain. The remaining attributes are those that affect the prediction process in the model. These attributes are location, career_level, experience_level, employment_type, job function, company_process_time, and company_industry, including the new attribute, salary level. Standardizing the numerical attributes is carried out, and then the data is divided into three parts. In the book The Elements of Statistical Learning (2009), it is stated that to build a good predictive model, data can be divided into a training set for model building, a validation set for choosing the best model, and a testing set for the final model assessment [19]. Thus, this research uses a workflow to allocate new data to existing splits using a serial waterfall workflow. New data is divided into three parts: testing set, validation set, and training set sequentially. The testing and validation sets will always be fresh to avoid overfitting. In addition, the training set allows learning from mistakes [20]. Around 60% of the data is used as a Training Set (5605 rows), 20% of the data is used as a Testing Set (1869 rows), and 20% of the data is used as a Validation Set (1869 rows). This research uses several algorithm methods: Individual Classifiers (Logistic Regression, KNN, Decision Tree, and SVM) and Ensemble Methods (Voting Classifier, Bagging Classifier, Random Forest, and Boosting Classifier). The model will be trained using Training Data, then evaluated using a Validation Set. A model with the best performance will then be used to predict the Testing Set.

The macro average is calculated from a confusion matrix to evaluate the classifier's performance. This research uses a macro average because the number of records from every class is imbalanced, and the method treats all classes (high, medium, and low) equally. Therefore, measuring the accuracy of every model's prediction result is essential for both high-paying jobs and medium and low-paying jobs.

Accuracy shows the ratio of correct predictions out of all predictions made by the model [21]. The Accuracy formula can be seen in equation (1).

$$Accuracy = \frac{True\ High + True\ Medium + True\ Low}{Total\ Predicted\ (High + Medium + Low)} \quad (1)$$

Precision shows how accurate or precise the model is. It calculates the ratio of actual positive class (True Positive) out of all optimistic predictions [22]. The formula to calculate Precision can be seen in equation (2).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} = \frac{True\ Positive}{Total\ Predicted\ Positive} \quad (2)$$

The recall will calculate how many positive classes the model correctly predicts out of all positive classes [22]. The Recall formula stated in equation (3).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{True\ Positive}{Total\ Actual\ Positive} \quad (3)$$

F1-score combines a model's precision and recall score [22]. F1-score is used to measure each model's performance because the number of records for each class from the target variable is imbalanced. F1-score will give a better measure of the incorrectly classified cases than the Accuracy Metric. The formula of the F1-score can be seen in equation (4).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

After that, the Hyperparameter Tuning technique is carried out to improve the model's performance. The difference between default parameters and hyperparameters can be seen in Table II.

TABLE II. THE DIFFERENCE BETWEEN DEFAULT PARAMETER AND HYPERPARAMETERS

Algorithm/ Metric	Default Parameters	Hyperparameters
Logistic Regression	C: 1, penalty: None, solver: lbfgs	C: 10, penalty: 12, solver: sag
K-Nearest Neighbor	n_neighbors: 5, weights: uniform	n_neighbors: 2, weights: distance
Decision Tree	criterion: gini, max depth: None, max features: None, max leaf nodes: None, min samples split: 1	criterion: gini, max depth: 100, max features: None, max leaf nodes: 1000, min samples split: 3
Support Vector Machine	C: 10, gamma: scale, kernel: rbf	C: 10, gamma: 0.1, kernel: rbf, verbose: 3, refit: True
Voting Classifier	voting: hard, weights: None	voting: hard, weights: (1,1,2)
Bagging Classifier	n_estimators: 10, n_jobs: None, criterion: gini, max depth: None, max features: 1, max leaf nodes: None, min samples split: 1, bootstrap: True	n_estimators: 100, criterion: entropy, max depth: 100, max features: auto, max leaf nodes: 1000, min samples split 3, bootstrap: False
Random Forest	criterion: gini, max depth: None, max features: sqrt, max leaf nodes: None, min samples split: 2, bootstrap: True	criterion: gini, max depth: 100, max features: auto, max leaf nodes: 1000, min samples split: 2, bootstrap: False
Boosting Classifier (AdaBoost)	criterion: gini, max depth: None, max features: sqrt, max leaf nodes: None, min samples split: 1	criterion: gini, max depth: 100, max features: log2, max leaf nodes: 1000, min samples split: 4

The performance before and after the Hyperparameter Tuning of each model can be seen in Table III.

