Proceedings of Science and Mathematics
Faculty of science, universiti teknologi malaysia

https://science.utm.my/procscimath
Volume 21 (2024) 190-198

Comparative Study Using Kaplan-Meier and Cox Models in Analyzing Employee Attrition

Muhammad Izzat Irfan Ahmad, Zarina Mohd Khalid*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: zarinamkhalid@utm.my

Abstract

This research explores into the field of human resource management, focusing on the critical issue of employee attrition. The departure of employees not only disrupts operational dynamics but also incurs substantial costs associated with recruitment and training. The primary aim of this study is to explore the efficacy of survival analysis models, specifically the Kaplan-Meier (KM) estimator and the Cox Proportional Hazards (CPH) model, in comprehending and predicting employee turnover. Methodologically, the research employs the KM model for univariate analysis, presenting step-function survival curves, and the CPH model for multivariate analysis, providing a nuanced understanding of factors influencing attrition. The investigation entails utilizing KM models to dissect individual factors contributing to attrition and conducting log-rank tests to discern significant disparities based on specific variables. Additionally, the CPH model is employed to scrutinize the cumulative effects of various factors on attrition. Python libraries and statistical packages are leveraged for data processing and model implementation. The results reveal several covariates significantly influencing employee retention survival, including age, grey wage, self-control, anxiety, and industry sectors such as technology, services, and consulting. Additionally, the mode of transportation to work, method of job application, and employer contact via job sites or online portals play crucial roles in attrition prediction. Specifically, the CPH model identifies significant covariates, with confidence intervals indicating the range of their impact. For instance, age (95% C.I: 0.01~0.03), grey wage (95% C.I: 0.29~0.79), and self-control (95% C.I: -0.11~-0.02) are among the influential factors. In response to these findings, preventive actions such as proactive agediverse hiring practices, fair wage structures, and comprehensive mental health wellness programs are recommended. Additionally, targeted interventions tailored to specific industries and transportation modes can address unique challenges and promote job satisfaction. By addressing these factors, organizations can mitigate the risk of employee turnover and enhance overall organizational success.

Keywords: s Employee attrition; Kaplan-Meier model; Cox Proportional Hazards model; survival analysis; human resource management; predictive analytics; retention strategies

Introduction

Employee attrition, or turnover, is a significant challenge for organizations, impacting their operational efficiency, financial stability, and overall productivity. High attrition rates lead to increased costs related to recruitment, training, and loss of institutional knowledge. Additionally, frequent turnover disrupts team dynamics and lowers employee morale. Therefore, predicting employee attrition is crucial for developing effective retention strategies and ensuring organizational stability.

One effective method to analyze and predict employee attrition is survival analysis. Survival analysis is particularly useful for time-to-event data, where the event of interest is the termination of employment. This method can provide insights into the timing and likelihood of employee departures, which are essential for proactive talent management.

This study employs two prominent survival analysis models: the Kaplan-Meier (KM) estimator and the Cox Proportional Hazards (CPH) model. The KM estimator is a non-parametric statistic that estimates survival probabilities over time and provides clear visual representations of survival functions. It is particularly effective for univariate analysis, where the impact of a single factor on survival can be examined without considering other variables. In contrast, the CPH model is a semi-parametric model that allows for multivariate analysis, assessing the effect of multiple covariates on the time to an event. The CPH model provides hazard ratios that quantify the relative impact of each variable, making it valuable for understanding the combined effects of several factors on employee attrition.

Previous studies have applied various methods to analyze employee turnover, such as regression models, decision trees, and machine learning techniques. However, survival analysis offers a distinct advantage by incorporating the time dimension, providing a more comprehensive understanding of turnover dynamics.

This research aims to (1) compare the efficacy of the KM and CPH models in analyzing employee attrition and (2) identify significant predictors of employee turnover. The study leverages a comprehensive dataset containing demographic, professional, and behavioral information of employees. By applying the KM estimator, the study visualizes survival probabilities and identifies significant differences in attrition rates across various groups. Subsequently, the CPH model is employed for multivariate analysis, identifying key predictors and quantifying their impact on employee turnover.

Literature Review

Survival Analysis

Survival analysis predicts the time until an event occurs within a specific timeframe. Traditionally statistical, its applications now include machine learning approaches, enhancing prediction accuracy ([15]). It’s widely used in fields like medicine, computer science, and engineering ([1];[20]). For example, it predicts survival times for patients and user engagement in online communities ([2];[3]).

