I. J. Education and Management Engineering, **2022, 3, 22-29**
Published Online June 2022 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2022.03.03

Determining Factors Resulting to Employee Attrition Using Data Mining Techniques

Jennifer Anne A. Repaso
Bulacan State University
Email: jenniferanne.repaso@bulsu.edu.ph

Elenita T. Caparino
Bulacan State University
Email: elenita.caparino@bulsu.edu.ph

Mary Grace G. Hermogenes
Bulacan State University
Email: marygrace.hermogenes@bulsu.edu.ph

Joann G. Perez
Bulacan State University
Email: joann.galopo@bulsu.edu.ph

Received: 26 October 2021; Accepted: 25 January 2022; Published: 08 June 2022

Abstract: Business Process Outsourcing is a budding industry which currently employs millions of workers in the Philippines which draws applicants from undergraduate to professionals. It provides high-quality, well-paying jobs to millions of Filipinos while inspiring economic activity and investments all around. However, attrition rate of around 50 percent in the current year is a big challenge to predict employee turnover. This study came up with a model that can be adopted in the organization to predict possible attrition and guide the employers particularly the HR team in determining first-hand the type of applicant that they have by applying Data Mining techniques. The authors extracted significant predictors among the given data from a BPO company. Fast Correlation-Based Filtering Algorithm was performed to remove irrelevant data and increase learning accuracy. 1470 records with 21 attributes were initially provided and 17 were identified as significant after filtering and preprocessing of data was performed. The preprocessed data was used for model building with the application of Naïve Bayes Algorithm. The resulting model predicted percentage probability of hoppers and stayers. Among the 17 given variables, Total Working Years, Marital Status and Age ranked as the top predictors in determiningpossibility of attrition. The data was split into 60% training data or a total of 882 records and 40% testing data or a total of 588 records. The predicted number of stayers is 542 or 92.2% and the predicted hoppers or who likely to resign are 46 or 7.8% The model was tested and evaluated to check accuracy of result through confusion matrix cross validation technique which yielded an accuracy percentage of 84.69%.

Index Terms: Data Mining Techniques, Employee Attrition, Filtering Algorithm, Naïve Bayes Algorithm, Confusion Matrix.

1. Introduction

Philippines serves as one of leading destination for Business Process Outsourcing (BPO). The growth is driven by a lot of factors such as low labor cost, skilled workforce, and neutrally accented English language [1]. Its role in the growth and evolution of APEC economies is undeniably very important and one that warrants attention in a regional forum such as APEC [2]. The BPO sector currently employs millions of workers in the Philippines and still growing which is an indication of future development and a great source of job creation.

Despite all the benefits, this industry is facing the challenge of finding quality human resources given the high attrition rate and this is a common issue that most HR managers face today [3]. According to the IT and Business Process Association of the Philippines (IBPAP), average yearly industry turnover for its members was 38% in 2009; that figure decreased to 19 per cent in 2015 but remained high compared to industry standards. With the seemingly increasing attrition the Business organizations are really interested to come up with plans on how to retain their employees and to initially select proper employees who will be a potential stayer in the company. The most important asset of the company

is their employees, so if attrition rate of employees continuously increase that will be a big problem to the company. Understanding and forecasting turnover at the firm and departmental levels is essential for reducing attrition as well as for effectively planning, budgeting, and recruiting in the human resource field [3]. Furthermore, turnover disrupts social and communication structures and causes productivity loss, and it also demoralizes the remaining employees and leads to additional turnover. Failure to predict employee turnover reduces the performance and profits of the organization and interrupts the organizational structure. The high turnover rate and the low replenishment rate are both symptoms of deeper problems in this industry, mainly in human resources management and industrial relations (IR/HRM) [4]

With all this issues a need to come up with a quicker decision concerning how to retain its employees is vital. It has become imperative to identify the factors that lead to attrition, hereby finding out some strategies and model to be adopted in the organization to minimize the attrition rate and empower or at least guide the employers particularly the HR team in determining first-hand the type of applicant that they have through Data Mining.

