Predicting Employee Turnover in IT Industries using Correlation and Chi-Square Visualization

Bagus Priambodo[1], Yuwan Jumaryadi[2], Sarwati Rahayu[3], Nur Ani[4], Anita Ratnasari[5],
Umniy Salamah[6], Zico Pratama Putra[7], Muhamad Otong[8]
Faculty of Computer Science, Universitas Mercu Buana, Jakarta, Indonesia[1, 2, 3, 4, 5, 6]
Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia[7]
Department of Electrical Engineering, Universitas Sultan Ageng Tirtayasa, Serang, Indonesia[8]

***Abstract*—Employee turnover in the IT industry is among the highest compared to other industries. Knowing factors that influence the turnover may help reduce this issue in future. One of these factors is job satisfaction, which are composed of two important factors, status and seniority. In this study, the correlation and chi-square visualization are utilized to determine the factors that affect employee turnover. The experiment was carried out to predict turnover using a private IT consultant dataset comparing three classification algorithms (decision tree, Naïve Bayes, and Random Forest). The result shows that job duration and positioning are factors that influence employee turnover in a software company.**

Keywords—Employee turnover; turnover factors; chi-square; classification algorithm

I. Introduction

According to the information and communications technology (ICT) workers survey, 43.6% of ICT professionals work in IT Industries, with an annual growth rate of 14.7% [1]. Given the sharp increase in demand for skilled IT professionals, supply of qualified computer specialists, especially college graduates, is on the rise [2]. This situation has given the IT industry a competitive advantage in hiring an ever-growing number of professional software engineers, as each company tries to lure the best employees by offering them better opportunities in terms of pay, expertise, and working atmosphere. The IT industry therefore has one of the highest turnover rates in comparison to other industries [1].

Poor financial compensation, management style, and career opportunities are the main factors contributing to dissatisfaction with current employer and their decision to leave [3]. Staff turnover is a major disadvantage for the company on direct and indirect costs. The direct costs are related to the time and resources spent on recruiting, hiring and training employees [4]. The indirect costs include the decrease in production and services caused by the company failing to find replacements. Hiring and training new employees is also a large investment because the higher the turnover rate, the higher the cost [5].

In general, turnover is due to employee dissatisfaction with a number of factors, including working conditions, salary, support quality, co-workers, nature of work, employment security, and career prospects. The factors that affect job satisfaction are divided into two factors, namely organizational factors - which include company policies and work environment - and individual factors [6]. The status and seniority are one of these individual factors that affect job satisfaction, as a perception that a lack of employment status causes many workers seeking another job [6]. In addition, attitudes toward human resource management indirectly affect turnover, which is fully mediated by job satisfaction [7].

In this paper, visualization and feature selection are proposed to identify factors that influence employee turnover. The paper is organized as follows. In Section II, this paper first gives an overview of previous work on this topic. This is followed by a presentation of the method used in Section III, followed by a discussion of the experimental results in Section IV. Finally, the paper is summarized in Section V.

II. Related Work

Some turnover studies suggest using machine learning to design retention policies [8]. A study was carried out by [9] to predict employee turnover in one of Indonesia's renowned telecommunication company, comparing three classification algorithm (Naïve Bayes, Decision Tree and Random forest) based on 12 attributes. Another study by [10] implemented twelve features to predict employee turnover, while other use 19 attributes [8]. Similar study by [11] evaluate six classification algorithm (KNN, SVR, Naïve Bayes, Decision Tree, Random forest) to predict employee turnover using a dataset from Kaggle. In contrast to the previous study, this study will evaluate the individual and seniority factors, which are represented by length of service and job level as factor of employee turnover in IT consultant firm.

A. *Naïve Bayes*

In supervised machine learning [12][13][14][15], Naïve Bayes is a widely used model of classification due to its simplicity and efficiency. Naïve Bayes computes the posteriori probability for a class by observing the churn problem, i.e., the churn and non-churn observation probabilities. The posteriori probabilities are computed by using the rule of Bayes and the assumption of naive Bayesians with unique employee backgrounds for each class. The goal of the Bayesian decision rule is to set a record of a new hire to the class with the highest probabilities posteriori as [16] suitable model for predicting employee turnover.

