Jurnal CoreIT, Vol.8, No.1, June 2022
ISSN 2460-738X (Print)
ISSN 2599-3321 (Online)

Application of Predictive Analytics To Improve The Hiring Process In A Telecommunications Company

Luh Putu Saraswati Devia Jayanti[1]*, Meditya Wasesa[2]
[1,2]Master of Business Administration, School of Business and Management,
Institut Teknologi Bandung, Indonesia
luhputusaraswati_deviajayanti@sbm-itb.ac.id (*Corresponding Author), meditya.wasesa@sbm-itb.ac.id

Abstract. Industry 4.0 refers to the increasing tendency towards automation and data exchange in technologies like Big Data and AI. The existence of technology means telecommunication companies have to adapt. Therefore, it takes great people so that the company can continue to survive. The problem that companies often face in hiring great people is that it costs a lot and takes a long time to recruit. Predictive analysis can assist in identifying system issues and solutions. This study aims to develop predictive analytics that can improve recruitment screening based on CVs and find the best predictive model for the company to reduce costs and long recruitment cycles using technology. The authors built an analytical prediction model in four stages: data collection, data preprocessing, model building, and model evaluation. This technique uses Random Forest and Naive Bayes classification algorithms. Both systems properly predicted more data sets with 70% accuracy, 70% precision, and a recall rate above 80%. Compared between the two techniques, Random Forest outperforms Naive Bayes for this predictive model. A lot of people are talking about predictive analytics for hiring, but there aren't many data mining frameworks that can help to find rules based on the CVs of people who have worked for companies before.

Keywords: HR Analytic, Human Resource, Naïve Bayes, Predictive Analytic, Random Forest, Recruitment

Received April 2022 / **Revised** June 2022 / **Accepted** June 2022

INTRODUCTION

The Fourth Industrial Revolution (or Digital Era 4.0) is now underway. Industrial Revolution 4.0 "Business 4.0" refers to the rising automation and data interchange in industrial technology and operations. Industry 4.0, Cyberphysical Systems, Artificial Intelligence, Smart Factories (Big Data), Cloud Computing, and other technologies will help manufacturers and businesses. Nowadays, everyone must be tech-savvy. New devices and technology are revolutionizing telecom. Telecommunications is a fast-growing industry. The telecom business must adapt, change, and grow to remain competitive. Indonesia's telecom industry is vital. The Telecommunications Company's strategy and ideas continue to serve Indonesia. The deployment of sustainable and accessible infrastructure for all Indonesians has also increased. Because of this, the corporation will give priority to employing top digital talent to help boost the nation's digital capabilities and increase digital adoption.

To find the best candidate, telecommunications companies must also think about the cost and time so that the process runs effectively. In [1], According to respondents (19%), the cost (budget) issue became the third most significant factor affecting recruitment circumstances. According to respondents' explanations, most enterprises in Indonesia are currently (at the time of this research) working on reducing costs. According to [2], If recruiters do not pay enough attention to specific details, it may result in unrealistic job analysis and long recruitment cycles. This makes it harder to find outstanding talent and increases costs in terms of money, labor, and opportunity. In [3], screening eligible candidates can consume a significant amount of time and resources, which is exacerbated when a large number of applicants respond.

In addition to identifying potential and matching it to an organization's needs, advanced algorithms can also locate team players based on core traits and personality matching, making it an effective way to avoid the need for costly and time-consuming preliminary screening [4]. HR analytics can be used for predictive purposes. Predictive analytics uses sophisticated methods like machine learning to forecast future events. It is the use of historical data to forecast future recruitment strategies, hiring decisions, and workforce planning [5]. Predictive algorithms such as regression, decision trees, random forests, and Bayesian statistics are widely used nowadays.

DOI: 10.24014/coreit.v8i1.16915

Previous studies have explored using predictive analytics for the recruitment process. In [6], The research predicts that joining efficient candidates on before resume selection and the total process is to be done in an efficient way with minimal cost and minimal time. In [7], Naïve Bayes is used in the next step to determine or predict employee placement based on their characteristics. In [8], the authors proposed system aims to analyze the performance and possible suitable candidates for the job using the random forest method. According to [9], the authors set up a job predictor based on the candidate's resume.

There are many articles exploring predictive analytics for recruitment, but there are few data mining frameworks for extracting rules based on historical CVs of company applicants with limited data. Thus, this study aims to develop predictive analytics that can improve recruitment screening based on CVs and find the best predictive model for the company to reduce costs and long recruitment cycles using technology. The benefit is that organizations can find the right people the first time, which leads to better results.

