Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 6, No. 4, November 2021, Pp. 345-350

Employee attrition and performance prediction using univariate ROC feature selection and random forest

Aris Nurhindarto*[1], Esa Wahyu Andriansyah[2], Farrikh Alzami[3], Purwanto[4], Moch Arief Soeleman[5], Dwi Puji Prabowo[6]
Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia[1,2,3,4,5,6]

Article Info
Keywords:
Employee Attrition and Performance, Feature Selection, Univariate ROC, Receiver Operating Characteristics Curve, Decision Tree, Random Forest

Article history:
Received: September 20, 2021
Accepted: November 30, 2021
Published: March 08, 2022

Cite:
Nurhindarto, A., Andriansyah, E. W., Alzami, F., Purwanto, P., Soeleman, M. A., & Prabowo, D. P. (2021). Employee Attrition and Performance Prediction using Univariate ROC feature selection and Random Forest. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 6(4). https://doi.org/10.22219/kinetik.v6i4.1345

*Corresponding author.
Aris Nurhindarto
E-mail address:
arisnurhindarto@dsn.dinus.ac.id

Abstract
Each company applies a contract extension to assess the performance of its employees. Employees with good performance in the company are entitled to future contracts within a certain period of time. In a pandemic time, many companies have made decisions to carry out WFH (Work from Home) activities even to Termination (Attrition) of Employment. The company's performance cannot be stable if in certain fields it does not meet the criteria required by the company. Thus, due to many things to consider in contract extension, we are proposed feature selection steps such as duplicate features, correlated features and Univariate Receiver Operating Characteristics curve (ROC) to reduce features from 35 to 21 Features. Then, after we obtained the best features, we applied into Decision Trees and Random Forest. By optimizing parameter selection using parameter grid, the research concluded that Random Forest with feature selection can predict Employee Attrition and Performance by obtain accuracy 79.16%, Recall 76% and Precision 82,6%. Thus with those result, we can conclude that we can obtain better prediction using 21 features for employee attrition and performance which help the higher management in making decisions.

1. Introduction

The rapid development of technology can increase company productivity, the merging of computers with telecommunications causes a revolution in the field of information systems, namely applications that can help a company. By utilizing the application, it can help the Human Resources Department (HRD) in evaluating the performance of each HR or employee easier and making decisions faster. Human Resources is the most vital component in an organization that will move and carry out activities to achieve goals by realizing the vision and mission of the company that has been set. Having qualified human resources can determine the success of a company or organization, and also make it easier for organizational leaders to direct them to achieve goals, besides that quality human resources can be expected to encourage the achievement of organizational competitive advantage [1]. Employee performance is often also referred to as work performance, which is the actual behavior that shows the results of the relationship between the efforts, abilities and perceptions of each employee on the work results or achievements achieved. Employee performance can be defined as the quality and quantity carried out by employees to carry out their duties in accordance with the assigned responsibilities. To find out the performance results of employees to what extent the level of professionalism of employees that has been achieved can develop in accordance with expectations or vice versa [2]. Each company applies a contract extension to assess the performance of its employees. Employees with good performance in the company are entitled to contracts in the future. The extension is carried out for a certain period of time, and only employees with good potential or performance are entitled to sign further contracts [2]. In terms of providing a contract extension, the company must evaluate the performance of all its employees. Due to the small frequency of face-to-face meetings between managers and employees, it is difficult to evaluate employee performance and it is difficult to determine the accuracy of the prediction level so that the company can obtain the best human resources [3].

Moreover, we need a prediction if the well performed employee will leave the work due to there is no more room for the employee to growth or there are better opportunity in the other companies [4]. Also, by knowing which employee who have less performance, could save us from financial loss [5]. Thus, by knowing these, we could prevent the employee attrition which will save us from: 1) find suitable replacements for employees, particularly those with high

experience and special skills; 2) takes time and efforts for new employees to achieve the same levels of expertise and productivity. 3) takes time, effort and money to recruit new employees [6][7].