TABLE III. COMPUTATIONAL RESULTS

Algorithm/Metric			Default parameter	Hyperparameters
Individual Classifier	Logistic Regression	Accuracy	0.46	0.46
		Precision	0.45	0.45
		Recall	0.50	0.49
		F1-score	0.44	0.44
	K-Nearest Neighbor	Accuracy	0.57	0.64
		Precision	0.54	0.59
		Recall	0.59	0.62
		F1-score	0.54	0.60
	Decision Tree	Accuracy	0.70	0.72
		Precision	0.65	0.66
		Recall	0.67	0.69
		F1-score	0.65	0.67
	Support Vector Machine	Accuracy	0.47	0.67
		Precision	0.47	0.63
		Recall	0.49	0.62
		F1-score	0.45	0.62
Ensemble Method	Voting Classifier	Accuracy	0.62	0.65
		Precision	0.59	0.61
		Recall	0.63	0.65
		F1-score	0.58	0.61
	Bagging Classifier	Accuracy	0.75	0.76
		Precision	0.70	0.71
		Recall	0.72	0.73
		F1-score	0.71	0.72
	Random Forest	Accuracy	0.75	0.76
		Precision	0.71	0.71
		Recall	0.71	0.72
		F1-score	0.71	0.72
	Boosting Classifier	Accuracy	0.73	0.75
		Precision	0.69	0.72
		Recall	0.69	0.70
		F1-score	0.69	0.71

The particular classifier method produces a decision tree as a model with the highest accuracy in the default parameters and hyperparameters based on the accuracy of the F1 score. The decision Tree has an F1-score of 65% (obtained from a precision of 65% and recall of 67%) using default parameters and 67% (obtained from a precision of 66% and recall of 69%) after using hyperparameters. Therefore, the Decision Tree is used as a base estimator in the Bagging and Boosting Classifier methods. Meanwhile, logistic regression performed the worst compared to the remaining algorithms in this research, with an F1-score accuracy value of 44% on both

default parameters and hyperparameters (obtained from a precision of 45% and recall of 50%). Logistic regression performs worst because it is considered statistical learning and not machine learning algorithms. Statistical learning works better in finding relationships between features rather than doing predictive modeling [23]. Therefore, machine learning algorithms are better than statistical learning for this research.

Ensemble methods produce two algorithms with the highest accuracy values: the Bagging Classifier and Random Forest. Using default parameters, the F1-score of the Bagging Classifier and Random Forest is 71%. F1-score is obtained from a precision of 70% and recall of 72% from Bagging, while for Random Forest from a precision of 71% and recall of 71%. Bagging Classifier and Random Forest with hyperparameters achieved an accuracy is 72%. F1-score is calculated from a precision of 71% and recall of 73% for Bagging and for Random Forest from a precision of 71% and recall of 72%. However, the Voting Classifier produces the worst performance compared to other algorithms in the Ensemble Method with default parameters and hyperparameters of 58% (obtained from a precision of 59% and recall of 63%) and 61% (obtained from a precision of 61% and recall of 65%) respectively.

Overall, the results show that all algorithms work better with hyperparameters than with default parameters. Table II also shows that Ensemble Methods perform better than Individual Classifiers, which makes sense since the ensemble method combines many individual classifiers. Out of all classifiers in Ensemble Method, Bagging Classifier and Random Forest give the best performance in terms of F1-score. The metric results of the Bagging Classifier and Random Forest algorithms are similar, so another perspective is needed to compare the performance of these two algorithms.

In Bagging Classifier, the splitting process to create each decision tree considers all features. Meanwhile, Random Forest uses different subsets of features by selecting only some features randomly so that each tree will differ. Thus, the variance of the trees that have been created can be averaged, and the model will have a lower bias rate. Therefore, Random Forest will be the best classifier, which will be applied to predict the testing set. The prediction results of the Random Forest classifier and the actual results can be seen in Table IV.

