Predictive Analytic on Human Resource Department Data Based on Uncertain Numeric Features Classification

https://doi.org/10.3991/ijim.v15i08.20907

Asrul Huda (✉), Noper Ardi
Universitas Negeri Padang, Padang, Indonesia
asrulhuda@gmail.com

Abstract—Business Intelligence is very popular and useful for a better understanding of business progress these days, and there are many different methods or tools being used in Business Intelligence. It uses combination of artificial intelligence, data mining, math, and statistic to gain better understanding and insight on the business process performance. As employees have an important role in business process, the desire to have a tool for classifying and predicting their wages are desirable. In this research, we tried to analyzed dataset from Human Resource Department, and this dataset can be used to analyst the data in order to draw a conclusion about whether any employees would prematurely leave the company, and then, a preventive action based on those parameters can be proposed. This is a kind of predictive analytic system which bases on Naïve Bayes, and it can predict whether an employee would leave or stay according to his or her characteristics. But the Naïve Bayes itself does not enough. So we develop a way to solve the problem using uncertain Numeric features classification on it. The accuracy of the result is depended on the amount and effectiveness of the training sets.

Keywords—Predictive Analytic; Naïve Bayes; Business Intelligence; Human Resource

1 Introduction

Human analytical is an application of math, statistics and data modeling related to employees in the company to view and predict future employee performance based on the data. This analysis is commonly known as HR (Human Resource) Analytic. Improving business performance is the main goal of HR strategy, and HR analytic is used to make a better decision about it.

Employees play a substantial role in the business process. The longer employees work in a company, the greater their value for the company. The value means in this context is such as their familiarity of the company, work experience, tool used in the company, even a project that they work in it before. So, keeping them comfortably staying in the company is essential[1][2]. Occasionally, they leave the company without any notification. This can be happening due to many factors such as salary, over workload, promotion, or even better opportunity from other company. If this situation

is not taken seriously, it will be detrimental to the company. When they walk out the door, their substantial value also gone with them. Not surprisingly then, the tool for predicting this kind of situation is substantial. In an organization, the employee's data stored in HR which can be used to make a predictive modelling as a preventive action. The ultimate goal of the model is to predict and significantly reduce employee turnover[3].

There are several conditions that can be used as a reference of Human Resource Department (HRD) to predict their employees will be leave of the company. From these predictions, the HRD can do preventive action for the most valuable company's assets are not out of the company.

2 Descriptive Analytic

Usually, the first activity that can be used for analytics is descriptive analytics. This is the starting point of further more complex analytics like predictive and prescriptive. The preliminary stage of data processing is descriptive analytic. it generates a summary of historical data to produce useful information.

Descriptive analytics is data simplification[4]. Descriptive analytic is a field of statistics focusing on gathering and summarizing raw data to be easily interpreted. It can be used to analyze the past event based on the available data and get insight about how to deal with the future.

In contrast to predictive models which focus on predicting a single customer behavior, descriptive models can identify many different relationships between customers or products. Essentially, in descriptive analytic, seeking answer about what happened can be done easily without performing complex analysis like in diagnostic and predictive models.[5] Descriptive analytics can be used to analyzed the reasons behind past failure or success by mining its historical data and studies its performance. There are many management reporting data available, such as, marketing, operation, finance, and sales that uses this type of analysis. If these available data used efficiently, it will improve business performance. Classifying or prospecting costumer into groups can also be done by quantifying relationship in data using descriptive models.[6]

The using of descriptive models is very broad these days, for example, to classify costumers by their specific references, such as, life stage, marital, or product references. For further development, descriptive modeling tools can be used to make prediction and simulate large number of individualized agents.[3]

3 Predictive Analytic

Predictive analytics uses historical data, statistical algorithm and Machine learning to predict future events. Predictive analytics turns data into actionable and valuable information. Predictive model is build based on historical data using mathematical model to capture its significant trends. Predictive analytics uses this trends to set on the probable future result of a prospect or an event of a currently happening situation.[7]

3.1 Naïve bayes

Naive Bayes is a collection of classification algorithms which are based on Bayes Theorem[8]. Probabilistic classifier is used in Naïve Bayes classifier. Statistical classifiers that used in Bayesian classifier can be used to predict the probability of a class membership, such as, probability that a given sample is a property of a specific class. Naive Bayes is also known for another name, Simple Bayes or independence Bayes.

