ISSN 2278-3091
Volume 9, No.1, January – February 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse111912020.pdf
https://doi.org/10.30534/ijatcse/2020/111912020

Prediction of Employees' Lateness Determinants using Machine Learning Algorithms

John Louis D. Mercaral[*], Allemar Jhone P. Delima[2], Ramcis N. Vilchez[3]
[1,2] Professional Schools, University of Mindanao, Matina, Davao City, Philippines
[3] College of Computer Education, University of Mindanao, Matina, Davao City, Philippines
[1]johnlouis_mercaral@umindanao.edu.ph, [2]allemardelima@umindanao.edu.ph,
[3]ramcis_vilchez@umindanao.edu.ph

ABSTRACT

The lateness of employees greatly affects the organization from manpower, to the financial cost, and even to its production. There are a lot of factors which cause the employees to arrive late at work; hence, it is still a hot topic for research. Data mining techniques have good possibilities to address the issue. Through this study, C4.5 and Naïve Bayes, which are both classification algorithms, were used to test in determining the best technique measured by various criteria, which include accuracy, precision, area under the ROC curve (AUC), Recall, and F-Measure. The results of the study were obtained that Naïve Bayes got higher outcomes with 83.3333% accuracy rate compared to C4.5 with 72.2222% and concluded as potential Data Mining Techniques for employees' tardiness. Finally, in this study, the lateness of employees is often done by employees going late to bed.

Key words : C4.5, Classification, Data Mining, Naïve Bayes

1. INTRODUCTION

The lateness of employees has been unfavorably giving a bad impact on the organization's financial cost [1]. Loss of efficiency to the late employee and workers connected to him and loss of time to the management in dealing with the employee's conduct are among the result of this behavior [1][2]. Some factors ended in tardiness is predicted by commuting, family issues over the work, personality, and commitment to the company [1]-[3].

Both in the organization and studies, lateness has a little highlight compared to other behaviors [2][4]. Absenteeism and turnover are distinguishable whereas the lateness is possibly not present in employee records [2]. Studies of [3]-[5] claimed that there are only small numbers of approaches for studying and works, especially Data Mining prediction targeting employees' lateness. From large chunks of data, data mining provides an automatic or semi-automatic extraction of patterns and knowledge that are probably helpful [3]. One way it is useful is in the management of employees, starting from recruitment, assignment, and behaviors [6].

This study identifies the factor that most influences the employees' tardiness in an IT company in Tagum City, Davao del Norte, through data mining. The processes are done using classification tools such as Naïve Bayes and C4.5 decision tree algorithms. Lastly, this study presents the potential Data Mining Techniques for employees' lateness behavior.

2. RELATED LITERATURE

Data Mining as a newly prominent technique, provides significant solutions in the society specifically the information technology sector [6]-[8]. Widely known as knowledge discovery, associations, clustering, classifying, and forecasting are some of the methods linked with Data Mining. Decision Support Systems (DSS) with the support of data mining, help organizations in producing better decisions [6][7].

Classification and prediction are some of the major data mining techniques. The two data analysis techniques provide define and forecast data classes [7]. The classifiers are used in producing a prediction model in which a new case belongs. The accuracy of prediction changes depending on the classifier, type of attributes, and classes of datasets [9]. There are five categories of classification: statistic, distance, decision tree, neural network, and rule-based in which the most commonly used algorithms are C4.5, Naïve Bayes, K-Nearest Neighbors, and Backpropagation Neural Network [10].

In [9], Naïve Bayes and the C4.5 decision tree for classification are examined in predicting lung cancer survivability based on a 15-year data set of patient records. The study showed that C4.5 is better than Naïve Bayes in predictive effectiveness in data from the SEER database despite the paper emphasize that there is no decision in the comparison of the two in literature. Naïve Bayes performed poorly compared to C4.5 as it got lower accuracy for five, seven, and ten years of training datasets [9].