This study aims to apply a predictive model to identify whether an applicant in the BPO Company is a possible stayer or a hopper type of employee using Naïve Bayes Algorithm. Specifically, the following are to be met:

1. To identify predictors using Fast correlation-based Filtering algorithm for feature selection;
2. To develop a model to identify if applicant is a stayer or hopper type of employee using Naïve Bayes Algorithm;
3. To evaluate the accuracy of the model using confusion matrix cross validation

2. Related Works

Data mining is widely used now in academic researches and studies because it deals mostly with several given data that needs to be analyzed, filtered, classified and categorized to make this data meaningful to be used in the specific study or research. In Data Mining, discovery of patterns and relationship of data is part of the process called knowledge discovery which describe steps to be taken to ensure meaningful result. In other words, data mining is another way in finding meaning in data.

In the study by [1] they predicted student's performance by basing it on diverse factors like personal, social, and psychological and their environmental they were able to attain their objective through data mining.

Data mining is widely used to extract useful patterns or rules from a large database through an automatic or semi-automatic exploration and analysis of data. Prior to model creation as a major part of a data mining process, data pre-processing is also essential. Data pre-processing involves data cleaning like imputation, feature selection, discretization, and data reduction. This enables building models achieving a higher accuracy rate. Studies show that real data sets which have undergone data pre-processing achieve efficient results [2, 3] With the use of data mining techniques, processors are no longer limited to passively storing or collecting data. They can also help the users to actively extract the key points from large amounts of data and make use of analysis or prediction.

Data mining problems are generally categorized as association, clustering, classification, and prediction. In the study by [4] they defined each categories in data mining, Association is the discovery of association rules showing attribute-value conditions that occur frequently together in a given dataset. Clustering is the process of dividing a dataset into several clusters in which the interclass similarity is maximized while the inter-class similarity is minimized. Classification derives a function or model that identifies the categorical class of an object based on its attributes. Prediction is a model that predicts a continuous value or future data trends.

This paper will focus on identifying or classifying applicants in the BPO industry whether they will be stayers or hopper type of employees. Since attrition in the BPO industry is getting higher in time the study will aid the employers especially the HR management in immediately identifying or predicting the type of applicant coming and prioritize those that will likely to be a stayer type so that trainings and efforts of the training team will not be put into waste. Amongst all the algorithm used in data mining the proponent chose to use Naïve Bayes algorithm to serve as a tool in coming up with logical solutions to classify the given data sets and identify which set belongs to either of the 2 classifications, the hopper type, or the stayer type of employees.

According to the book "How the Naïve Bayes Classifier Works in Machine Learning" by [5], In term of classification Naïve Bayes Algorithm is a fast, highly scalable algorithm that can be used for Binary and Multiclass classification. It is a simple algorithm that depends on doing a bunch of counts but can easily be used on small to large dataset. It is also a popular choice for text classification problems. In the study conducted by [4] they used Naïve Bayes classification algorithm to predict a target class depending on in its calculations on probabilities, namely Bayesian theorem. It was also stated that because of this use, results from classifier are more accurate and effective, and more sensitive to new data added to the dataset. Table-1 shows the result of their study using different classification technique.

A research conducted by [6] explored different classification techniques for predicting career classification. According to the results, when compared to other classification algorithms used in the dataset, Naive Bayes and Random Forest produced high accuracy results. Naïve Bayes algorithm has been used in the study conducted by different authors [7]–[16]. Also, in another study conducted by [17] Naive Bayes classifier are used for the prediction of COVID-19. The accuracy of NB got a 99%. Another study applying the Naïve Bayes is a research by [18] applying the soil nitrogen mapping with a smart prototype using the TCS3200 sensor combined with Naive Bayes algorithm and GIS (Geographical

information systems). A research by [19] utilized Na ïve bayes in sentiment analysis. From the result obtained it was seen that in case of movie reviews Naïve Bayes' got better results than K-NN. Another study by [20] shows higher accuracy rate of Na ïve Bayes Algorithm as compared to other data mining algorithm. Table 2 shows the result in terms of classification accuracy result.