METHODS

A. Research Framework

In order to define the solutions, a research framework must first be developed. This research method is modified based on research [10] to build predictive models from data collection to evaluation. The research framework is shown in Figure 1. The research includes four major processes that are included in this research: data collection, preprocessing of data, model building, and the evaluation of the model. Using secondary data from the employer over a four-year period, the candidate's CV is compiled. It is possible to extract valuable and representative information from a dataset by preprocessing it. A prediction model is then developed, and finally an evaluation is performed on it. This model will be built with the RapidMiner app.

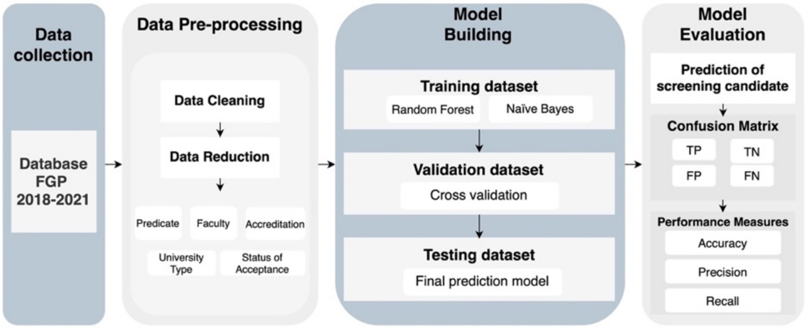


Figure 1. Research framework

B. Model Algorithm

This study used random forest and Naïve Bayes models. The two models will be compared to see which is more predictive.

Table 1. Prediction models choice

Classifiers Category	Classifiers by Groups	Model
Ensemble	Bagging	Random Forest
Singular	Probability	Naïve Bayes

1. Random Forest

Random forest is a sort of classifier that is often used in both regression and classification classifiers in supervised learning. Additionally, the RF classifier is an ensemble classifier, which means it creates predictions by integrating many CARTs [11].

DOI: 10.24014/coreit.v8i1.16915

2. Naïve Bayes

Based on the simple probability Bayes theorem, the Naïve Bayes classifier employs a small amount of training data to estimate the mean and variance of a variety of variables. This algorithm can be expressed formally as equation 1. According to another definition, the Naïve Bayes classifier is a classification system that employs a probability method and statistics that were developed by British scientist Thomas Bayes, and it forecasts future opportunities based on previous experiences [12].

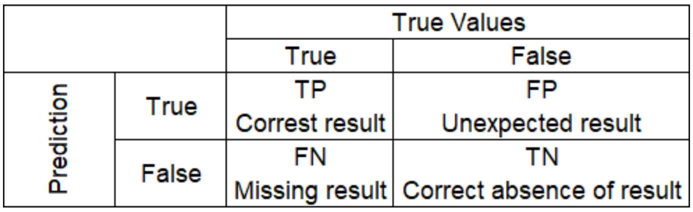
$$P(H\,\big|\,\boldsymbol{E}) = \ (P(E|H)P(\boldsymbol{H}))/(P\ (\boldsymbol{E})) \tag{1}$$

Description:
P(H|E) = Probability of being in a certain place (conditional probability)
P(E|H) = Probability of occurrence of parameter E under hypothesis H
P(H) = Hypothesis H is the prior probability (prior) hypothesis.
P(E) = E specifies the initial probability (previous) parameter.

C. Model Evaluation

1. Confusion Matrix

Confusion matrix are a machine learning tool widely used to test or show the behavior of models in supervised classification settings, such as classification exercises. This method uses a square matrix, where each row represents an individual instance's actual class, and each column represents a predicted class. These columns and rows in this matrix include all of the basic information regarding predictions generated by a classification model after being applied to a given dataset. It is common practice to assess a model's generalization accuracy with data not utilized during the model's learning phase [13].



		True Values	
		True	False
Prediction	True	TP Correst result	FP Unexpected result
	False	FN Missing result	TN Correct absence of result

Figure 2. Classification confusion matrix

2. Performance Measures

Confusion matrix for making a variety of performance indicators to see how well the performance model that was made works. Performance measures like accuracy, precision, and recall were used in this experiment to see how the results worked out. Precision data is data that is based on not having enough information. If you use a binary classification algorithm, as the one shown, precision in binary classification can be the same as positive predictive values. Deleted information that can be found and used again is called "recall data." It comes from data that is relevant to the query. The term "sensitivity" refers to the ability to remember information in a two-step process. With recall, it is possible to see how the relevant data taken was set up and approved by the query and how it was taken. Accuracy is a percentage of the total data that has been found and looked at [14].

$$Presicision = \left(\frac{TP}{TP+FP}\right) X\ 100\%. \tag{2}$$

$$Recall = \left(\frac{TP}{TP+FN}\right) X\ 100\%. \tag{3}$$

$$Accuracy = ((TP + TN)/(TP + TN + FP + FN))X\ 100\% \tag{4}$$

RESULT AND DISCUSSION

1. Data Collection

This company collects data for research purposes using secondary sources, specifically the CV dataset from 2018 to 2021. This data includes basic demographic information (ID number, gender, and age), education background (degree level, major field of study, and college GPA), English language skills (score), work experience, chosen field of labor, and admission status. It contains 125,309 candidates.