In machine learning, a simple probabilistic classifier is used for Naive Bayes classifiers. Supervised learning method as well as a statistical method for classification is represented by Naive Bayes Classification. The model is a probabilistic model that permit to capture uncertainty about the model in a contentious way by determining probabilities[9].

Bayesian classification provides useful combination of observed data, past knowledge and learning algorithms[10]. The classification also provides a useful perspective for better understanding and also evaluating many learning algorithms. For hypothesis, this will helps to determine exact probabilities and also it is robust to noise in input data[11].

$$\rho(A|B) = \frac{\rho(B|A).\rho(A)}{\rho(B)} \quad (1)$$

$\rho(A|B)$ is conditional probability of an event A occurred given the event B is true, $\rho(B|A)$ is probability of the event B occurred given the event A is true and $\rho(A)$ and $\rho(B)$ is probabilities of event A and B happened respectively.[12]

4 Experimental Design

There are several stage will be conducted in this research. The Stage are mainly consisting of two main stages, data preparation stage and testing data stage. The rest of the stage can be seen in Figure 1.

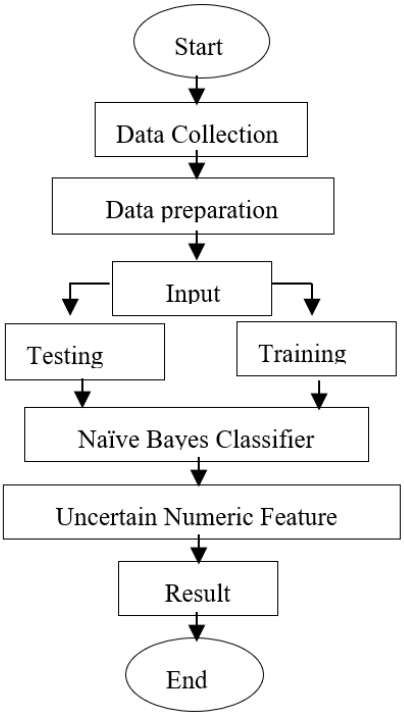

Fig. 1. Research Flowchart

The initial stage starts from data collection phase. The data used in this research are collected from open data at Kaggle.com. The raw data acquired will be processed in advance at data preparation stage. In This stage, the raw data format will be adjusted to a readable format for java programming. The results of this stage will be used as input data in the model.

In the training and testing stage, the random data will be choosing for Testing and the rest of the data is used for training. Training data is used to form the initial classification model which will then be tested with test data[13][14].

5 Experimental Analyst

5.1 Dataset

On this project, we use "Human Resource Analytic" dataset from https://www.kaggle.com. This dataset is simulated by a company who studied about "Why their best and most experienced employees are leaving prematurely". With this database, we can decide what the most valuable parameters are, and then we will be able to predict which valuable employees will possibly leave the company next.

There are ten parameters in this dataset:

1. Satisfaction level, indicates the satisfaction level of the employee.
2. Last evaluation, indicates the data about achievement from the last evaluation.
3. Number of projects, indicates the number of projects taken by an employee that assign by company.
4. Average monthly hours, indicates the average monthly hours of work provided by the company.
5. Time spent at the company, the total time the worker works at the company.
6. Work accident, indicates whether the employee has an accident. The value for this parameter is binary (1 = yes, 0 = no).
7. Promotion last 5 years, indicates whether the employee has a promotion in last 5 years. The value for this parameter is binary (1 = yes, 0 = no).
8. Departments (column sales), indicates the department where the employee work. The value for this parameter is categorical which is consist of sales, technical, marketing, support, management, accounting, product manager, accounting, and HR.
9. Salary, indicates The Employee salary level based on their works. The value for this parameter is categorical which is consist of high, medium, and low.
10. Left, indicates whether the employee has left the company. The value for this parameter is binary (1 = yes, 0 = no)