B. Decision Tree

A decision tree is divided by decision rules into some classes [17]. It represents a fluctuating graph resembling a tree structure that implies a test in an attribute for each internal node (not leaf node). A branch represents a test result, with a class label for each leaf node (or end node). The root node is the top node in a tree [18]. The C4.5 decision tree classifier is employed to build predictive models due to its ability to segment dynamic decision processes [19].

C. Random Forest

Random Forest (RF) proposed by Brieman [20], uses aggregation and bootstrap ideas to introduce random forests based on a decision tree [21]. The trees are constructed separately in RF, using bootstrap samples of various data sets. The concept of RF is to construct multiple decision trees using only a subset of attributes based on sample data [16]. Predictions are taken by taking a simple majority decision by dividing the individual nodes into a subset of predictors using the best distribution. At that node, the predictors are selected randomly and this randomness makes it robust against overfitting [22][23].

III. Methodology

In this section, a proposed framework is presented, which includes preprocessing, relationship visualization, chi-square calculation, correlation, and prediction (Fig. 1).

A. Sourcing and Pre-processing

This is the initial phase of the framework. The purpose of preprocessing the data is to convert the raw data into an easier and more efficient format to use for future processing steps. In the first phase, a min-max method is used to normalize the data. Normalization can shorten the training time as all the data used for training share the same scale. Collecting the dataset from private IT consultants between January 1, 2015 and December 31, 2018, the data consists of gender, entry and exit dates, job title, and department, as shown in Table I.

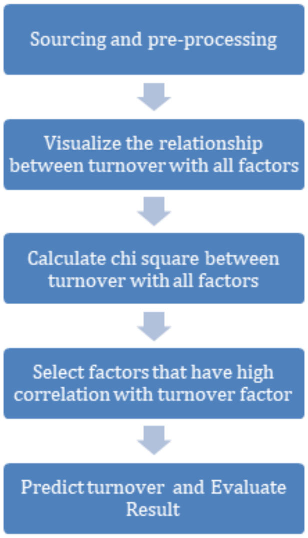
Sourcing and pre-processing
Visualize the relationship between turnover with all factors
Calculate chi square between turnover with all factors
Select factors that have high correlation with turnover factor
Predict turnover and Evaluate Result

Fig. 1. Framework of predicting employee turnover by identifying the two most influence factors in turnover, i.e., work duration and department.

TABLE I. Attribute Employee Turnover

No	Gender	Date of accession	Date of withdrawal	...	Job title
1	Male	01/12/2016	05/05/2017	...	Head of the Finance Department
2	Female	16/08/2017	29/03/2018	...	Accounting officer
3	Male	01/03/2016	14/01/2017	...	Driver
4	Female	19/09/2016	14/07/2017	...	Recruiter
149	Male	01/07/2015	-	...	Functional consultant
150	Male	01/07/2015	-	...	Software platform manager

B. Visualization of Correlation between Turnover and All Factors

The correlation across pairs of time series indicates the degree to which the variance in one time series may be the same as the variance in the other. This alone does not imply a causal relationship between the two series. However, this is generally grounds for further investigation of a possible relationship. Relationships of correlation are often expressed as values between -1 and 1, figures close to 1 indicating a high correlation, figures close to 0 indicating a low relationship, and figures close to -1 indicating an inverse relationship. An analysis is made of the correlation between turnover and all factors. The visualization is selected to analyze the most correlated factor with the resignation factors.

C. Calculation of the Chi-Square between Turnover and All Factors

For classification problems having both categorical input and output, a chi-square test ("chi") is useful in determining how relevant the input variables are to the output variable. In this study, the correlation between turnover and all factors is analyzed using chi-square, calculating the *p-value* to determine which factor has the lowest *p-value*. The formula for the chi-square is shown on (1).

$$X^2 = \sum \frac{(O-E)^2}{E} \quad (1)$$

Σ means summarize

O = any observed (actual) value

E = any expected value.