2. Data Pre-processing

Pre-processing converts unusable raw data to a more useful and efficient format. The outcome of this pre-processing data will be a dataset that is ready for machine learning processing. Pre-processing is the process of cleaning and reducing data.

The dataset was cleansed in this study by correcting any inaccurate data collected, such as typos. During data cleansing, missing values are also deleted. Missing values are prevalent in datasets. A missing value in a dataset is a characteristic or variable that lacks data in a sample of observations. Before processing the dataset, the missing value must be removed. The dataset utilized in this study has several missing values for each attribute. Relevant data lacks necessary information, as in the following examples:

1. The data appears to be blank or lacking attribute data (GPA, Department, University, and so on)
2. unclear data
3. D4/S1 schooling is below average.
4. Non-GPA data, specifically (1-4).

In this study, missing values are removed from the data by removing them from the data set. After removing the missing value from the data, outlier identification is performed. The data in this study will be discarded if machine learning detects outliers. After removing any missing values, the dataset is reduced.

The company's dataset for this study contains information on basic demographics (ID number, gender, and age), educational attainment (degree level, major field of study, and GPA), English language ability (score), work experience, chosen field of work, and acceptance status. On the other hand, this strategy entails reducing the number of dimensions or utilizing only the necessary data rather than the entire set of data. GPA, major, university, and candidate admission status are all variables considered. It is superfluous to include information about a candidate's English language proficiency and work experience, as the business may not take these things into consideration. These are GPA values that have been simplified to predicate forms (summa cum laude, magna cum laude, and cum laude) to correspond to the shapes of the other qualities. To make machine learning easier to process, major field data is converted to faculty data. The data on universities is classified into two categories: type and accreditation. Universities can be evaluated through accreditation. The higher the accreditation value of an institution, the more qualified it is. Since Indonesia has both private and public universities, the university data type is used. Additionally, candidate acceptance status data is streamlined from passing, not passing, and not passing to enter the final step to yes or no data. This pre-processing converts raw data to machine-learning-suitable data. The output from Tables 2 and 3 will be descriptive statistics about the datasets and variables.

Table 2. Choice of variable

Category	Variable/Data Types	Variable Description
GPA	Predicate / Ordinal	Labels based on established criteria for student graduation
Major of Field	Faculty / Nominal	A university's academic program or major
University	Type of University / Nominal	Indonesian Higher Education Types
	Accreditation / Ordinal	The evaluation of an educational institution using predetermined criteria

Table 3. Descriptive statistics of the datasets

Data	Attribute	Statistics	Raw Data	Filtered Data
Basic Information	ID Number	Count	125309	123788

DOI: 10.24014/coreit.v8i1.16915

	Birth of date	Count	125309	123788
	Gender	Count	125309	123788
	GPA	Count	125309	123788
	Degree level	Count	125309	123788
Educational Background	Major of Field	Count	125309	123788
	University	Count	125309	123788
	Type of English test	Count	125309	123788
English Profiency	Score	Count	125309	123788
	Work Experience	Count	125309	123788
Work Experience	Field of Work	Count	125309	123788
Field of Work	Acceptance Status	Count	125309	123788
Candidate acceptance status				

3. Model Building

At this stage, a data model is required for training, validating, and testing machine learning algorithms. By extracting critical information from datasets, data training teaches models or generates predictions. Because the data set already has a label, this study employs supervised machine learning. This data training technique makes use of 902 yes and 902 no records from the data period 2018-2020. The yes data for 902 comes from previously collected data, whereas the no data comes from machine learning data sampling. After that, the model will be trained on the specified attributes. To train a machine learning model, this data set will be combined with the goal attribute acceptance status (YES/NO) in order to determine whether or not a candidate screening will pass.

Validation of the dataset is required following training. This validation ensures that the model is deployable and fits well, avoiding over- or under-fitting. At this stage, cross-validation is used to evaluate the learning model's statistical performance. They divide the training data into ten equal segments and then repeat the training process on each segment to ensure that the model produces identical results.