	A	B	C	D	E	F	G	H	I	J
1	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	departments	salary
2	0.38	0.53	2	157	3	0	1	0	sales	low
3	0.8	0.86	5	262	6	0	1	0	sales	medium
4	0.11	0.88	7	272	4	0	1	0	sales	medium
5	0.72	0.87	5	223	5	0	1	0	sales	low
6	0.37	0.52	2	159	3	0	1	0	sales	low
7	0.41	0.5	2	153	3	0	1	0	sales	low
8	0.1	0.77	6	247	4	0	1	0	sales	low
9	0.92	0.85	5	259	5	0	1	0	sales	low
10	0.89	1	5	224	5	0	1	0	sales	low
11	0.42	0.53	2	142	3	0	1	0	sales	low
12	0.45	0.54	2	135	3	0	1	0	sales	low
13	0.11	0.81	6	305	4	0	1	0	sales	low
14	0.84	0.92	4	234	5	0	1	0	sales	low
15	0.41	0.55	2	148	3	0	1	0	sales	low

Fig. 2. The original dataset

5.2 Prediction and implementation

The main goal is to predict whether an employee would leave or stay. From the dataset and the main goal, it clearly shows that there are two classes which are “leave” and “stay”, and each class has nine features which are the other parameters in the dataset.

To implement the system, we created a simple project using Java programing language with Jetbrain IntellijIDEA IDE. And there is no specific structure of the project packages. However, in order to make it easy to manage and understand, we created two different packages which one of them is for the training set or data, and another is for the classes of Java.

We set up the training sets by creating a file named “TrainingSets.txt” which is basically the dataset in text format, and inside, there are many lines of dataset in which each line consists nine features and the defined class “leave” or “stay” at the end of

the line like in the fig. 2. below. Be noted that the column of G (left) in fig. 1. will be shifted to the last order and all values of 0 are changed to "stay" and 1 to "leave".

```
0.45 0.69 5 193 3 0 0 sales low stay
0.78 0.82 5 247 3 0 0 sales low stay
0.49 0.6 3 214 2 0 0 sales low stay
0.36 0.95 3 206 4 0 0 sales low stay
0.54 0.37 2 176 2 0 0 sales low stay
0.99 0.91 5 136 4 0 0 sales low stay
0.5 0.75 6 127 3 0 0 sales low stay
0.74 0.64 4 268 3 0 0 sales low stay
0.56 0.58 4 258 3 0 0 sales medium stay
0.34 0.39 2 136 3 0 0 sales medium stay
0.48 0.94 5 255 6 0 0 accounting medium stay
0.73 0.62 3 218 3 0 0 accounting medium stay
0.59 0.87 3 268 4 0 0 accounting medium stay
0.81 0.57 3 224 2 0 0 hr medium stay
0.9 0.66 3 231 3 0 0 hr medium stay
```

Fig. 3. The training set of the dataset

Naive Bayes classifier assumes that all the features are unrelated to each other. The presence or absence of a feature does not influence the presence or absence of any other feature, and this is the reason why Naive Bayes works well with any type of training set with specific classes.[12] However, it still could lead to inaccuracy with any feature which have uncertain value representing as various numbers, or in the other words, the uncertain numeric features need to be classified and grouped in order to provide high accuracy.

There are five uncertain numeric features, and in order to fix the issue above, we find the minimum and maximum of each of numeric feature, and then classify into four different levels or quarters of its rank. Those five uncertain numeric features are classified as satisfaction_level, last_evaluation, number_profect, average_monthly, time_spent. The complete data for those five uncertain numeric features is shown in the table 1 below.

Table 1. Uncertain numeric features classification

Feature	Min	Max	Q1	Q2	Q3	Q4
Satisfication_level	0	1	0-0.24	0.25-0.49	0.5-0.74	0.75-1
Last_evaluation	0	1	0-0.24	0.25-0.49	0.5-0.74	0.75-1
Number_projecr	2	7	0-1	2-4	5-7	8-10
Average_monthly	96	310	75-149	150-224	225-299	300-375
Time_spent	2	10	0-1	2-4	5-7	8-10

Last but not least, our training set (Fig. 2) does not contain any note or mark to tell which feature each value belongs to; therefore, the system needs to differentiate the values between features by indexing the orders of the values of each line for example, "0.91 0.68 3 218 3 1 0 accounting medium" will be marked as "0-q4 1-q3 2-q2 3-q2 4-q2 4-y 6-n accounting medium".

Note:

1. For features “Whether they have had a work accident” and “Whether they have had a promotion in the last 5 years”, their values are marked as “y” if they equal to “1”, and as “n” if they equal to “0”.
2. If any value is unknown or not is in any group of the classify, it will be marked as “u”.