D. Predict Turnover

In classification, a mean value of grouping data based on a label or target class is calculated in the study. The three classification algorithms are used to compare the impact of the proposed framework on classification performance. Using Decision Tree, Naive Bayes, and Random Forest as pilot methods, all of them are used to predict turnover based on all factors. Next, a forecast of sales based on selected factors is also made using these three algorithms. Both results are then compared to determine which has the better performance.

IV. Result and Discussion

A. Pre-Processing

Experiments are conducted using several machine learning algorithms with additional tests on the preprocessed data and feature selection to achieve better accuracy. The preprocessed data for job level, department and gender category are converted into numbers as shown in Table II.

B. Visualize the Correlation between Turnover and All Factors

The correlation between gender, department and degree of resignation is visualized in a bar chart (Fig. 2). The correlation between duration and turnover factor (resignation) is visualized by a boxplot. As shown in Fig. 3, gender does not have a significant impact on employee turnover. The trend of gender is similar for terminated and non-terminated employees.

TABLE II. Data Set after Pre-processing

No	Duration (month)	Job level	Department	Gender
1	5	5	13	1
2	7	1	1	0
3	10	1	8	1
4	9	1	17	0
...	...	...	...	...
149	42	1	14	1
150	42	5	41	1

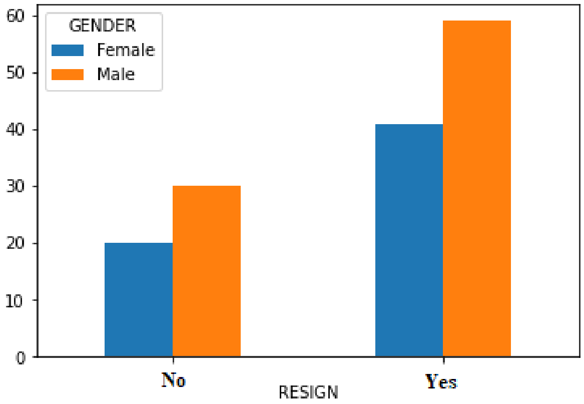

Fig. 2. The relationship between gender and the turnover Factor (resignation).

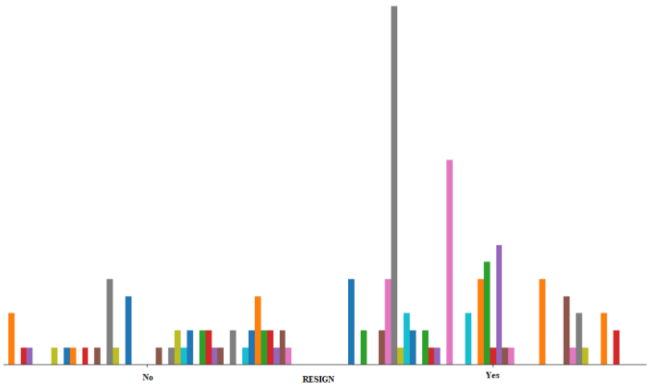

Fig. 3. The relationship between the department and the factor of turnover (resignation).

Fig. 3 shows the relation between the department and the turnover factor and reveals that department factors have a significant impact on employee turnover. The trend of the employee leaves and the employee stays are different from various departments.

Fig. 4 shows the relationship between the workplace level and the turnover factor. It indicate that the workplace level has no significant influence on employee turnover. Further, the trend of the job level is similar for terminated and non-terminated employees.

However, the boxplot relationship between duration and turnover (resignation) factor from Fig. 5 shows that the length of employment (duration) has a significant impact on employee turnover.