The two algorithms are also investigated in the study of [10] to determine which is a better algorithm with various measurement metrics in direct marketing. The two algorithms produced a forecast for the client's subscription of a term deposit in a banking institution in Portugal. With a total of 45,

211 records in the dataset, C4.5 have outperformed in terms of accuracy, precision, and AUC [10]. Both C4.5 and Naïve Bayes algorithms are also applied in the study of [11]. The study used a bank dataset with 300 instances and nine attributes taken from the UCI repository. Results showed that C4.5 gave simpler and higher accuracy in the performed classification and proven that more cost-efficient compared to Naïve Bayes [11].

Moreover, data mining techniques are also utilized in perceiving of employees' behavior. C4.5 decision tree along with Random Tree and Random Forest algorithms are used in obtaining the most influential determinants for teachers arriving late at school. A 2.5-month data is collected in the Kalam Kudus Christian School III Jakarta, and various measurements such as accuracy and AUC are assessed, Random Tree took at the top out of the three algorithms [3].

The local dataset of 1000 employee records with thirty features from different institutions and firms in the Kurdistan region in Iraq is utilized in the study of [11], which used FR nearest neighbor (FRNN), C4.5, Naïve Bayes, and Convolution Neural Network (CNN). These classification techniques are examined for the improvement of predicting employees' behavior in which concluded that both FRNN and CNN classifiers considered producing a higher accuracy rate [12].

Further, employees' performance forecast is also addressed with data mining. The study of [6] focused on the Human Resource functions, specifically talent management, and used data from the higher education institution employees, which contain 53 related attributes. Classifier algorithms such as Random Forest and C4.5 for decision tree and Multilayer Perceptron (MLP) and Radial Basic Function Network for the neural network are analyzed to determine the appropriate classifier for the HR department. The results show that both the decision tree and neural network are suitable for talent management in performance prediction, and C4.5 has the highest accuracy [6].

The table 1 shows the summary of the literature review findings, which includes the problem, used machine learning algorithms, and recommended technique of the study.

3. METHODOLOGY

3.1 Datasets

The survey, as a suitable approach for an organized process of collecting data and allowing a greater number of participants, produces direct and accurate results in which it gives way to access the participants' confidentiality [12]. Consequently, in this study, the dataset is from the data collected through a survey that was given to the participants in an IT company. The organization is a systems integration company under the Digital Exchange Community specializing in transaction management systems that have over 100 employees. The company has other locations and headquartered in Manila, Philippines. In this study, it only covered their location in Tagum, City, Philippines, which has over 37 employees. The survey administration application developed by Google, namely, Google Forms, is leveraged for the survey, and each employee of the institution answered based on the form to obtain the dataset. The survey consists of items that equate as the total attributes for the dataset. Some of these attributes covered in the data collection are age, position, sex, and employee's lateness determinants such as distance from home and office, organizational commitment, going late in bed, with vehicle or not, punctuality, and work-family conflict.

Table 1. Related Works

Research Authors	Problem studied	Used Data Mining Techniques	Recommend
G. Dimitoglou, J. A. Adams, and C. M. Jim [9]	Prediction of Lung Cancer Survivability	C4.5 and Naïve Bayes	C4.5
M. Karim and R. M. Rahman [10]	Classification and generation of Actionable Knowledge for Direct Marketing	C4.5 and Naïve Bayes	C4.5
A. Goyal and R. Mehta [11]	Classification algorithm performance	C4.5 and Naïve Bayes	C4.5
W. Gata et al. [3]	Prediction of teachers' tardiness	C4.5, Random Tree, and Random Forest	Random Tree
T. A. Rashid and A. L. Jabar [12]	Prediction of employees' behavior	FR nearest neighbor (FRNN), C4.5, Naïve Bayes, and Convolution Neural Network (CNN)	FRNN and CNN
H. Jantan, A. R. Hamdan, and Z. A. Othman [6]	Prediction of employees' performance	C4.5, Random Forest, Multilayer Perceptron and Radial Basic Function Network	C4.5

The efforts of identifying issues, which include noise, outliers, errors, and missing values can result in better training responses in data mining and machine learning algorithms [12]. In this case, the data collected in the study undergo a pre-processing activity aiming to guarantee that data is suitable, well-formatted, and complete.