Table 1. Accuracy percentage on prediction performance

Classification Algorithm	Accuracy %
Na ïve Bayesian	84.3187
K-Star	71.4286
Random Forest	73.6264
Zero R	75.8242

Table 2. Accuracy of classification algorithms

Algorithm	Evaluation using Cross Validation	Evaluation using Hold-out
ID3	50%	43.7%
C4.5 (J4.8)	60.5%	56.2%
Na ïve Bayes	65.8%	68.7%

With these results, Na ïve Bayes algorithm is used in the study to come up with a better outcome. This algorithm may also be helpful in obtaining the objective to come up with logical prediction of the type of employee whether they will become a stayer hopper type of employee. This algorithm uses conditional probability, through this we can calculate the probability of an event using its prior knowledge. Which means that we make prediction based on prior knowledge and current evidence. Below is the formula for calculating the conditional probability using Na ïve Bayesian

$$\frac{P(H|E=P(E|H*P(H)}{P(E)} \quad (1)$$

Where:

P(H) is the probability of hypothesis H being true. This is known as the prior probability or the categorical outcome
P(E) is the probability of the evidence (regardless of the hypothesis) or Predictors.
P(E|H) is the probability of the evidence given that hypothesis is true.
P(H|E) is the probability of the hypothesis given that the evidence is there.

3. Framework of the Study

The study aims to develop a predictive model using applicable attributes to predict the employee turnover in the BPO industry Figure 1 shows the framework which is divided into 3 major activities which are the Data Preparation, which includes pre-processing and Feature Selection using Filtering, Classification and Evaluation and Validation of data. The feature selection is the data cleansing part of the gathered data wherein data will be analyzed and identify correct or inconsistent data and remove noisy data. This is the preprocessing stage. In this stage the proponent will use Fast correlation-based filtering algorithm using RapidMiner software to filter data and select the best attributes or variables to be included in the classification stage.

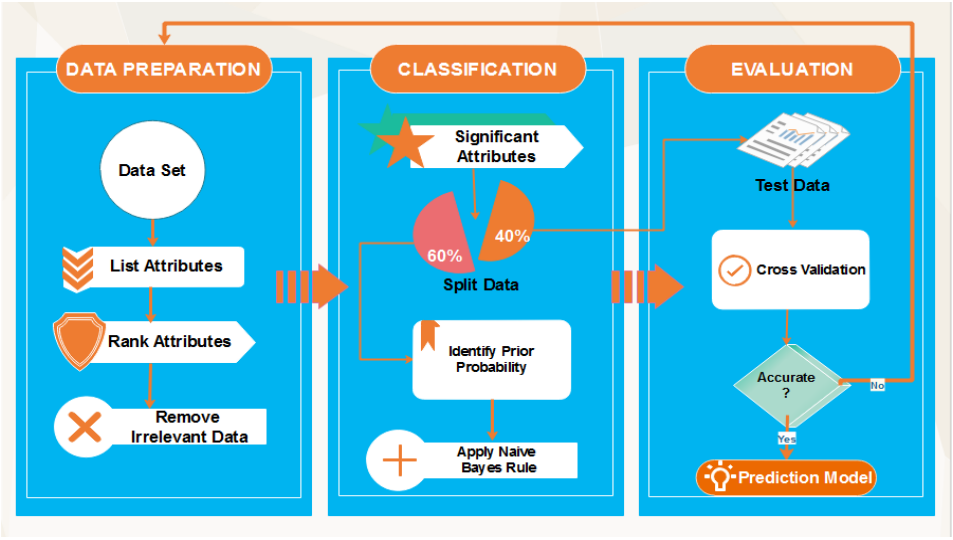

Fig. 1. Predictive Model for Employee Attrition

The table below shows the attributes derived from the given data from a BPO company. The given attributes were provided by the company from their employees' profile with 1,233 employee records which was filtered in the process. The filtered data was used for classifying data using Naïve Bayes algorithm.