6 Result and Discussion

The system designed in this research based of java programming. The implemented system does not have any user interface. It requires command line Windows to test the data. Input data test can be put as the argument after the jar file, and the result of prediction will be show after.

1. Test 1: Test whether the program is working well. Input: “0.91 0.68 3 218 3 1 0 accounting medium” from Class “stay” line 3615. The data will be processed in the model and the result is correct as shown as below.

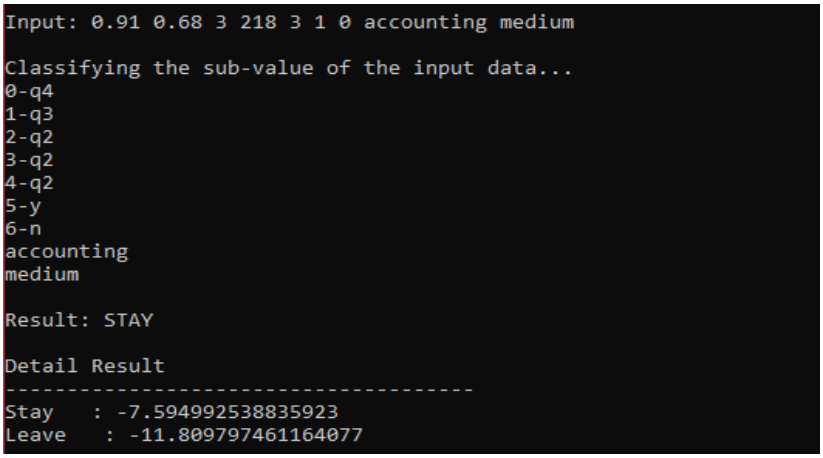

Fig. 4. Test 1

2. Test 2: Test an input which does not exist in the training set but closely similar to line 3178 from class “stay”. Input: “0.55 0.85 6 210 4 0 0 support medium” The result is expected to be “STAY”, and it does come as expected. The result is shown below.

```
Input: 0.55 0.85 6 210 4 0 0 support medium

Classifying the sub-value of the input data...
0-q3
1-q4
2-q3
3-q2
4-q2
5-n
6-n
support
medium

Result: STAY

Detail Result
-------------------------------------
Stay    : -6.171852538835923
Leave   : -8.884927461164077
```

Fig. 5. Test 2

3. Test 3: Test an incorrect or incomplete input. Input: "0.85 6 210 4 0 0 support medium". The system will check the input and give some advices for the input. The result is shown below.

```
Input: 0.85 6 210 4 0 0 support medium

Classifying the sub-value of the input data...
The input is not correct.
Please check your input again!
Make sure that there are 9 features in their orders.
If any feature is unknown, please just add -1.
Be noted that there is empty space before and after the quote.

Result: null

Detail Result
-------------------------------------
Stay    : null
Leave   : null
```

Fig. 6. Test 3 with incorrect input

The system asked to put add "-1" for any unknown features. So the input should be corrected as "-1 0.85 6 210 4 0 0 support medium". And the result is shown as in the fig. 6.

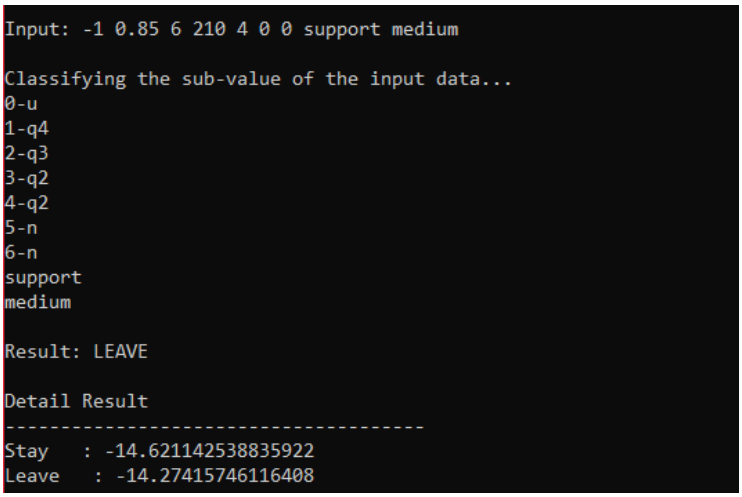

Fig. 7. Test 3 after corrected the input

7 Conclusion

The accuracy of the result is depended on the amount and effectiveness of the training sets. Naive Bayes classifier assumes that all the features are unrelated to each other. The presence or absence of a feature does not affect the other features, and this is the reason why Naive Bayes works very well with any type of training set with specific classes. However, it still could lead to inaccuracy with any features which have uncertain value representing as various numeric numbers, or in the other words, the uncertain numeric features need to be classified and grouped in order to provide high accuracy.