Data testing is a subset of model testing that involves simulating new datasets to determine how well the model performs on them. In 2021, 306 data have a yes status, while another 306 have a no status. According to the number of findings, data sampling using machine learning yielded 306 data from the current data collection.

4. Model Evaluation

After data processing is complete, the results of running machine learning algorithms (Table 4) can be seen. As predicted with static data, the highest value is obtained. In Table 5, the results of the evaluation using the confusion matrix on two tested models are shown.

Table 4. Top value in prediction model

Category	Top Value
Accreditation	A
Faculty	Engineering
Predicate	Magna Cum Laude
University Type	State

Table 5. Confusion matrix result

	Random Forest	
	True NO	**True YES**
Pred. NO	79	28
Pred. YES	59	140
	Naive bayes	
	True NO	**True YES**
Pred. NO	81	32
Pred. YES	57	136

Table 6. Predictable amount of data

Prediction Model	Description	Amount of Data	Accuracy
Random Forest	The amount of data that was successfully predicted correctly	219	71.5%
	The amount of data that failed to predict correctly	87	
Naïve Bayes	The amount of data that was successfully predicted correctly	217	70.9%
	The amount of data that failed to predict correctly	89	

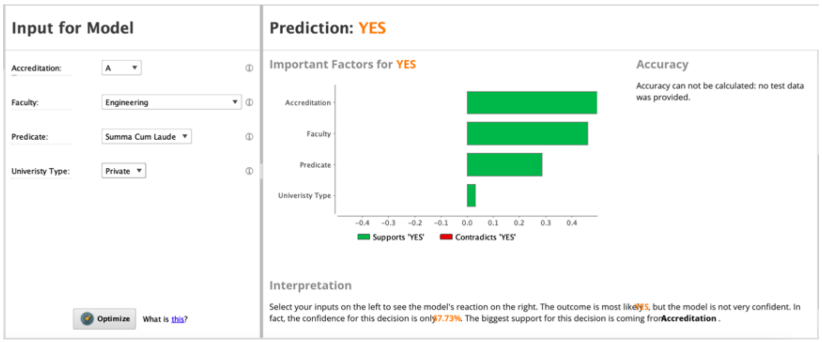
Table 7. Performance measures result

Model	Value ± SD		
	Accuracy	Precision	Recall
Random Forest	71.5% ± 3.9%	70.3% ± 6.4%	83.4% ± 6.3%
Naïve Bayes	70.9% ± 1.8%	70.6% ± 4.6%	81.4% ± 6.7%

Based on the confusion matrix and performance measures, the random forest classifier is a good model to deploy. As a result, when using pre-existing features and labels, random forest outperforms naive bayes. 219 data items were successfully predicted using the four criteria established at the outset of the process. This model is 71.5 percent accurate, according to the results. This reveals that the final model is capable of accurately categorizing 70% of things. Precision measures the difference between expected and observed outcomes, whereas recall measures the model's ability to adapt to new data. Precision is not synonymous with recall. According to the study's findings, the model's recall is greater than its precision. When more information is given to the model, it can predict the results with a success rate of up to 70%. According to [15] the authors stated that by adding many attributes that are considered in the calculation, it can predict more prediction accuracy.

The company simply needs to add new criteria to the registration process, such as faculty, predicate, university type, and university accreditation, and then execute the model. Figure 3. It displays an easy-to-complete candidate data entry form. Based on their registration information, this model can predict whether applicants will pass or fail. Additionally, this simulator can display variables that influence forecast results. Because they are based on past performance, they can assist businesses in selecting more qualified candidates in the future. This is a more effective method because it allows candidates with similar probability requirements to pass. This enables businesses to save money and time with a relatively small dataset.

DOI: 10.24014/coreit.v8i1.16915



Input for Model

Accreditation: A

Faculty: Engineering

Predicate: Summa Cum Laude

Univeristy Type: Private

Prediction: YES

Accuracy

Accuracy can not be calculated: no test data was provided.

Interpretation

Select your inputs on the left to see the model's reaction on the right. The outcome is most likely YES, but the model is not very confident. In fact, the confidence for this decision is only 67.73%. The biggest support for this decision is coming from **Accreditation** .

Figure 3. Model simulator prediction

Additionally, the recruitment process can be cost-effective when predictive analytics is used. This is because, following registration, not all candidates can advance to the selection stage simply by passing the screening stage using a prediction model. As a result, fewer candidates are accepted, and associated expenditures are also lowered. Additionally, by reducing the number of candidates, the organization avoids the expense of hiring additional staff to conduct this recruitment procedure. Along with cost, the company's time to hire is a factor. If the recruiter reduces the number of candidates obtained as a result of the screening process, the recruiter can also minimize recruiting time, particularly if the recruiter sorts the candidate from stage to stage, which takes time.