Table 3. Attributes containing employee information

List of attributes			
1	Age	14	Over18
2	Department	15	Overtime
3	DistanceFromHome	16	PerformanceRating
4	Education	17	RelationshipSatisfaction
5	EducationField	18	StandardHours
6	EmployeeCount	19	TotalWorkingYears
7	EnvironmentSatisfaction	20	TrainingsAttended
8	Gender	21	WorkLifeBalance
9	JobInvolvement	22	YearsAtCompany
10	JobSatisfaction	23	YearsInCurrentRole
11	MaritalStatus	24	YearsSinceLastPromotion
12	MonthlyRate	25	YearsWithCurrManager
13	NumCompaniesWorked		

The next stage to perform is classification. In this process the proponent will be using Naïve Bayes Algorithm to classify respondents (employees) into stayer type or hopper type based on the selected features or attributes. This is where all attributes are analyzed based on the chosen algorithm to come up with logical classification. The last stage of the entire process is to test and evaluate the accuracy of the result from the classification phase to come up with accurate prediction tool. Confusion matrix cross validation will be used in this process.

4. Results and Discussion

In the data filtering phase, Fast Correlation based Filtering algorithm was used in the process. This is also referred to as Pearson correlation coefficient in statistics. Among 21 attributes, the variables with 0 weight, missing and and insignificant values has been removed since it was interpreted as insignificant data and will not be helpful in the classification. 3 attributes were removed namely employee count, over 18 and Standard hours. 17 were identified as significant variables and 1 variable set as label which is attrition. The Resulting attribute and its weight value is listed in Table 4.

Table 4 Significant attributes ranked by weight

Weight	Attribute
0.172	TotalWorkingYears
0.162	MaritalStatus
0.161	Age
0.134	YearsAtCompany
0.108	MonthlyRate
0.103	JobSatisfaction
0.103	EnvironmentalSatisfaction
0.078	DistanceFromHome
0.075	EducationField
0.064	Dapartment
0.064	WorkLifeBalance
0.046	RelationshipSatisfaction
0.043	NumCompaniesWorked
0.031	Education
0.030	YearsSinceLastPromotion
0.029	Gender
0.003	PerformanceRating

After the data preparation phase the filtered data or the final data is loaded using Rapid miner tool for model building andtesting. Then attrition attributes are set to label for prediction and selected for model building. The data gathered was split into training (60%) and testing data (40%). Figure 2 shows the process performed in Rapid Miner Tool.

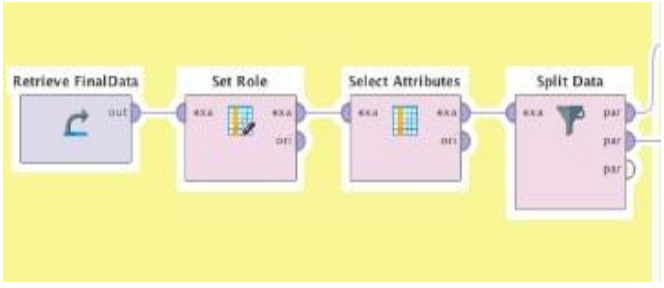

Fig.2. Data Preparation using Rapid Miner

The next step is to use Naïve Bayes algorithm for classification of data. The model is tested using confusion matrix cross validation to get performance accuracy of the model. There are 2 classifications identified which are hoppers or employees who are likely to resign and stayers or employees who are likely to stay. The result of the classified data was generated wherein 92.2% of the total employees (542) was predicted as stayer or willlikely to stay in the company and 7.8% of the employees will likely to hop or resign in the company. The process for model building is shown in Figure 3.

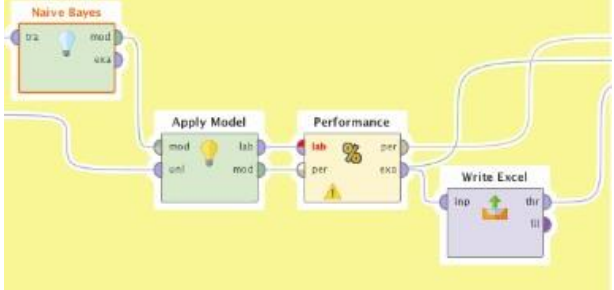

Fig.3. Model building using Naïve Bayes algorithm

The test result is 84.69% using confusion matrix cross validation techniques and the accuracy percentage result is shown in Table 5.