TABLE IV. RANDOM FOREST CONFUSION MATRIX IN VALUES

Predicted Values / **Actual Values**	**Low income**	**Medium income**	**High income**
Low income	130	76	3
Medium income	96	616	139
High income	10	121	678

Based on the Random Forest's Confusion Matrix, the actual high-income class, which is correctly predicted by our model, is 678 records, with 121 records being wrongly predicted as the medium-income class and ten records being wrongly predicted as the low-income class. The actual medium-income class, which is correctly predicted by our model, is 616 records, with 139 records being wrongly predicted as the high-income class and 96 records being wrongly predicted as the low-income class. Lastly, the actual low-income class, which is correctly predicted by our model, is 130, with three records being wrongly predicted as the high-income class and 76 records being wrongly predicted as the medium-income class.

This means that the Random Forest model works well when predicting high-income and medium-income classes and works well when predicting low-income classes because around 38% of the actual low-income classes are misclassified as either high-income classes or medium-income classes. This is because there are imbalanced classes in the training set, where it has many records with high-income class and medium-income class (more than 2500 records) but only has a few records with low-income class (less than 700 records). Therefore, it is balanced up by using SMOTE technique instead of using actual data.

Random Forest will be used to predict the Testing Set. There are five most popular jobs based on how many times the job function occurred in the dataset, which can be seen in Table V.

TABLE V. JOB FUNCTION VALUE COUNTS

Rank	**Job Function**	**Value Count**
1	Penjualan / Pemasaran,Penjualan Ritel	196
2	Komputer/Teknologi Informasi,IT-Perangkat Lunak	178
3	Akuntansi / Keuangan,Akuntansi Umum / Pembiayaan	144
etc.	…	…

The Salary Level prediction result of one of the most popular jobs, which is 'Penjualan / Pemasaran, Penjualan Ritel,' can be seen in Table VI.

TABLE VI. PREDICTION RESULT – PENJUALAN / PEMASARAN,PENJUALAN RITEL

Career Level	**Location**	**Salary Level**
CEO/GM/Direktur/Manajer Senior [CEO/Director/Senior Manager]	DKI Jakarta	High income
	Sulawesi Utara	High income
Lulusan baru/Pengalaman kerja kurang dari 1 tahun [Fresh Graduates/Workers with less than 1 year experience]	Banten	Medium income
	DKI Jakarta	High income
	etc.	…
Pegawai (non-manajemen & non-supervisor) [Employee (non-management & non-supervisor)]	Bengkulu	High income
	DIY	Low income, Medium income
	DKI Jakarta	Medium income
	etc.	…

Supervisor/Koordinator [Supervisor/Coordinator]	Bali	High income
	DKI Jakarta	Medium income
	etc.	…

'Penjualan / Pemasaran, Penjualan Ritel' (Retail Sales) is the most popular job in the dataset, where each career level is available in various provinces. Most people with a career in this field get medium to high income. Only the new graduate career level has low income in some provinces, which makes sense because of their little experience, so companies tend to give low salaries. However, in some provinces, fresh graduates can have an average or even a high income. An employee in this field is most likely to get an average income, some might get high income, and some might get low income, like in some companies in DIY. Career levels higher than an employee, supervisor, and manager tend to get high income. However, some provinces provide an average income for the supervisor workers.

CONCLUSION

By classifying and predicting the salary range of jobs, it can be ascertained whether the salary is appropriate or not to be able to fulfill a person's life needs, moreover, whether it is under the field of work, work experience, or the location of work. This topic is helpful and essential for employers and employees out there to be able to estimate the expected salary for a particular position in a specific field in their location. The best model is from Random Forest, with an accuracy rate of 72% with the help of hyperparameters. Therefore, the best classifier is Random Forest based on the performance metrics. Furthermore, the top 3 most popular jobs are Penjualan Ritel (Retail Sales), IT-Perangkat Lunak (IT-Software), and Akuntansi Umum / Pembiayaan (Accounting). They mostly have a salary level ranging from medium income (average) to high income (high), most job locations are on Java Island (Especially at DKI Jakarta), and they have job positions/career levels primarily as employees.

