ISSN 2278-3091
Volume 8, No.6, November – December 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse59862019.pdf
https://doi.org/10.30534/ijatcse/2019/59862019

Employee Performance Prediction using Naïve Bayes

Riyanto Jayadi[1], Hafizh M. Firmantyo[2], Muhammad T. J. Dzaka[3], Muhammad F. Suaidy[4], Alfitra M. Putra[5]
Information Systems Management Department, BINUS Graduate Program - Master of Information Systems Management,
Bina Nusantara University
Jakarta, Indonesia 11480.
[1]riyanto.jayadi@binus.edu, [2]hafizh.firmantyo@binus.ac.id, [3]muhammad.dzaka@binus.ac.id,
[4]muhammad.suaidy@binus.ac.id, [5]alfitra.putra@binus.ac.id

ABSTRACT

Employee performance drives company success to achieve its goals. Predicting employee performance in the future is a necessity for companies to success. This paper presented employee performance prediction in a company using machine learning. The machine learning process follows Cross-industry standard process for data mining (CRIPS-DM). The Naive Bayes classification method is employed to create the prediction model. The result shows that Naïve Bayes successfully correctly classified instances as high as 95.48 percent.

Key words : Human Resources, Employee Performance, Naive Bayes

1 INTRODUCTION

For all organizations around the world, the turnover rate of employees has gradually increased and become one of the significant issues for the organization. This turnover problem consume cost for the company including the cost severance pay, cost of hiring new employee cost and training them [1].Companies, institution and organizations are paying attention to how to reduce these turnover as it cost a lot of resource for them. The first step to reduce this turnover is identifying which employee will resign.

One major reason of employee's reassignment is when the company failed provide enough satisfaction. One of main drive in employee performance is their satisfaction. According to [2] , the employee's satisfaction correlates to the capability or ability of employee's performance for each task is accomplished by the employees would be achieved. Each employee's performance boosts their company competitive advantage. Human resource (HR) in each company are tasked to assure their employee's satisfaction so that it can provide good performance. Each individual goals of their employee are part companies goals to achieve companies objectives and vision [3]. A successful company are companies that can identify their employee performance.

Another way to predict employee resign is by monitoring their performance is using key performance indicator (KPI). KPI is a necessary tool to measure and determine the success of the company. Company performance toward its visions and how well it executed missions and strategies can be monitored using KPIs. These KPIs can be broken down and assigned to every divisions and each employee in the company. Each employee KPIs performance is part of their organizational success. These KPIs can provide evidence on how far each employee and the company reaching their objectives [4]. Periodically, every stakeholders and employees in a company have to have agreement regarding their strategies and vision interpretation and their KPI measurement[5].

All employees should not be treated the same as others because every employee has a different performance regarding where the employee is positioned in the organization. Employee performance drives company success to achieve its goals. Predicting employee performance in the future is a necessity for companies to success [2].

Based on these key performance indicator and satisfaction, the employee performance can be predicted. Thus, the employee turnover also can be predicted, to make sure that employee turnover prediction has a high accuracy so using machine learning can help the organization. Every machine learning had their algorithms and techniques. This kind of prediction fall into classification task in machine learning. Several previous have demonstrated machine learning application in many business area including human resource[1][6][7][8][9][10][11].

Classification is a technique of analyzing data and to predict which class a data belongs. Classification is one of

technique[12]. Two steps generally exist during classification process. The first step is learning process where the machine learning algorithm learn and train itself from the data, try to find pattern and generate model. Then, at the second step, the algorithm predict a new instance of occurrence based the learned model. [13]. Two most well-known classification algorithm are decision tree and naïve bayes.

Decision is the most common and most understandable technique in classification. It uses information entropy and information gain to learn and train from a given dataset. Then, it automatically gives some weights to attributes that being used, which is called information gain. The most influential variables, which has the highest information entropy, are identified from these process [14].

Meanwhile, naive bayes is one of the purest forms of bayesian type. This algorithm based on that all the independent attributes are appointed from the value of the conditional independence. The naïve bayes algorithm computes its learning model from the set of conditional independences and its frequency from the dataset [15]. Bayes proportion is used in this algorithm. It regard each from variable. Naïve bayes algorithm is well-known in the area of text mining as a good algorithm to solve classification problems [16][17].

In this paper, a classification methodology is proposed to predict the performance of employees using machine learning. The machine learning process follows the CRIPS-DM methodology. Naïve bayes classification algoritm is employed to predict the employee's performance. From the prediction, the company may decide which employees are deserved to behold or not.

In this paper, we proposed a naïve bayes based prediction methodology classification is proposed to predict the performance of employees using machine learning. The machine learning process follows the CRIPS-DM methodology. The goal is to predict the employee's performance. From the prediction, the company may decide which employees are deserved to behold or not.