Almost all industries use machine learning to improve work processes. Machine learning has evolved into a powerful function that can support a wide range of business solutions [8]. For example Machine learning is used for heart attack classification [9], sentiment analysis [10], tourist arrival [11] and so on.

In the dataset we are used, there are features as follows: age, Business Travel, Daily Rate, Department, Distance From Home, Education, Education Field, Employee Count, Employee Number, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, Marital Status, Monthly Income, Monthly Rate, Number Companies Worked, Over 18 years old, Take Overtime, Percent Salary Hike, Performance Rating, Relationship Satisfaction, Standard Hours, Stock Option Level, Total Working Years, Training Times Last Year, Work Life Balance, Years at Company, Years in Current Role, Years Since Last Promotion, Years With Current Manager and finaly, attrition as label. Those features are usually used in companies, several features perhaps should be updated due to COVID-19 Pandemic, but nevertheless, the features still practical in Work from Home scenario, due to many employee already mature and have experience in companies.

With many attributes that need to consider, feature selection is needed to obtain important features for calculate the performance of Empoyee. Good feature selection should be able to eliminating irrelevant, redundant, constant, duplicated, and correlated features. Here, we consider using Univariate Receiver Operating Characteristics curve (ROC) which uses machine learning models to measure the dependence of two variables. It's suitable for all variables, and also makes no assumptions about their distribution [12].

Univariate filter method is where individual features are graded according to certain criteria. Univariate filter methods are ideal for removing constant and quasi-constant features from data [13]. One of the disadvantages of this method is that it can select redundant features because the relationships between individual features are not taken into account when making decisions[14].

With the dataset obtained from Kaggle with title "IBM HR Analytics Employee Attrition & Performance", we are proposed using feature selection steps such as duplicate features, correlated features and Univariate Receiver Operating Characteristics curve (ROC) to reduce features from 35 to 21 Features. For the machine learning, we are compare using Random Forest and Decision Tree to find the suitable machine learning.

Decision Tree are chosen because: 1) Compared to other algorithms decision trees requires less effort for data preparation during pre-processing; 2) does not require normalization and scaling of data; 3) is very intuitive and easy to explain to technical teams as well as stakeholders [15]. Then, Random forests are chosen because: 1) reduces overfitting in decision trees; 2) works well with both categorical and continuous values; 3) automates missing values present in the data [16].

Thus, by optimizing parameter selection using parameter grid, the research concluded that Random Forest with feature selection can predict Employee Attrition and Performance by obtain accuracy 79.16%, Recall 76% and Precision 82,6%.

Our contribution in this research is: 1) using applied removed constant features; 2) removed duplicate features; 3) variance threshold to remove quasi-constant features; 4) random under sampling; 5) removed correlation features; and 6) Univariate ROC feature selection. From those steps, our research obtains higher performance than the model which only using Univariate ROC feature selection alone or not using feature selection.

For summary, these manuscripts are written in the following order: section 2 describe the research method, the results and discussion are listed in section 3, concluding and future research listed in section 4.

2. Research Method

This research stages can be seen in Figure 1 and described as follows: 1) preprocessing the data; 2) split the data into training data and testing data; 3) utilizing Univariate ROC feature selection to obtain reduced features; 4) Using 5 cross validations to reduce the randomness effect and obtain the best parameter for either random forest and decision tree in fitting module. Then, comparing the original data with reduced features data using random forest and decision tree to obtain prediction performance. The detailed research method can be explained in following subsection.