3.2 Classification

Classification serves the method in prediction, and its performance will depend on the organization's dataset [3][12]. Therefore, applied algorithms such as C4.5 and Naïve Bayes classifiers of related findings suggested and particularized. These algorithms are used to examine and predict employees' lateness to address the problem of the study.

3.2.1 C4.5

The C4.5, the extension of ID3 developed by Quinlan, is one of the decision tree algorithms [3][10][12]. C4.5 is the leading and most productive algorithm compared to other members of the decision tree family. The algorithm is composed of nodes that represent the tree roots and leaves. C4.5 applies a recursive pruning method in which the approximated error rate is used to remove part of the tree, and the process iterates and stops if nodes can't be tested and produce leaves [3][10].

3.2.2 Naïve Bayes

Naïve Bayes prominently known for its simplicity and performance as a probabilistic typical classifier in the calculation of combinations of values and frequency counting of a data set [9][10][13]. The algorithm implements Bayes theorem, which presumes that all of the attributes are independent of one another [9][13]. Naïve Bayes only requires a slight number for training data in estimating parameters for the classification [13].

The logic basis behind this Bayes' rule algorithm is to train P (Y|X) and utilized to acquire approximations of P (X|Y) and P(Y) [13]. The posterior probability P (Y|X) calculated for all classes in which the instance's label is the class with the highest probability [10].

4. RESULTS AND DISCUSSION

In this study, both C4.5 and Naïve Bayes algorithms are used for the training and testing data through the leverage of Waikato Environment for Knowledge Analysis (WEKA.) The comparison of the two techniques can be predicted, which performed better through the performance table below. Table 2 shows the algorithm test performance results, which include five criteria, namely the accuracy, precision for yes and no, AUC, Recall, and F-Measure.

The accuracy sets the number of total correct predictions. Naïve Bayes with 83.3333% has a greater percentage compared to C4.5 with 72.2222%. Figure 1 presents the True Positive (TP) rate for two classification algorithms whereas Figure 2 describes the graphical representations of the accuracy of the two algorithms.

Aside from the confusion matrix, Receiver Operating Characteristic (ROC) graph to evaluate the machine learning algorithms. The ROC graph consists of two coordinates which are the X-axis and Y-axis that are plotted with the false-positive rate and the true positive rate, respectively. To come up with the measurement of the ROC, the area under the ROC curve (AUC) is a method to be used in which it is considered perfect if the AUC resulted in 1. On the other hand, 0.5 AUC result defines the prediction is random [10].

Table 2. Results of two algorithms.

Criteria	Naïve Bayes	C4.5	Difference	Better
Accuracy	83.33%	72.22%	11.111	NB
Precision for yes	0.85	0.727	0.123	NB
Precision for no	0.813	0.714	0.099	NB
AUC	0.8812	0.7219	0.1593	NB
Recall	0.833	0.722	0.111	NB
F-Measure	0.833	0.72	0.113	NB

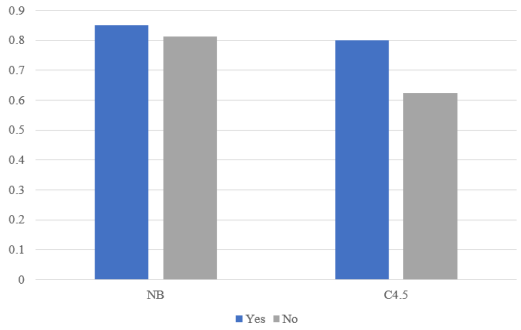

Figure 1: Positive rate of the two algorithms.

Figure 3 and Figure 4 depicts the ROC curve for yes and no classes of Naïve Bayes. The figures describe the technique has good classification characteristics where the ROC of no is less close compared to the class yes. Besides, the AUC for both classes is 0.8812. However, the ROC and AUC of C4.5 gained lower opposed to Naïve Bayes. Figures 5 and 6 show the results for C4.5 on ROC and AUC.