2 PREVIOUS WORKS

Several previous have demonstrated machine learning application in many business area including human resource[1][6][7][8][9][10][11] and financial[18][19]. Previously, research on employee performance was carried out based [1]. Howerver, where the paper uses the reliability test method using Cronbach's alpha while in this paper uses Naive Bayes classification, and also the difference in the number of datasets and the number of attributes.

The study in [6] demonstrates that the XGBoost classifier had high accuracy for predicting employee turnover. The formulation makes it a reliable technique. It capable to handle the noise from the data. Meanwhile, the study in [8] suggests that at every organizational level at least have the actual turnover rate and turnover intention rate and then report to the leader of the organization at each level.

In [7], they investigated how effective the performance appraisal system is and its effect on employee motivation at work. The main goal is to establish the appropriate role of performance appraisal as a motivational tool and as a potential challenge in the future. His findings show excellent results when organizations use performance appraisal as a motivational tool for employees. Meanwhile, the research in [9] shows that using not only one valuation technique, but more than that, helped produce better satisfaction and higher motivation. Key aspects of the performance appraisal system (PAS) that can help increase motivation, including the relationship between work and appreciation; use PAS to be able to help in setting goals and benchmarks; and use of CL to be able to identify the strengths and weaknesses of employees in organization.

The authors of [10] experiments and simulated human resource data set explaining organizations from small size, medium-sized, and large size employee population handle by 1. Decision tree method, 2. Naïve bayes method, 3. Random forest method, 4. Logistic regression method, 5. Support vector machines, 6. Neural networks, 7. An extreme gradient boosting method, 8. A gradient boosting tree method, 9. Linear discriminant analysis, 10. K-nearest neighbor method. From a complete and robust evaluation process, the use of all of these methods to predict employee turnover is being analyzed or established using statistical methods. Another previous work by [11] presented a method to solve employee turnover problems using machine learning techniques. Appropriate predictions allow organizations to make decisions for employee succession planning. However, the data needed in this modeling problem comes from Human Resource Information Systems (HRIS) and is usually under-funded compared to other divisions in the organization, which directly relate to its priorities for organization. Then guide to the prevalence of noise in the data predictive models that tend to be over-fitting and inaccurate.

3 METHODOLOGY

This research follows CRIPS-DM methodology. CRIPS-DM offers standardized steps to conduct machine learning project. It is a well-know methodology in many industry and can be used with different tools and techniques. It has six phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment[20]. In the business understanding step, business objectives, planning, success criteria, risk assessment, cost-benefit of the machine learning task is studied extensively. In this research, business objectives is to reduce cost turnover rate of employee.

With many information linked with this issue, also with the targeted company, it would have required some experience to apply machine learning. Typically it referred for stride over to the earlier stages, and it can improve the

overall quality of the datasets to balance the required model that has been selected.Data understanding explore the data definition and its quality. Data cleaning for preparing data to be more easy, accurate and representative is done in data preparation. Meanwhile, modelling consists the activity of model training, selection, setting parameter, and model selection itself. Lastly evaluation is to assest the model accuracy and other performance metrics and whetever the model has reach success requriment.. In this study, the next subsection explain data understanding, data preparation, modelling and evaluation. We did not deploy model. This study is be finished in the evaluation phase of CRISP-DM.

3.1 Data Understanding

The data from [21] is used as dataset in this study. There is 310 employee data in the dataset. The attributes in the dataset including geneder, marital status, employee status, performance score, position and termination status. There are 28 variables in the dataset. Table 1 shows the attributes and its description.

Table 1: Meta Data

Attribute	Type	Description
Employee Name	Nominal	Containing varchar of employee name
Employee Number	Numeric	The employee number generated by the system
MarriedID	Numeric	Referral number of married or not
MaritalStatusID	Numeric	Referral number of Marital Status
GenderID	Numeric	Referral number of employee gender
EmptStatus_ID	Numeric	Referral number of employment status
DeptID	Numeric	Referral number of the department where employee stationed
Perf_ScoreID	Numeric	Referral number of current employee performance
Age	Numeric	Containing number of employee current age
Pay Rate	Numeric	Containing number of employee current salary
State	Nominal	Containing varchar of which state employee does live
Zip	Nominal	Containing varchar of employee current zip code
DOB	Nominal	Containing varchar of employee birthdate
Sex	Nominal	Containing varchar of employee sex type
MaritalDesc	Nominal	Containing varchar of employee marital condition desc
CitizenDesc	Nominal	Containing varchar of employee citizenship description
Hispanic/Latino	Nominal	Containing varchar of employee race type
RaceDesc	Nominal	Containing varchar of employee race description
Date of Hire	Nominal	Containing varchar of date of an employee is hired
Days Employed	Numeric	Containing number of days current employment
Date of Termination	Nominal	Containing varchar of the termination date
Reason for Term	Nominal	Containing varchar of employee termination description
Employment Status	Nominal	Containing varchar of current employment status
Department	Nominal	Containing varchar of the current department which employee stationed
Position	Nominal	Containing varchar of current employee position in the office
Manager Name	Nominal	Containing varchar of current employee supervisor
Employee Source	Nominal	Containing varchar of a company which employee is hired from
Performance Score	Nominal	Containing varchar of employee current performance score