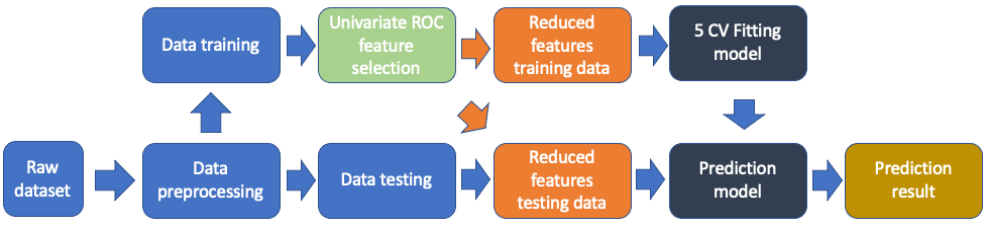

Figure 1. Research Method Overview

2.1 Obtaining raw dataset and Undersampling

The dataset we are using is IBM HR Analytics Employee Attrition & Performance. The dataset is a fictional data set which created by IBM data scientists to Uncover the factors that lead to employee attrition and explore important questions such as 'show me a breakdown of distance from home by job role and attrition' or 'compare average monthly income by education and attrition'. The dataset contains 1470 records, 34 features and 1 label. The dataset is belonged to imbalanced data, due to the attrition label contains 237 records it true and 1237 is false which can be seen at Figure 2. The dataset is obtained from https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset.

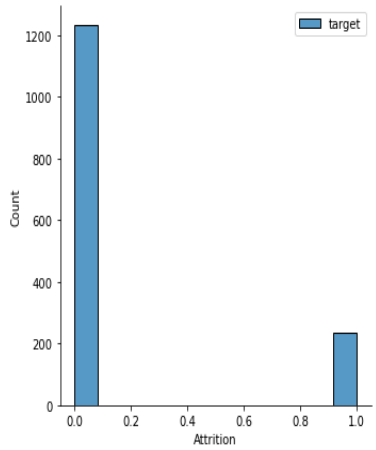

Figure 2. Attrition Label Distribution

As we can see, the dataset distribution is imbalanced due to real world nature, it can be proven that not many people have lower employee performance. Many machine learning cannot perform optimally if the dataset distribution is not balanced [17]. Thus, imbalanced data distribution should be normalized. There are several methods to tackle the imbalanced data distribution, such as: Random Over Sampling and Random Under Sampling. Random oversampling involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset. Random undersampling involves randomly selecting examples from the majority class and deleting them from the training dataset [18]. In this research, we split into training dataset and testing dataset by 90% and 10%. The reason we are using 90% dataset to training dataset because the dataset due to we are using random undersampling. Then, the random undersampling only applied to the training dataset due it is more convenient to obtain the balanced distribution.

2.2 Data Preprocessing – Cleaning data for training data

In this research, we used several procedures to clean the dataset. As worth mention, we used all the preprocessing data in the training data. after we obtain the pattern, we applied the preprocessing pattern into test data.

First, we transform the categorical features into ordinal features, such as: 1) we transform Business Travel into ordinal Travel_Rarely into one, and Travel_Frequently into two, else is zero; 2) Department Division; 3) Education Field; 4) Job Role; 5) Marital Status; 6) Over time; and 7) Gender.

Second, we applied removed constant features. This method can be achieved by check the standard deviation for each column, if found the standard deviation is equal zero, then remove the respected features, can be seen in Equation 1.

$$remove\ features\ if\ \sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} = 0 \tag{1}$$

Third, we removed duplicate features, after that, we are using variance threshold to remove quasi-constant features.in here, we set the threshold into 0.1 indicates 99% of observations approximately.

Fourth, we removed correlation features. Here we took 0.8 as threshold. We are using Equation 2 as follows.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{2}$$

2.3 Data Preprocessing – Univariate ROC feature selection

Univariate ROC feature selection using machine learning models to measure the dependence of two variabels. The procedure can be seen in Table 1.

Table 1. Univariate ROC Feature Selection Algorithm

Algorithm 1: Univariate ROC feature selection
Require: Empty list of ROC values Training data features Training data labels Testing data features Decision tree classifier
Ensure: Split the training data (training data features and labels) into ***training features set 1***, ***training labels set 1, testing features set 1*** and ***testing labels set 1*** For feature in training features: Fit ***training features set 1*** to Decision Tree classifier with ***training labels set 1*** Calculate the prediction probability as ***y_scored*** append ROC values with ROC AUC score using ***y_scored*** and ***testing labels set 1*** EndFor select features with ROC values > 0.54 assign selected features to training data features as **reduced training data** assign selected features to testing data features as **reduced testing data**
Output: - selected features - reduced training data - reduced testing data

2.4 Fitting Model

Here, we are using 5 cross validation and parameters grid to obtain best parameters and reduce the randomness effect of reduced training data. Decision tree [19] and Random forests [20] is used as main classifiers. The experiment parameters we are using are described as Table 2.