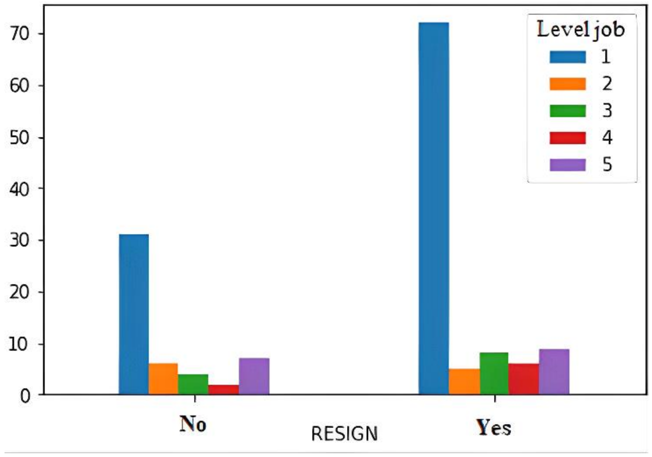

Fig. 4. The relationship between the workplace level and the turnover factor (resignation).

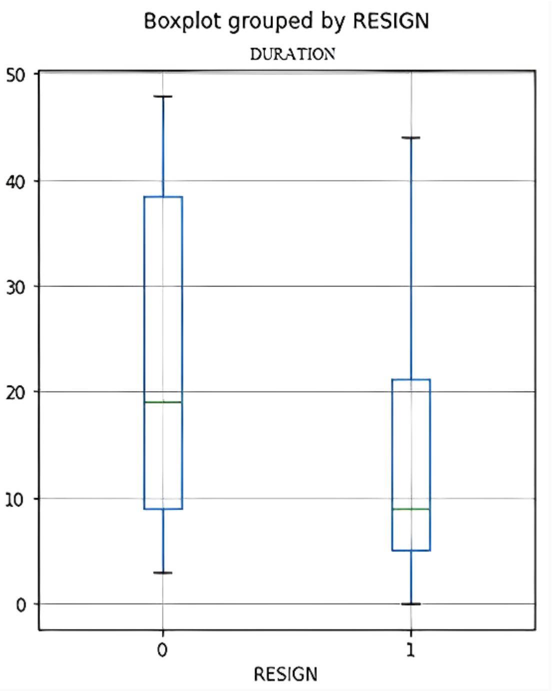

Fig. 5. The relationship between duration and turnover (resignation) factor.

C. Calculate the Chi-Square between Turnover and All Factors

The term statistically significant has become synonymous with a *p-value* <= 0.05. The table of *p-value*s shows that the *p-value* for duration (length of work) and department is less than 0.05, as shown in Fig. 6. Table III and Fig. 6 show that duration of work (length) and department are factors that affect employee turnover.

D. Turnover Forecast and Result Evaluation

The dataset is divided by 80% for training data and 20% for testing data. The result is shown in Table IV. The result of prediction with all factors is shown in the first column (with all factors) and the result of prediction with only selected factors is shown in the second column (only selected factors). Based on the visualization of the correlation between the turnover and all factors and the calculation of the chi-square. Duration and department factor have high correlation with employee turnover. The prediction result shows that the prediction using only selected factors has better performance compared to the prediction using all factors.

TABLE III. P-VALUE BETWEEN TURNOVER AND ALL FACTORS

No	Factors	*p-value*
1	Gender	0.94
2	Duration	**<0.01**
3	Level Job	0.36
4	Department	**<0.01**

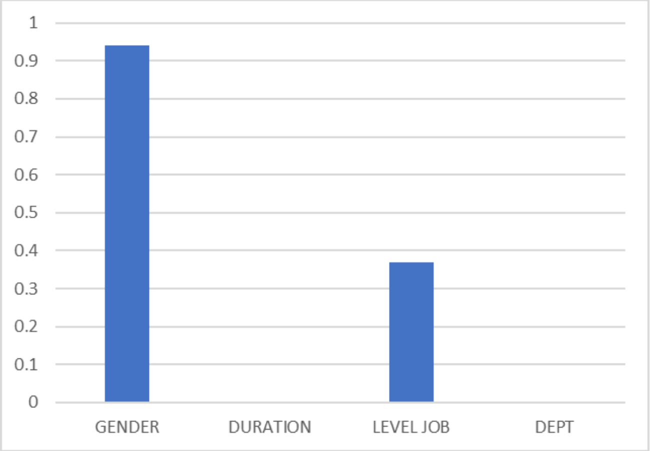

Fig. 6. Bar chart of the p-value between turnover (resignation) and all factors.