8 References

[1] D. P. R.Shiva, "Customer behavior analysis using Naive Bayes with bagging homogeneous feature selection approach," J. Ambient Intell. Humaniz. Comput., no. Published on Springer, 2020. https://doi.org/10.1007/s12652-020-01961-9

[2] M. V. Sebt, E. Komijani, and S. S. Ghasemi, "Implementing a Data Mining Solution Approach to Identify the Valuable Customers for Facilitating Electronic Banking,". International Journal of Interactive Mobile Technologies (iJIM). vol. 14, no. 15, 2020, pp. 157–174. https://doi.org/10.3991/ijim.v14i15.16127

[3] Z. Jaffar, "Predictive Human Resource Analytics Using Data mining Classification Techniques," Int. J. Comput., vol. 32, no. 1, pp. 9–20, 2019.

[4] S. Loeb, S. Dynarski, D. McFarland, P. Morris, S. Reardon, and S. Reber, "Descriptive analysis in education: A guide for researchers," U.S. Dep. Educ. Inst. Educ. Sci. Natl. Cent. Educ. Eval. Reg. Assist., no. March, pp. 1–40, 2017.

[5] N. Ardi and Isnayanti, "Structural Equation Modelling-Partial Least Square to Determine the Correlation of Factors Affecting Poverty in Indonesian Provinces," IOP Conf. Ser. Mater. Sci. Eng., vol. 846, no. 1, 2020. https://doi.org/10.1088/1757-899x/846/1/012054
[6] J. Strickland, Predictive Analytics using R. Lulu, 2015.
[7] A. Fuentes, Hands-On Predictive Analytics with Python: Master the complete predictive analytics process, from problem definition to model deployment. Packt, 2018.
[8] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," pp. 266–270, 2020. https://doi.org/10.1109/smart46866.2019.9117512
[9] S. P. Huma Parveen, "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm," 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol., 2017.
[10] I. M. Obeidat, N. Hamadneh, M. Alkasassbeh, M. Almseidin, and M. I. Alzubi, "Intensive Pre-Processing of KDD Cup 99 for Network Intrusion Classification Using Machine Learning Techniques,". International Journal of Interactive Mobile Technologies (iJIM). Vol 13. No.1, 2019. https://doi.org/10.3991/ijim.v13i01.9679
[11] S. P. Sujata Butte, "Big Data and Predictive Analytics Methods for Modeling and Analysis of Semiconductor Manufacturing Processes," Microelectron. Electron Devices, no. IEEE Workshop on, 2016. https://doi.org/10.1109/wmed.2016.7458273
[12] S. Kaliya Meiyyar, "The Comparative Study for Diagnosing Heart Disease Using KKN and Naïve Bayes," Int. J. Adv. Res. Comput. Sci. Manag. Stud., vol. 3, no. 8, 2015.
[13] N. Ardi, N.A Seatiawan and T.B Adji "Analytical Incremental Learning for Power Transformer Incipient Fault Diagnosis Based on Dissolved Gas Analysis," pp. 3–6, 2020. https://doi.org/10.1109/icst47872.2019.9166441
[14] A. Huda, N. Azhar, K. Anshari, and S. Hartanto, "Practicality and Effectiveness Test of Graphic Design Learning Media Based on Android,". International Journal of Interactive Mobile Technologies (iJIM). Vol.14. No.4, 2020 pp. 192–203. https://doi.org/10.3991/ijim.v14i04.12737

9 Authors

Asrul Huda works in Universitas Negeri Padang in Indonesia. Email: asrulhuda@gmail.com

Noper Ardi works for Universitas Negeri Padang in Indonesia. Email: noper.ardi@gmail.com

Article submitted 2021-01-03. Resubmitted 2021-02-26. Final acceptance 2021-02-26. Final version published as submitted by the authors.