Table 2. Machine Learning Parameters Setting

Machine Learning	Parameters	Values
Decision Tree	Max depth	1,2, 3, …, 20
	Min sample leaf	1, 4, 8, … 100
	Min sample split	2, 3, 4, … 10
	Criterion	Gini, entropy
Random Forest	Max features	Auto, squre root
	Max depth	10, 20, 30, …, 110
	Min sample split	2,3,4, … 10
	Min sample leaf	2,3,4, … 10
	Bootstrap	True, False
	Criterion	Gini, entropy
	Number classifier	100

3. Results and Discussion

In our experiment, we conduct the experiment as follows: 1) we are using original dataset and applied into decision tree and Random Forest; 2) we are using ROC feature selection and applied into decision tree and Random Forest. Thus, we obtained optimized parameter using cross validation parameters grid and presented into Table 3. Then the final result of Model Performance presented into Table 4.

Table 3. Optimized Parameters

Decision Tree	Decision Tree ROC	Random Forest	Random Forest ROC
criterion: 'entropy' max_depth: 3	criterion: 'entropy' max_depth: 3	bootstrap: True criterion: 'entropy'	bootstrap: False criterion: 'entropy'

min_samples_leaf: 21 min_samples_split: 2	min_samples_leaf: 21 min_samples_split: 2	max_depth: 10.0 max_features: 'auto' min_samples_leaf: 7 min_samples_split: 2	max_depth: 10.0 max_features: 'auto' min_samples_leaf: 5 min_samples_split: 2

Table 4. Model Performance Results

Model	Accuracy	Recall	Precision
Decision Tree	66%	64%	69.56%
Decision Tree + ROC	68.75%	64%	72.72%
Decision Tree our model ROC	75%	64%	**84.21%**
Random Forest	72.91%	68%	77.27%
Random Forest + ROC	66.67%	**76%**	65.52%
Random Forest our model ROC	**79.16%**	**76%**	82.61%