CONCLUSION

In this study, companies can use the HR analytical system using predictive analytics. This predictive analytics can identify possible future outcomes based on historical data. This can be done in four stages: data collection, data preprocessing (cleaning and reduction), model building, and evaluation. Algorithm models that can be used are Random Forest and Naïve Bayes. Based on the evaluation results, the Random Forest model managed to predict more correctly. The resulting performance measure has an accuracy of 70%, a precision of 70%, and a recall of 83%. So that these predictive analytics can generate new criteria that are appropriate based on historical data of candidates who have been accepted as employees to screen candidate recruitment at the company. This can help the company to get a candidate with a high probability of meeting the criteria needed by the company to be processed again to the next stage. This system can help to speed up the recruitment process and decrease company costs because there is already a screening stage at the beginning because not all candidates who apply can be accepted like in the manual recruitment process system.

REFERENCES

[1] E. P. Wiroko, “Tantangan dan Strategi Rekrutmen di Indonesia,” *psy*, vol. 4, no. 2, pp. 193–204, Dec. 2017, doi: 10.15575/psy.v4i2.1442.

[2] S. D. Rozario, S. Venkatraman, and A. Abbas, “Challenges in Recruitment and Selection Process: An Empirical Study,” *Challenges*, vol. 10, no. 2, p. 35, Aug. 2019, doi: 10.3390/challe10020035.

[3] U. C. Okolie and I. E. Irabor, “E-Recruitment: Practices, Opportunities and Challenges,” *European Journal of Business and Management*, p. 7, 2017.

[4] G. Walford-Wright and W. Scott-Jackson, “Talent Rising; people analytics and technology driving talent acquisition strategy,” *SHR*, vol. 17, no. 5, pp. 226–233, Oct. 2018, doi: 10.1108/SHR-08-2018-0071.

[5] V. Kumar and M. L., “Predictive Analytics: A Review of Trends and Techniques,” *IJCA*, vol. 182, no. 1, pp. 31–37, Jul. 2018, doi: 10.5120/ijca2018917434.

[6] D. Jagan Mohan Reddy, S. Regella, and S. R. Seelam, “Recruitment Prediction using Machine Learning,” in *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, Patna, India, Oct. 2020, pp. 1–4. doi: 10.1109/ICCCS49678.2020.9276955.

[7] F. A. Bachtiar, F. Pradana, and R. D. Yudiari, “Employee Recruitment Recommendation Using Profile Matching and Naïve Bayes,” in *2019 International Conference on Sustainable Information Engineering and Technology (SIET)*, Sep. 2019, pp. 94–99. doi: 10.1109/SIET48054.2019.8985988.

[8] S. Gupta, A. Hingwala, Y. Haryan, and S. Gharat, “Recruitment System with Placement Prediction,” in *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, Mar. 2021, pp. 669–673. doi: 10.1109/ICAIS50930.2021.9395768.

[9] H. Chaudhari, N. Yadav, and Y. Shukla, “A predictive analysis on job recruitment,” in *International Conference on Recent Trends in Engineering, Science & Technology - (ICRTEST 2016)*, Hyderabad, India, 2016, p. 6 (5 .)-6 (5 .). doi: 10.1049/cp.2016.1474.

[10] D. T. Andariesta and M. Wasesa, “Machine learning models for predicting international tourist arrivals in Indonesia during the COVID-19 pandemic: a multisource Internet data approach,” *JTF*, Jan. 2022, doi: 10.1108/JTF-10-2021-0239.

[11] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[12] I. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, “Implementation of Data Mining To Predict Period of Students Study Using Naive Bayes Algorithm,” *IJEET*, vol. 2, no. 1, p. 53, Sep. 2017, doi: 10.24843/IJEET.2017.v02.i01.p11.

[13] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognition*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

[14] F. Rahmad, Y. Suryanto, and K. Ramli, “Performance Comparison of Anti-Spam Technology Using Confusion Matrix Classification,” *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 879, no. 1, p. 012076, Jul. 2020, doi: 10.1088/1757-899X/879/1/012076.

[15] A. A. Mahmoud, T. AL Shawabkeh, W. A. Salameh, and I. Al Amro, “Performance Predicting in Hiring Process and Performance Appraisals Using Machine Learning,” in *2019 10th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, Jun. 2019, pp. 110–115. doi: 10.1109/IACS.2019.8809154.