3.2 Data Preparation

In this study, the data is prepared as the following. The target variable is regrouped from seven variable into two variable in order to avoid the tree branch grow bigger in the result of the model. The regrouping process is accomplished using Microsoft Excel 2016 by using nested if formula to regroup the class into two. The datasets is randomized to create more valid result.

3.3 Modeling

In this study, naive bayes algorithm is employed. The data contained numerical data, nominal and continuous data. The data includes age, pay rate, and days employed. The naive bayes classifier was used to predicted data from class entropy.

Confusion matrix is used to evaluate how many instances that wrongly classified as a result of this model. Furthermore, AUROC score was used to judged this model whether it was right or not. The classification technique that we used in every test is evaluated using 10-folds cross-validation. Moreover, the number of instances for data training was 155 because we applied the 50 percent split.

A cross-validation using 10-folds is executed to evaluate the model performance result more evenly. Meanwhile, fifty-percentage split used until the instance is reduced but not fewer than ten instances to because it would error if it lower than ten instance. A learning curve is created to validate the algorithm have enough instance to predict the employee performance itself.

In this study, we used various tools and hardware to process data that already exists and found the results obtained from this study. The hardware that we used was a laptop that had specifications of the 8th generation Intel Core i5 processor, accompanied by a Video Graphics Adapter (VGA) in the form of Nvidia Mx150 and 8GB RAM. The tools used in this paper were Weka with version 3.8.3 for pre-processing data and classified the data using various methods, and one of them was Naïve Bayes.

4 RESULT & DISCUSSION

With the Naïve Bayes method and measure an updated performance score as the objective variable, with 96.77% acuracy. This Naïve Bayes model takes 0.01 seconds to build with hardware described previously. Figure 1 displays the number of instances with a degree of accuracy. It showns that the number of instances is related to the level of accuracy. As number of instances getting increased, the higher the accuracy became. Table 2 shows the confusion matrix result. There are more true positives than false positives and more false negatives than false positives.

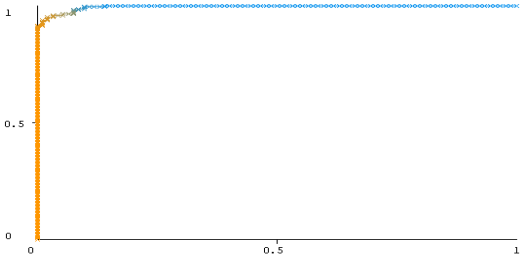

Figure 1: Percentage of instance compare to accuracy

Table 2: Confusion Matrix

	Positive	Negative
True	86,651	1,410
False	48,726	49,1621

Figure 2 describes ROC (Receiver Operating Characteristics) and the number of instances in dataset As the higher the number of instances, the higher the ROC becomes. Figure 3 shows AUC (Area Under the Curve) for the dataset, where it's almost 1. Based on the results of the Naive Bayes classification method, it can be seen that each attribute has an influence on employee performance found in the company. The most essential attributes in employee performance are Pay Rate, Employee status, and Days employed.

Figure 2: Learning rate

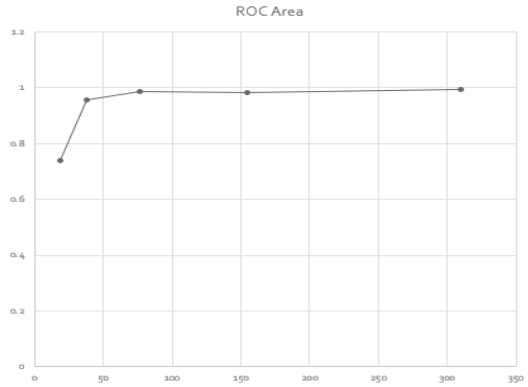

Figure 3: AUC Curve of true positive to false positive rate

5 CONCLUSION

This study shows that human resource can plays an essential role in company growth. A human resource departments need an assessment whether the employee would comply company's wants. They can use of machine learning technology to predict employee's resigment before it happen and can decide in advance how to face it. From the evaluation, correctly classified instance is 95.48% using the proposed model of Naïve Bayes. This is shows that the naïve bayes technique is very good at predicting. Alongside, based on the confusion matrix, it found a slight amount of false-positive result that means the cost of using the naïve bayes technique is small.