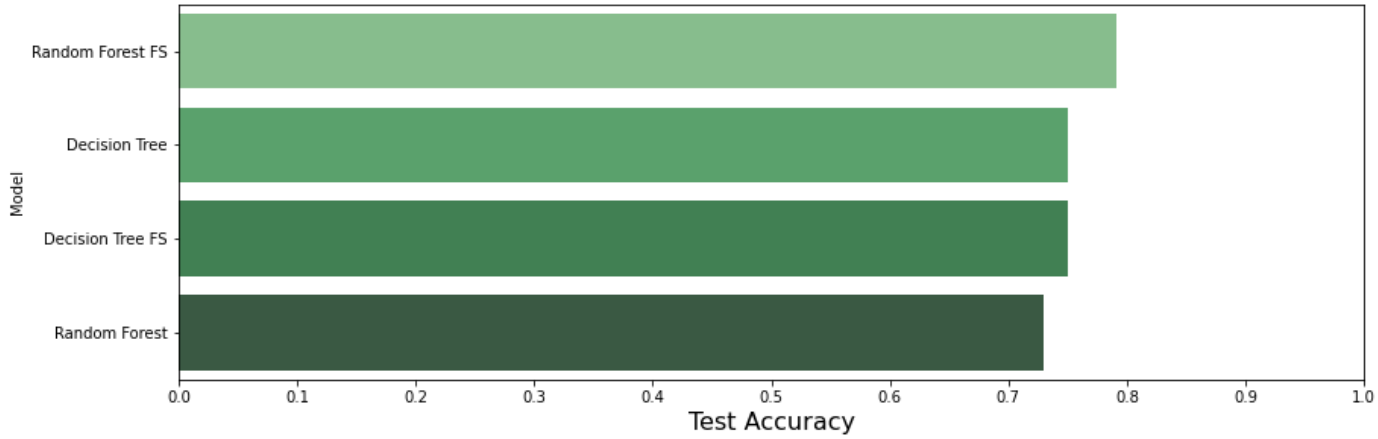

Figure 3. Model Accuracy Comparison

From Table 3 and Figure 3, we can see, the Univariate ROC feature selection is not helping the decision tree performance. but in Random Forest, together using applied removed constant features, removed duplicate features, using variance threshold to remove quasi-constant features removed correlation features with Univariate ROC feature selection is proven to help improve the random forest performance in Accuracy and Recall. the possible reason is that our methods can improve the Random Forest due to our methods prevents the random forest calculate the sparse data and omit unimportant features.

As remainder, this research using random undersampling, which is every class, have 237 records, total have 474 records. Thus, in next research, several considerations can be addressed, such as: 1) using random oversampling to obtain balanced data in training dataset; 2) using one-hot encoding to convert categorical values into numerical values. As worth mention, when we are using one hot encoding, the number of features will grow as much as features members. Then, by using one hot encoding, the dataset will become sparse. Thus, we need to consider machine learning which can handle sparse dataset.

Also worth mentioning, the IBM HR Analytics Employee Attrition & Performance is widely used dataset for practice the employee attrition. Here, we want to deliver our approach based on this dataset. Even though the dataset is syntethic dataset, it is served the purpose to gain the understanding of current event. The problem with dataset is it do not have the time series behaviour so our machine learning approach can be used only for yearly report, not on monthly based report for prediction.

4. Conclusion

This research presents the Employee Attrition and Performance Prediction using several cleaning processes such as: using applied removed constant features, removed duplicate features, variance threshold to remove quasi-constant features, random under sampling, removed correlation features, and Univariate ROC feature selection. From those steps, our research obtained higher performance than the model which only using Univariate ROC feature selection alone or not using feature selection in matter of accuracy and recall.

Cite: Nurhindarto, A., Andriansyah, E. W., Alzami, F., Purwanto, P., Soeleman, M. A., & Prabowo, D. P. (2021). Employee Attrition and Performance Prediction using Univariate ROC feature selection and Random Forest. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 6(4). https://doi.org/10.22219/kinetik.v6i4.1345

Because we are using Undersampling and found several findings in discussion section, future works that can be explored is using random oversampling method to balance the training data; using one hot encoding to convert categorical into numerical values; propose a method to create attrition dataset which contains time series behaviour so we can predict in different months in the future; Then, using machine learning that able to process the sparse data which resulted from one hot encoding methods.

Notation

The example of notation can be described with the following description:

Equation 1

σ : population of the features
μ : population means of features.
x_i : each value from the population of features.
N : size of population

Equation 2

r : correlation coefficient
x_i : values of the x-variable in a sample.
$\bar{x}$: mean of the values of the x-variable.
y_i : values of the y-variable in a sample
$\bar{y}$: mean of the values of the y-variable

Acknowledgement

Thank you to Program Doktor Ilmu Komputer Research Lab Universitas Dian Nuswantoro, which assist the authors in using Data Science Server.