The salary level predictions are determined by each province's UMP (minimum wage) with low-income, medium-income, and high-income levels. The job field, career level, and work location highly influence the predictions. Even though the job and career level are the same, locations also influence the salary level. For example, from Table V, it can be seen that Retail Sales workers with an employee's career level have a low salary in Jambi. However, employees will get a high salary in West Java (Jawa Barat) and an average salary in Central Java (Jawa Tengah). Therefore, this research still needs improvement so that the author can work with a more extensive dataset containing variables such as working age, credentials, and achievements.

REFERENCES

[1] "Economic Needs and Wants: Definition & Concept". September 2015. Retrieved from https://study.com/academy/lesson/economic-needs-and-wants-definition-lesson-quiz.html

[2] "Economic Need Definition". Retrieved from https://www.lawinsider.com/dictionary/economic-need.

[3] "Economic and Social Inclusion Corporation". 2008. Retrieved from https://www2.gnb.ca/content/gnb/en/departments/esic/overview/content/what_is_poverty.html

[4] Ranggasari, R. "Indonesia's Inflation Rate at 1.17% September; Highest in 9 Months". October 2022. Retrieved from https://en.tempo.co/read/1641075/indonesias-inflation-rate-at-1-17-september-highest-in-94-months.

[5] Jones, A. "Why Workers are Choosing Big Pay Packets Over Flexibility". May 2022. Retrieved from https://www.bbc.com/worklife/article/20220428-why-workers-are-choosing-big-pay-packets-over-flexibility

[6] "Eight Factors That Can Affect Your Pay". Retrieved from https://www.salary.com/articles/eight-factors-that-can-affect-your-pay/

[7] Wisneski, C. "7 Reasons Why Machine Learning Forecasting is Better Than Traditional Method". January 2022. Retrieved from https://www.akkio.com/post/5-reasons-why-machine-learning-forecasting-is-better-than-traditional-methods.

[8] Gopal, K., Singh, A., Kumar, H., & Sagar, S. "Salary Prediction Using Machine Learning," India: Galgotias University, vol. 8 Issue 1, June 2021, pp. 381-383.

[9] Lothe, D. M., Tiwari, P., Patil, N., Patil, S., & Patil, V. (2021). Salary Prediction using Machine Learning. *INTERNATIONAL JOURNAL*, *6*(5).

[10] Hansen, L. & Rennold, N. "7 Key Factors that are Holding You Back from Growing Your Salary". March 2022. Retrieved from https://www.businessinsider.com/7-factors-stopping-you-from-growing-your-salary-2022-3

[11] "What is Random Forest?". Retrieved from https://www.ibm.com/topics/random-forest

[12] Wibowo, C. P. "Job Description and Salary in Indonesia". March 2022. Retrieved from https://www.kaggle.com/datasets/canggih/jog-description-and-salary-in-indonesia

[13] Wijaya, C. Y. "5 SMOTE Techniques for Oversampling Your Imbalance Data". September 2020. Retrieved from https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5

[14] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. "SMOTE: synthetic minority over-sampling technique". Journal of artificial intelligence research, 16, June 2002, pp. 321-357.

[15] Aurélien, G. (2017). Hands-on machine learning with scikit-learn & tensorflow. Geron Aurelien. 134, 145-150, 191.

[16] Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11), 218. https://doi.org/10.21037/atm.2016.03.37

[17] (n.d.). *What is Decision Tree?*. IBM. https://www.ibm.com/id-en/topics/decision-trees.

[18] Shahane, S. (n.d.). *Voting Classifier*. Kaggle. https://www.kaggle.com/code/saurabhshahane/voting-classifier

[19] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics (2009)

[20] Raykar, V.C., Saha, A. (2015). Data Split Strategies For Evolving Predictive Models. In: Appice, A., Rodrigues, P., Santos Costa, V., Soares, C., Gama, J., Jorge, A. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2015. Lecture Notes in Computer Science(), vol 9284. Springer, Cham. https://doi.org/10.1007/978-3-319-23528-8_1

[21] Naveen. "What is Precision, Recall, Accuracy and F1-score?". February 2022. Retrieved from https://www.nomidl.com/machine-learning/what-is-precision-recall-accuracy-and-f1-score/

[22] Shung, K. P. "Accuracy, Precision, Recall or F1?". March 2018. Retrieved from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[23] Ij, H. (2018). Statistics versus machine learning. Nat Methods, 15(4), 233. https://www.nature.com/articles/nmeth.4642.pdf