REFERENCES

[1] Y. F. Safri, R. Arifudin, and M. A. Muslim, **K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor**, *Sci. J. Informatics*, vol. 5, no. 1, p. 18, 2018. https://doi.org/10.15294/sji.v5i1.12057

[2] O. Olubiyi, G. Smiley, H. Luckel, and R. Melaragno, **A qualitative case study of employee turnover in retail business**, *Heliyon*, vol. 5, no. 6, p. e01796, 2019. https://doi.org/10.1016/j.heliyon.2019.e01796

[3] P. W. Hom, T. W. Lee, J. D. Shaw, and J. P. Hausknecht, **One hundred years of employee turnover theory and research**, *J. Appl. Psychol.*, vol. 102, no. 3, pp. 530–545, 2017. https://doi.org/10.1037/apl0000103

[4] R. F. Cox, R. R. A. Issa, and D. Ahrens, **Management's perception of key performance indicators for construction**, *J. Constr. Eng. Manag.*, vol. 129, no. 2, pp. 142–151, 2003.

[5] J. M. Cramer and B. Roes, **Total employee involvement: measures for success**, *Environ. Qual. Manag.*, vol. 3, no. 1, pp. 39–52, 1993. https://doi.org/10.1002/tqem.3310030105

[6] R. Punnoose and P. Ajit, **Prediction of Employee Turnover in Organizations using Machine Learning Algorithms**, *Int. J. Adv. Res. Artif. Intell.*, vol. 5, no. 9, pp. 22–26, 2016.

[7] A. Idowu, **Effectiveness of Performance Appraisal System and its Effect on Employee Motivation**, *Nile J. Bus. Econ.*, vol. 3, no. 5, pp. 15–39, 2017.

[8] G. Cohen, R. S. Blake, and D. Goodman, **Does Turnover Intention Matter? Evaluating the Usefulness of Turnover Intention Rate as a Predictor of Actual Turnover Rate**, *Rev. Public Pers. Adm.*, vol. 36, no. 3, pp. 240–263, 2016. https://doi.org/10.1177/0734371X15581850

[9] Z. Ahmed, S. Sabir, M. Khosa, I. Ahmad, and M. A. Bilal, **Impact of Employee Turnover on Organisational Effectiveness in Tele Communication Sector of Pakistan**, *IOSR J. Bus. Manag. Ver. IV*, vol. 18, no. 11, pp. 2319–7668, 2016.

[10] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, **Employee turnover prediction with machine learning: A reliable approach**, in *Proceedings of SAI intelligent systems conference*, 2018, pp. 737–758. https://doi.org/10.1007/978-3-030-01057-7_56

[11] P. Ajit, **Prediction of employee turnover in organizations using machine learning algorithms**, *Algorithms*, vol. 4, no. 5, p. C5, 2016.

[12] Q. A and E. Al, **Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance**, *Int. J. Adv. Comput. Sci. Appl.*, vol. 3, no. 2, pp. 144–151, 2012. https://doi.org/10.14569/IJACSA.2012.030225

[13] D. Singh, N. Choudhary, and J. Samota, **Analysis of Data Mining Classification with Decision Tree Technique**, *Glob. J. Comput. Sci. Technol.*, 2014.

[14] Y.-Y. Song and L. U. Ying, **Decision tree methods: applications for classification and prediction**, *Shanghai Arch. psychiatry*, vol. 27, no. 2, p. 130, 2015.

[15] S. Xu, **Bayesian Naïve Bayes classifiers to text classification**, *J. Inf. Sci.*, vol. 44, no. 1, pp. 48–59, 2018. https://doi.org/10.1177/0165551516677946

[16] F. C. Permana, Y. Rosmansyah, and A. S. Abdullah, **Naive Bayes as opinion classifier to evaluate students satisfaction based on student sentiment in Twitter Social Media**, *J. Phys. Conf. Ser.*, vol. 893, no. 1, 2017.

[17] M. M. Saritas, **Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification**, *Int. J. Intell. Syst. Appl. Eng.*, vol. 7, no. 2, pp. 88–91, 2019. https://doi.org/10.18201/ijisae.2019252786

[18] M. Tuga, A. S. Braza, and Fransisca, **Bank Marketing Data Mining using CRISP-DM Approach**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, pp. 2322–2329, Oct. 2019. https://doi.org/10.30534/ijatcse/2019/71852019

[19] J. C. Mogi, N. P. T. Rahmanto, C. Wiranto, and M. Tuga, **Lending Club Default Prediction using Naïve Bayes and Decision Tree**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, pp. 2528–2534, Oct. 2019. https://doi.org/10.30534/ijatcse/2019/99852019

[20] A. Clark, **The Machine Learning Audit—CRISP-DM Framework**, 2018.

[21] C. Patalano, **Human Resources Dataset**, *Kaggle*, 2019. .