TABLE IV. THE PREDICTION RESULT OF THE EMPLOYEE TURNOVER

	With all factors	Selected factors only
Random forest	0.833	0.867
Naïve Bayes	0.77	0.77
Decision tree	0.83	0.90

V. CONCLUSIONS

In this study, correlation and chi-square visualization are applied. Its key idea is based on a two-step process. A correlation between pairs of time series is first used for further investigation of possible relationships. Secondly, Chi-Square is used to determine how relevant the input variables are to the output variables. By forecasting employee turnover - compared to Decision tree, Naïve Bayes, and Random forest - predictions using only selected factors (duration of employment and department) indicated a notable improvement in various performance parameters compared to predictions using all factors. Using a dataset of IT consulting employees in Indonesia, the key factors affecting employee turnover were identified using the chi-square algorithm; specifically, years in the company, job level, department placement, and gender as the most important factors. Therefore, duration of employment and department are unarguably among the most influential factors for retaining employees.

By predicting upfront, the likely turnover of employees, companies are able to prepare plans and initiate measures to reduce that likelihood, hire replacements swiftly, and make other adjustments aimed at retaining individuals in important positions. By using correlation and chi-square visualization approaches, HR managers may anticipate employee turnover better and take action in a timely manner. Similar applications of this algorithm are possible for imbalanced data classification problems such as predicting customer attrition, preventive cancer care, and anomaly detection. These issues share common characteristics, like a focus on minority classes and the condition that "it is preferable to test wrong than to detect wrong." Future work would be to analyze theoretically the characteristics of the data in these areas and to verify and optimize the approach further.

Due to the limitations of the study, the extra process of feature selection and weight calculation implies that the cost of modeling is higher in time compared to other algorithms. Improvements to the operational efficiency and accuracy of the entire prediction can be analyzed in future research. Furthermore, for industries with high turnover rates, the algorithm is based on unbalanced data and therefore may not be suitable. Details on increasing the universality of the algorithm are yet to be studied.

ACKNOWLEDGMENT

This work was supported by Universitas Mercu Buana through KDN Research Grant Scheme [02-5/1355/B-SPK/VI/2022].

REFERENCES

[1] L. Jinadasa and V. Wickramasinghe, “IT Industry Labour Turnover : The Reality,” 10th Int. Conf. Sri Lanka Stud., no. December 2005, pp. 0–10, 2005.

[2] Sasmoko, Y. Indrianti, S. A. Widhoyoko, Y. Udjaja, and U. Rosyidi, “Performance Change with or Without ITEI Apps,” in The 6th International Conference on Cyber and IT Service Management (CITSM 2018), 2018, pp. 1–4, doi: 10.1109/CITSM.2018.8674055.

[3] R. Korsakienė, A. Stankevičienė, A. Šimelytė, and M. Talačkienė, “Factors driving turnover and retention of information technology professionals,” J. Bus. Econ. Manag., vol. 16, no. 1, pp. 1–17, 2015, doi: 10.3846/16111699.2015.984492.

[4] I. Jeffrey and A. B. Prasetya, “the Employee Performance Assessment and Employee Training, on Employee Intension,” J. Apl. Manaj., vol. 17, no. 1, pp. 56–65, 2019, doi: 10.21776/ub.jam.2019.017.01.07.

[5] R. Ameliya and H. Febriansyah, “The significant factors of employee turnover case study : ABC Hotel,” J. Bus. Manag., vol. 6, no. 2, pp. 239–249, 2017.

[6] Hojops Odoch and S. Nangoli, “The effect of organizational commitment on job satisfaction in Uganda Colleges of Commerce,” Issues Bus. Manag. Econ., vol. 2, no. 10, pp. 165–171, 2014.