Table 5. Accuracy result of the model

Accuracy: 84.69%			
	true Yes	true No	Class Precision
pred. Yes	22	24	47.83%
Pred. No	66	476	87.82%
Class	25.00%	95.20%	

Amongst all the significant attributes the top predictors that was identified was Total Working Years, Marital Status and Age. Figure 4, shows the age range, marital status and range of working years with the highest confidence level which means that these are the most likely to be hopper type of employees. Based on the result as shown in table 4-4, age range between 18-21 are the ages that are most likely to resign. And in terms of Marital status, those who are single has the highest probability of becoming hoppers type of employees. The result also shows that those employees who has 0-2 working years has a tendency to resign.

prediction(A...	confidence(Yes) ↓	Age	TotalWorkin...	MaritalStatus
Yes	0.963	19	2	Single
Yes	0.950	19	2	Single
Yes	0.905	19	0	Single
Yes	0.900	20	1	Single
Yes	0.868	20	2	Single
Yes	0.859	18	0	Single
Yes	0.854	21	2	Single
Yes	0.788	21	1	Single
Yes	0.752	19	1	Single
Yes	0.740	20	2	Single

Fig.4. Top Predictors and values with the highest confidence level generated using RapidMiner Tool

5. Conclusion and Future Works

Predicting employee attrition can be useful to the HR management so that they can predict percentage of attrition within their employees. Once they identified the possible attrition rate, they can immediately plan strategies on hiring and as to how many employees they need within the company. They can also plan comprehensive retention strategies for their employees. From the result of the model, it also provides the top variable (rank 1) which is the common predictor of attrition, and this is Age of the employee Based on the result as shown in Figure 4, age range between 18-21 are the ages that are most likely to resign. And in terms of Marital status, those who are single has the highest probability of becoming hoppers type of employees. The result also shows that those employees who has 0-2 working years has a tendency to resign.

Future researchers may incorporate this model into a real-world system that predicts which employees are likely to resign. Future researchers can also apply the model with more complex employee records from companies, as well as other predictor variables related to employee attrition.

References

[1] B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification," vol. 9, no. 4, 2012.
[2] J. Galopo Perez and E. S. Perez, "Predicting Student Program Completion Using Naïve Bayes Classification Algorithm," *Int. J. Mod. Educ. Comput. Sci.*, vol. 13, no. 3, pp. 57–67, 2021, doi: 10.5815/ijmecs.2021.03.05.
[3] E. T. Capariño and A. M. Sison, "Application of the Modified Imputation Method to Missing Data to Increase Classification Performance," *2019 IEEE 4th Int. Conf. Comput. Commun. Syst.*, pp. 134–139, 2019.
[4] Q. a. Al-Radaideh and E. Al Naqi, "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.
[5] R. Saxena, "How The Naive Bayes Classifier Works In Machine Learning," *Dataaspirant*. 2017.
[6] J. A. A. Repaso and E. T. Capariño, "Analyzing and predicting career specialization using classification techniques," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 1 Special Issue 3, pp. 342–348, 2020, doi: 10.30534/ijatcse/2020/5391.32020.
[7] E. T. Caparino, "Analyzing and Predicting IT Career Specialization Using Naïve Bayes Algorithm".