Kaplan-Meier Model

The Kaplan-Meier (KM) model estimates and graphs survival curves, useful in medicine, finance, and biology. In medicine, it estimates survival probabilities and handles incomplete data ([6];[19]). It’s used in cancer studies and implant failure analysis ([16],[17]). In finance, it designs profitable strategies (Sarwar et al., 2018). Developed in 1958, it remains a key tool for time-to-event data analysis (Etikan et al., 2017).

Cox Proportional Hazards Model

The Cox Proportional Hazards (CPH) model, or Cox regression, examines relationships between survival time and predictors. It’s extensively used in medical research for various conditions ([11];[12]). The model predicts survival probabilities while controlling for covariates ([15]) and is useful in gene selection for medical research ([20]). It remains robust across different survival predictions ([23]).

Log-Rank Test

The log-rank test compares survival distributions between groups, commonly used in medicine and clinical trials. It has shown effectiveness in studies like those on Spironolactone for heart failure (Galili et al., 2021). Efforts to improve it include addressing conservativeness in small samples and adaptively weighted versions for better performance ([10];[12]).

Employee Attrition

Employee attrition impacts organizational performance and budgets. Factors influencing turnover include job satisfaction, organizational commitment, and support ([18]). Attrition can result from resignations and retirements, leading to the loss of productive employees ([19]). Leadership, ethical climate, and corporate image significantly influence turnover intention ([16]). Advanced techniques like machine learning are being explored to predict attrition, showing growing interest in technology for this issue ([21]).

Methodology

Data Preparation

Data preparation is a crucial step in ensuring the dataset aligns with the requirements of the survival analysis models used in this study. The dataset consists of demographic, professional, and behavioral information of employees, focusing on factors influencing employee attrition. Key variables include age, gender, job role, department, salary, years at the company, and performance ratings. The dataset spans a period from January 2009 to April 2019.

Data preprocessing involves the following steps:

1) Cleaning: Removing any incomplete or inconsistent records to ensure data integrity.
2) Coding: Converting categorical variables into numerical formats suitable for analysis. For example, job roles and departments are assigned numerical codes.
3) Time-to-Event Variable: Calculating the time-to-event variable, which represents the duration of employment until attrition.
4) Event Indicator: Creating an event indicator variable, where '1' indicates employee attrition and '0' indicates continued employment.

Kaplan-Meier Analysis

The Kaplan-Meier (KM) model is used for univariate analysis to estimate and visualize survival probabilities over time. This method involves the following steps:

1) Estimation: Applying the KM estimator to calculate the survival function, S(t)S(t)S(t), which represents the probability that an employee will remain employed beyond time ttt.
2) Graphing: Plotting the KM survival curves for different groups based on categorical variables such as gender, job role, and department.
3) Comparison: Conducting log-rank tests to compare survival distributions between groups and identify significant differences.

The KM model provides clear visual representations of survival probabilities, helping to identify trends and patterns in employee attrition.

Cox Proportional Hazards Model

The Cox Proportional Hazards (CPH) model is employed for multivariate analysis to assess the impact of multiple covariates on the time to employee attrition. The methodology involves:

1) Model Fitting: Fitting the full CPH model to the dataset, including all relevant covariates.
2) Variable Selection: Identifying significant predictors based on their p-values. Variables with p-values less than 0.1 are considered significant.
3) Reduced Model: Creating a reduced CPH model using only the significant predictors identified in the previous step.
4) Hazard Ratios: Estimating hazard ratios for each predictor in the reduced model to quantify their impact on the hazard of employee turnover.

The CPH model's ability to handle multiple covariates simultaneously provides a comprehensive understanding of the factors influencing employee attrition.

Model Comparison

To compare the efficacy of the Kaplan-Meier and Cox Proportional Hazards models, the following steps are undertaken:

1) Survival Curves: Comparing the survival curves generated by the KM model for different groups.
2) Hazard Ratios: Analyzing the hazard ratios from the CPH model to understand the relative impact of each predictor.

3) Goodness-of-Fit: Assessing the goodness-of-fit for both models to evaluate their performance. This involves examining the concordance index (C-index) for the CPH model and visual inspection of the KM survival curves.

Results and discussion

Kaplan-Meier Analysis

The Kaplan-Meier (KM) survival curves were generated to visualize the survival probabilities of employees over time, considering various demographic and professional factors.