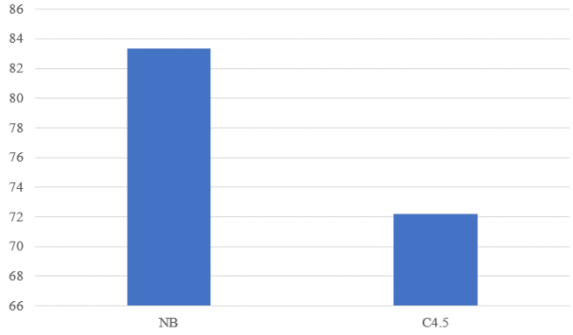

Figure 2: Accuracy of the two algorithms.

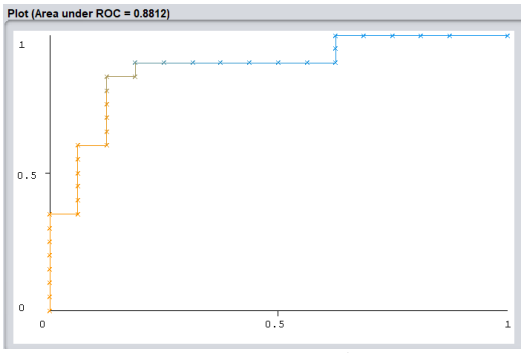

Figure 3: Naïve Bayes ROC for yes class.

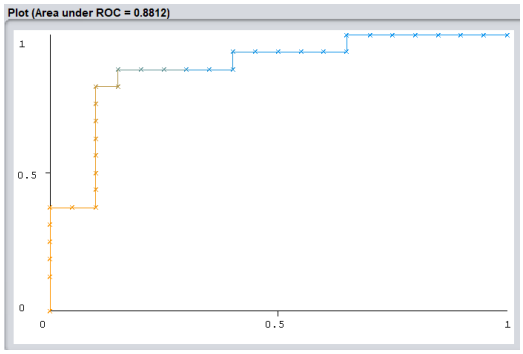

Figure 4: Naïve Bayes ROC for no class.

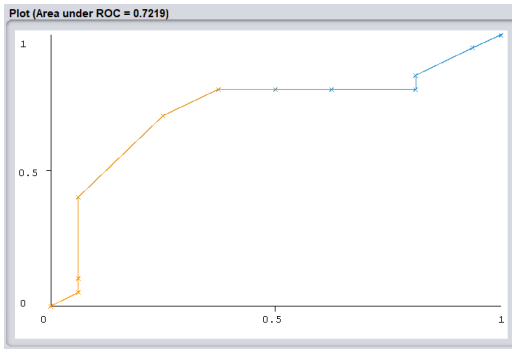

Figure 5: C4.5 ROC for yes class.

Other performance metrics are precision, recall, and F-measure. Precision defines the segment of related instances among the retrieved instances whereas recall defines the segment of the total number of related instances that were truly retrieved. The result of the calculation of precision and recall is the F-measure. Figure 7 depicts the graphical representations of precision, recall, and F-measure values of the two algorithms.

Additionally, the study also aims to provide the most influential factor for employee's tardiness. The determination of most affecting attribute is defined by the Gain Ratio value. Table 3 shows that going late in bed has the highest gain ratio value.

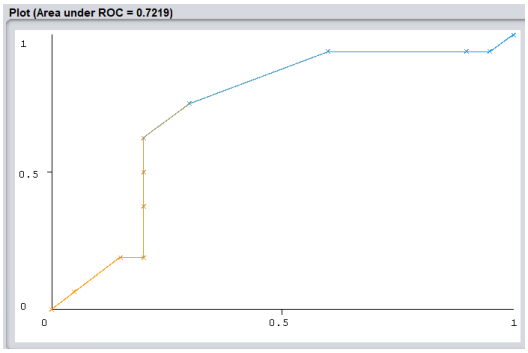

Figure 6: C4.5 ROC for no class.

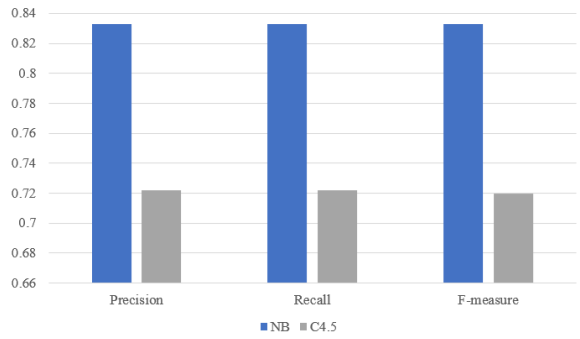

Figure 7: Naïve Bayes and C4.5 precision, recall, and F-measure results.