[7] C. Maier, S. Laumer, A. Eckhardt, and T. Weitzel, “Analyzing the impact of HRIS implementations on HR personnel’s job satisfaction and turnover intention,” J. Strateg. Inf. Syst., vol. 22, no. 3, pp. 193–207, 2013, doi: 10.1016/j.jsis.2012.09.001.

[8] E. Ribes, K. Touahri, and B. Perthame, “Employee turnover prediction and retention policies design: a case study,” 2017, [Online]. Available: http://arxiv.org/abs/1707.01377.

[9] A. Alamsyah and N. Salma, “A Comparative Study of Employee Churn Prediction Model,” Proc. - 2018 4th Int. Conf. Sci. Technol. ICST 2018, vol. 1, no. 2, pp. 1–4, 2018, doi: 10.1109/ICSTC.2018.8528586.

[10] S. Yadav, A. Jain, and D. Singh, “Early Prediction of Employee Attrition using Data Mining Techniques,” Proc. 8th Int. Adv. Comput. Conf. IACC 2018, pp. 349–354, 2018, doi: 10.1109/IADCC.2018.8692137.

[11] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, “Evaluation of machine learning models for employee churn prediction,” Proc. Int. Conf. Inven. Comput. Informatics, ICICI 2017, no. Icici, pp. 1016–1020, 2018, doi: 10.1109/ICICI.2017.8365293.

[12] I. Nurhaida et al., “Implementation of Deep Learning Predictor (LSTM) Algorithm for Human Mobility Prediction,” Int. J. Interact. Mob. Technol., vol. 14, no. 18, pp. 132–144, 2020, doi: 10.3991/ijim.v14i18.16867.

[13] A. Nugroho, H. L. H. S. Warnars, S. M. Isa, and W. Budiharto, “Comparison of Binary Particle Swarm Optimization And Binary Dragonfly Algorithm for Choosing the Feature Selection,” in 2021 5th International Conference on Informatics and Computational Sciences (ICICoS), 2021, pp. 24–28, doi: 10.1109/ICICoS53627.2021.9651779.

[14] B. Priambodo, A. Ahmad, and R. A. Kadir, “Predicting Traffic Flow Propagation Based on Congestion at Neighbouring Roads Using Hidden Markov Model,” IEEE Access, vol. 9, pp. 85933–85946, 2021, doi: 10.1109/ACCESS.2021.3075911.

[15] B. Jokonowo, M. A. Kulsum, and N. Komala, “Perbandingan Model Proses Algoritma Alpha dan Alpha++ Pada Aplikasi E-commerce,” J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 6, no. 1, pp. 123–129, 2022, doi: 10.29207/resti.v6i1.3732.

[16] V. V. Saradhi and G. K. Palshikar, “Employee churn prediction,” Expert Syst. Appl., vol. 38, no. 3, pp. 1999–2006, 2011, doi: 10.1016/j.eswa.2010.07.134.

[17] T. M. Mulyono, F. Natalia, and S. Sudirman, “A study of data mining methods for identification undernutrition and overnutrition in obesity,” ACM Int. Conf. Proceeding Ser., pp. 6–10, 2019, doi: 10.1145/3374549.3374565.

[18] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd Editio. San Fransisco: Morgan Kaufmann Publisher, 2014.

[19] H. Jantan, A. R. Hamdan, and Z. A. Othman, “Human Talent Prediction in HRM using C4 . 5 Classification Algorithm,” Int. J. Adv. Trends Comput. Sci. Eng., vol. 02, no. 08, pp. 2526–2534, 2010.

[20] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1201/9780429469275-8.

[21] R. Genuer, J. Poggi, C. Tuleau-malot, and N. Vialaneix, “Random Forests for Big Data,” Big Data Res., vol. 9, pp. 28–46, 2017.

[22] 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.

[23] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, “Students’ Orientation Using Machine Learning and Big Data,” Int. J. online Biomed. Eng., vol. 17, no. 1, pp. 111–119, 2021, doi: 10.3991/ijoe.v17i01.18037.