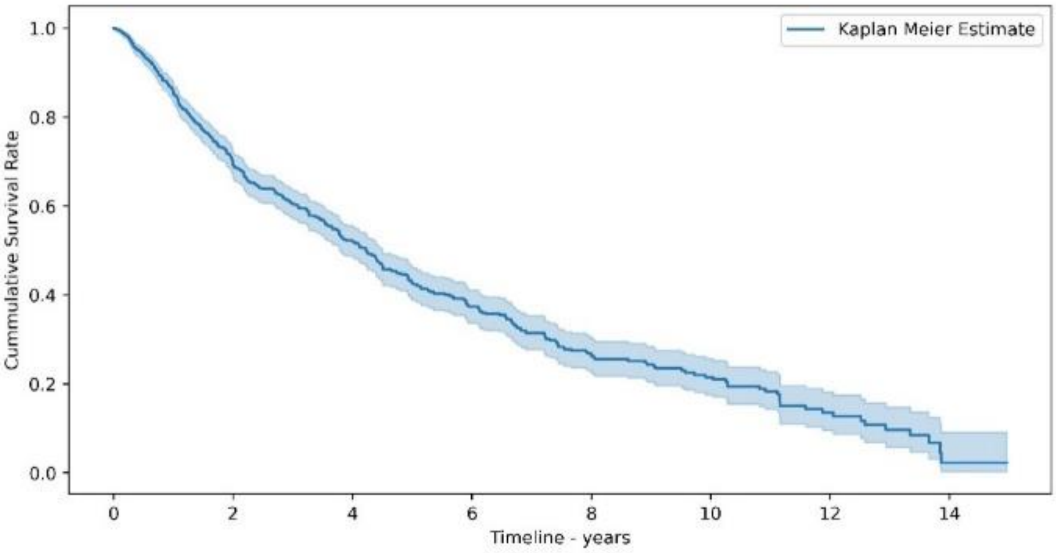

Figure 1 Survival Curve of Employee Attrition

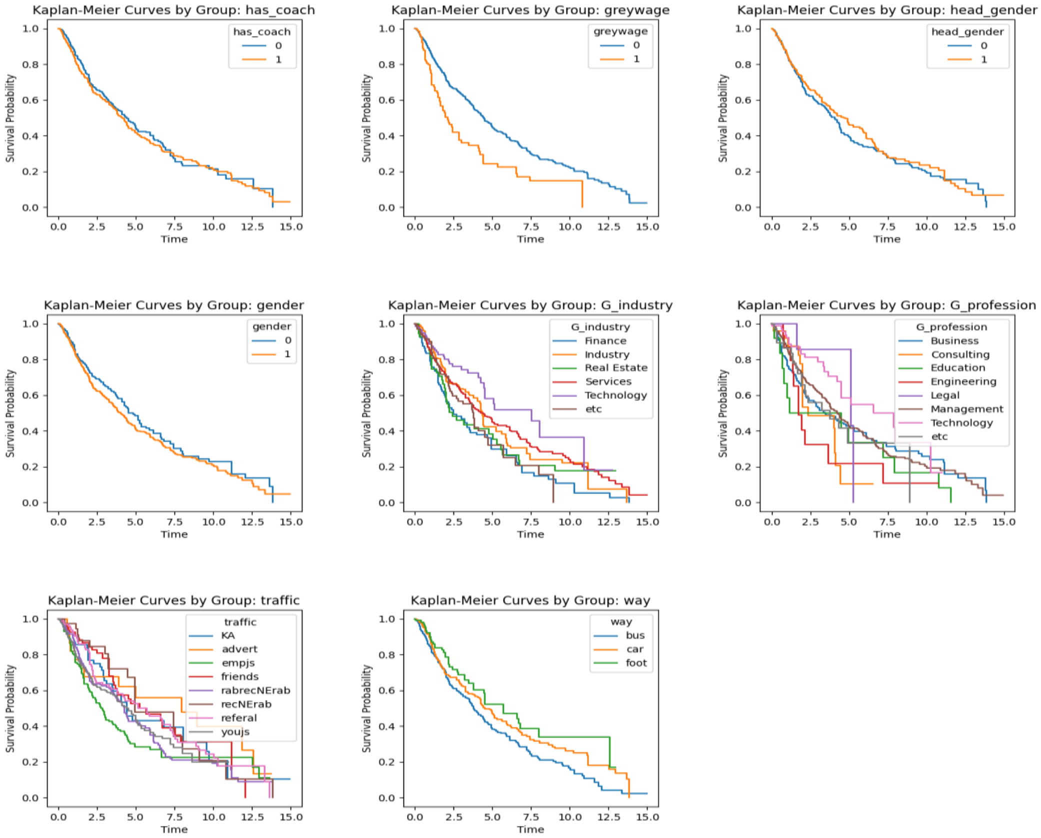

Figure 2 Kaplan-Meier Survival Curve for Categorical Variables

Gender

The KM survival curves for gender show distinct patterns. Female employees exhibit a slightly higher survival probability over time compared to male employees. The log-rank test indicates that this difference is statistically significant (p-value = 0.12545). Although not highly significant, this suggests a trend where gender might influence attrition rates.