Table 3. Gain Ratio

Attribute	Value
Late_in_bed	0.365
House_to_Office	0.343
Work_Family_Conflict	0.286
Organizational_Commitment	0.15

5. CONCLUSION

This study analyzes the strength of data mining and machine learning methods, particularly the C4.5 and Naïve Bayes, to accurately predict the lateness of the employees in the organization. The evaluation and comparison of such methods are also significant to take close to a satisfactory decision. With various measure metrics, Naïve Bayes gained an 83.3333% accuracy rate, which is higher compared to C4.5. It also determines that Naïve Bayes works better with small numbers of data compared to C4.5 Related findings show that C4.5 of the Decision Tree family outperformed in contrast to Naïve Bayes. Unlike this study, which has a small number of instances, these previous works have a greater number of data, including the studies of [9], [10], and [11] with a 15-year dataset, 45, 211 records, and 300 instances respectively.

Furthermore, based on the results gained in the study, late_in_bed is the attribute that has the most influence on employees' lateness factor. In conclusion, the employees' lateness in this study is often done by the employees who are late going to bed compared to employees who are early to sleep.

REFERENCES

[1] M. Dishon-berkovits and M. Koslowsky, "Determinants of employee punctuality," *J. Soc. Psychol.*, vol. 142, no. 6, pp. 723–739, 2002, doi: 10.1080/00224540209603932.

[2] M. Koslowsky, "A new perspective on employee lateness," *Appl. Psychol.*, vol. 49, no. 3, pp. 390–407, 2000, doi: 10.1111/1464-0597.00022.

[3] W. Gata *et al.*, "Prediction of Teachers' Lateness Factors Coming to School Using C4.5, Random Tree, Random Forest Algorithm," 2019, vol. 258, no. Icream 2018, pp. 161–166, doi: 10.2991/icream-18.2019.34.

[4] J. J. Bardsley and S. R. Rhodes, "Using the steers-rhodes (1984) framework to identify correlates of employee lateness," *J. Bus. Psychol.*, vol. 10, no. 3, pp. 351–365, 1996, doi: 10.1007/BF02249608.

[5] G. Blau, "Developing and Testing a Taxonomy of Lateness Behavior," *J. Appl. Psychol.*, vol. 79, no. 6, pp. 959–970, 1994, doi: 10.1037/0021-9010.79.6.959.

[6] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Towards applying Data Mining Techniques for Talent Mangement," in *International Conference on Computer Engineering and Applications*, 2009, vol. 2, no. 2011, pp. 476–481.

[7] J. Ranjan, D. P. Goyal, and S. I. Ahson, "Data mining techniques for better decisions in human resource management systems," *Int. J. Bus. Inf. Syst.*, vol. 3, no. 5, pp. 464–481, 2008, doi: 10.1504/IJBIS.2008.018597.

[8] I. S. Makki and F. Alqurashi, "An adaptive model for knowledge mining in databases 'EMO_MINE' for tweets emotions classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 7, no. 3, pp. 52–60, 2018, doi: 10.30534/ijatcse/2018/04732018.

[9] G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability," pp. 1–9, 2012.

[10] M. Karim and R. M. Rahman, "Decision Tree and Naive Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing," *J. Softw. Eng. Appl.*, vol. 06, no. 04, pp. 196–206, 2013, doi: 10.4236/jsea.2013.64025.

[11] A. Goyal and R. Mehta, "Performance comparison of Naïve Bayes and J48 classification algorithms," *Int. J. Appl. Eng. Res.*, vol. 7, no. 11 SUPPL., pp. 1389–1393, 2012.

[12] T. A. Rashid and A. L. Jabar, "Improvement on predicting employee behaviour through intelligent techniques," *IET Networks*, vol. 5, no. 5, pp. 136–142, 2016, doi: 10.1049/iet-net.2015.0106.

[13] 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, doi: 10.14569/ijarai.2016.050904.