References

[1] P. Altıok, “Applicable vision, mission and the effects of strategic management on crisis resolve,” *Procedia - Soc. Behav. Sci.*, vol. 24, pp. 61–71, 2011. https://doi.org/10.1016/j.sbspro.2011.09.057
[2] A. A. Davidescu, S.-A. Apostu, A. Paul, and I. Casuneanu, “Work Flexibility, Job Satisfaction, and Job Performance among Romanian Employees—Implications for Sustainable Human Resource Management,” *Sustainability*, vol. 12, no. 15, p. 6086, Jul. 2020. https://doi.org/10.3390/su12156086
[3] S. M. Hamidi, “Performance Appraisal and Its Effects on Employees Motivation: A Case Study of Afghan Wireless Communications in Kabul,” *SSRN Electron. J.*, 2019. https://dx.doi.org/10.2139/ssrn.3426851
[4] V. V. Saradhi and G. K. Palshikar, “Employee churn prediction,” *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1999–2006, Mar. 2011. https://doi.org/10.1016/j.eswa.2010.07.134
[5] V. Shapoval, “Organizational injustice and emotional labor in the hospitality industry: A theoretical review,” *Int. J. Hosp. Manag.*, vol. 83, pp. 56–64, Oct. 2019. https://doi.org/10.1016/j.ijhm.2019.04.002
[6] K. Haldorai, W. G. Kim, S. G. Pillai, T. (Eliot) Park, and K. Balasubramanian, “Factors affecting hotel employees’ attrition and turnover: Application of pull-push-mooring framework,” *Int. J. Hosp. Manag.*, vol. 83, pp. 46–55, Oct. 2019. https://doi.org/10.1016/j.ijhm.2019.04.003
[7] D. J. Madigan and L. E. Kim, “Towards an understanding of teacher attrition: A meta-analysis of burnout, job satisfaction, and teachers’ intentions to quit,” *Teach. Teach. Educ.*, vol. 105, p. 103425, Sep. 2021. https://doi.org/10.1016/j.tate.2021.103425
[8] A. Oussous, F.-Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 30, no. 4, pp. 431–448, Oct. 2018. https://doi.org/10.1016/j.jksuci.2017.06.001
[9] A. Purnomo, M. A. Barata, M. A. Soeleman, and F. Alzami, “Adding feature selection on Naïve Bayes to increase accuracy on classification heart attack disease,” *J. Phys. Conf. Ser.*, vol. 1511, p. 012001, Apr. 2020. https://doi.org/10.1088/1742-6596/1511/1/012001
[10] F. Alzami, E. D. Udayanti, D. P. Prabowo, and R. A. Megantara, “Document Preprocessing with TF-IDF to Improve the Polarity Classification Performance of Unstructured Sentiment Analysis,” *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, pp. 235–242, Aug. 2020. https://doi.org/10.22219/kinetik.v5i3.1066
[11] Purwanto, Sunardi, F. T. Julfia, and A. Paramananda, “Hybrid model of ARIMA-linear trend model for tourist arrivals prediction model in Surakarta City, Indonesia,” in *AIP Conference Proceedings*, 2019, vol. 2114, p. 060010. https://doi.org/10.1063/1.5112481
[12] P. X. Xiang Liu, “Feature Selection using Bootstrapped ROC Curves,” J. Proteomics Bioinform., vol. s9, 2014.
[13] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, “Benchmark for filter methods for feature selection in high-dimensional classification data,” *Comput. Stat. Data Anal.*, vol. 143, p. 106839, Mar. 2020. https://doi.org/10.1016/j.csda.2019.106839
[14] E. B. Nkemnole and O. Abass, “A t-distribution based particle filter for univariate and multivariate stochastic volatility models,” *J. Niger. Math. Soc.*, vol. 34, no. 2, pp. 227–242, Aug. 2015. http://dx.doi.org/10.1016%2Fj.jnnms.2014.11.002
[15] A. De Caigny, K. Coussement, and K. W. De Bock, “A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees,” *Eur. J. Oper. Res.*, vol. 269, no. 2, pp. 760–772, Sep. 2018. https://doi.org/10.1016/j.ejor.2018.02.009
[16] M. C. E. Simsekler, A. Qazi, M. A. Alalami, S. Ellahham, and A. Ozonoff, “Evaluation of patient safety culture using a random forest algorithm,” *Reliab. Eng. Syst. Saf.*, vol. 204, p. 107186, Dec. 2020. https://doi.org/10.1016/j.ress.2020.107186
[17] M. Kuhn and K. Johnson, *Applied predictive modeling*. New York, NY: Springer New York, 2013.
[18] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018.
[19] S. L. Salzberg, “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993,” *Mach. Learn.*, vol. 16, no. 3, pp. 235–240, Sep. 1994. https://doi.org/10.1007/BF00993309
[20] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324