[8] M. Shouman, T. Turner, and R. Stocker, "INTEGRATING NAIVE BAYES AND K-MEANS CLUSTERING WITH DIFFERENT INITIAL CENTROID SELECTION METHODS IN THE DIAGNOSIS OF HEART DISEASE PATIENTS," pp. 125–137, 2012, doi: 10.5121/csit.2012.2511.
[9] R. Shinde, S. Arjun, P. Patil, and P. J. Waghmare, "An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm," vol. 6, no. 1, pp. 637–639, 2015.
[10] H. Shaziya, R. Zaheer, and G. Kavitha, "Prediction of Students Performance in Semester Exams using a Naïve bayes Classifier," pp. 9823–9829, 2015, doi: 10.15680/IJIRSET.2015.0410072.
[11] M. June, W. Ahmed, and M. T. Scholar, "Available Online at www.ijarcs.info Performance Analysis of Naïve Bayes Algorithm on Crime Data using Rapid Miner," vol. 8, no. 5, pp. 683–687, 2017.
[12] E. Studies, "Enhancing Forecasting Performance of Naïve-Bayes Classifiers with Discretization Techniques," pp. 24–30.
[13] W. A. Van Eeden *et al.*, "Predicting the 9-year course of mood and anxiety disorders with automated machine learning : A comparison between auto-sklearn , naïve Bayes classifier , and traditional logistic regression," *Psychiatry Res.*, vol. 299, no. October 2020, p. 113823, 2021, doi: 10.1016/j.psychres.2021.113823.
[14] Muladi, U. Pujianto, and U. Qomaria, "Predicting high school graduates using Naive Bayes in State University Entrance Selections," *4th Int. Conf. Vocat. Educ. Training, ICOVET 2020*, pp. 155–159, 2020, doi: 10.1109/ICOVET50258.2020.9230336.
[15] K. Yadav, "Comparing the Performance of Naive Bayes And Decision Tree Classification Using R," no. December, pp. 11–19, 2019, doi: 10.5815/ijisa.2019.12.02.
[16] S. Maitra, S. Madan, R. Kandwal, and P. Mahajan, "Mining authentic student feedback for faculty using Naïve Bayes classifier," *Procedia Comput. Sci.*, vol. 132, pp. 1171–1183, 2018, doi: 10.1016/j.procs.2018.05.032.
[17] N. Deepa, J. S. Priya, and T. Devi, "Towards applying Internet of Things and Machine Learning for the Risk Prediction of COVID-19 in pandemic situation using Naive Bayes Classifier for improving Accuracy," *Mater. TODAY Proc.*, 2022, doi: 10.1016/j.matpr.2022.03.345.
[18] A. Yudhana, D. Sulistyo, and I. Mufandi, "GIS-based and Naïve Bayes for nitrogen soil mapping in Lendah, Indonesia," *Sens. Bio-Sensing Res.*, vol. 33, p. 100435, 2021, doi: 10.1016/j.sbsr.2021.100435.
[19] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment Analysis of Review Datasets Using Naïve Bayes' and K-NN Classifier," *Int. J. Inf. Eng. Electron. Bus.*, vol. 8, no. 4, pp. 54–62, 2016, doi: 10.5815/ijieeb.2016.04.07.
[20] S. Deepajothi and S. Selvarajan, "A Comparative Study of Classification Techniques On Adult Data Set 1," vol. 1, no. 8, pp. 1–8, 2012.

Authors' Profiles

Jennifer Anne Repaso is a faculty member from Bulacan State University Sarmiento Campus under the Information Technology Department. She has a degree of B.S.Ed. Computer Major taken from University of Santo Tomas Manila and earned her Master in Information Technology (MIT) from Technological Institute of the Philippines Q.C. Currently, she is finishing her Doctor in Information Technology also in Technological Institute of the Philippines, Manila Campus.

Elenita T. Caparino is currently a faculty member of Bulacan State University, Sarmiento Campus, with a rank of Associate Professor I. She earned her Doctor in Information Technology (DIT) degree from the Technological Institute of the Philippines, Quezon City. She was a recipient of the K-12 Graduate Scholarship Program sponsored by the Commission on Higher Education (CHED) from 2016-2019.

Mary Grace G. Hermogenes is currently a faculty member of Bulacan State University, Sarmiento Campus, with a rank of Associate Professor V. She earned her Master of Science in Information Technology at the Hannam University of South Korea as faculty scholar. She finished her Doctor of Philosophy in Education major in Education Leadership and Management at La Consolacion University Philippines.

Joann G. Perez earned her Master's Degree in Information Technology at Technological Institute of the Philippines and is now pursuing her Doctor in Information Technology Program at the same institution. Currently, she is a full-time faculty member of Bulacan State University-Sarmiento Campus. Her research interests include data mining and software engineering. Ms. Perez is a member of the Philippine Society of IT Educators Central Luzon Chapter (PSITE R3).

How to cite this paper: Jennifer Anne A. Repaso, Elenita T. Capariño, Mary Grace G. Hermogenes, Joann G. Perez, " Determining Factors Resulting to Employee Attrition Using Data Mining Techniques", International Journal of Education and Management Engineering (IJEME), Vol.12, No.3, pp. 22-29, 2022. DOI: 10.5815/ijeme.2022.03.03