Coaching

Employees with coaching show higher survival probabilities than those without coaching. The log-rank test reveals a p-value of 0.41493, indicating no significant difference between the two groups. This suggests that coaching, while beneficial, may not significantly impact overall attrition rates when considering the entire observation period.

Coach Gender

The KM survival curves for head gender indicate no significant differences in survival probabilities between employees supervised by male and female heads. The log-rank test yields a p-value of 0.266829, suggesting that the gender of the head does not significantly affect employee attrition rates.

Wage type

The survival curves for wage type (grey wage vs. white wage) show a pronounced difference. Employees with grey wages have a significantly lower survival probability compared to those with white wages. The log-rank test confirms this with a highly significant p-value of 0.000002. This finding underscores the importance of formalizing wage structures to improve employee retention.

Industry

The survival analysis across different industries reveals significant differences. Employees in the Technology industry exhibit the highest survival probabilities, whereas those in the Finance industry show the lowest. The log-rank tests for pairwise industry comparisons yield several significant p-values, such as Finance vs. Technology (p-value = 0.000052), indicating substantial differences in attrition rates across industries.

Profession

Analysis by profession shows that certain professions, such as Technology and Consulting, have higher survival probabilities compared to others like Management and Business. The log-rank test results, such as Management vs. Technology (p-value = 0.027865), confirm these differences, highlighting the influence of professional roles on attrition rates.

Mode of Transportation

The mode of transportation to work also impacts survival probabilities. Employees who commute by bus have lower survival probabilities compared to those who commute by car or on foot. The log-rank test shows significant differences, particularly between bus and foot commuters (p-value = 0.004621).

Cox Proportional Hazards Model

The Cox Proportional Hazards (CPH) model provides a multivariate analysis, identifying key predictors of employee attrition and quantifying their impact.

Full Model

The full CPH model includes all relevant covariates. The table below presents the significant predictors from the full model along with their hazard ratios and p-values:

Table 1: The Full Model of Cox Proportional Hazards Model upon Employee Atttrion

Predictor	*Coef*	*AHR*	*Coef 95% CI*	*AHR 95% CI*	*P value*
Age	0.02	1.02	0.01~0.03	1.01~1.03	<0.005***
Greywage	0.51	1.67	-0.25~0.77	1.29~2.15	<0.005***
Selfcontrol	-0.06	0.94	-0.13~0.01	0.88~1.01	0.08*
Anxiety	-0.06	0.94	-0.13~0.01	0.88~1.01	0.08*
Industry					
Finance (R.C)					
Industry	-0.51	0.60	-0.82~-0.20	0.44~0.82	<0.005***
Services	-0.59	0.55	-0.86~-0.32	0.42~0.73	<0.005***
Technology	-0.95	0.39	-1.57~-0.52	0.25~0.59	<0.005***
Way					
Bus (R.C)					
Car	-0.20	0.82	-0.39~0.00	0.67~1.00	0.05**
Foot	-0.35	0.71	-0.68~0.01	0.51~0.99	0.04**
Profession					
Business (R.C)					
Consulting	0.52	1.68	-0.04~1.07	0.96~2.92	0.07*

Engineering	0.56	1.75	-0.10~1.22	0.90~3.39	0.10*
Technology	-0.40	0.67	-0.88~0.07	0.41~1.07	0.10*
Traffic					
KA (R.C)					
Empjs	0.68	1.98	0.30~1.07	1.34~2.92	<0.005***
Youjs	0.39	1.48	0.00~0.78	1.00~2.17	0.05**

Table 1 illustrates the full Cox Proportional Hazards (CPH) model for employee attrition, detailing the impact of various predictors on the hazard of leaving the organization. Each predictor's coefficient (Coef), adjusted hazard ratio (AHR), confidence intervals (95% CI), and p-values are provided to understand their significance and influence on attrition.

Age is a significant predictor with a coefficient of 0.02 and an AHR of 1.02, indicating that each additional year of age increases the risk of attrition by 2% (p < 0.005).

Employees with grey wages have a coefficient of 0.51 and an AHR of 1.67, indicating a 67% higher risk of attrition compared to those with formal wages (p < 0.005).

Higher self-control and anxiety slightly reduce the risk of attrition by 6% each (AHR = 0.94, p = 0.08).

Industry type significantly affects attrition, with finance as the reference category. Employees in the industry sector have a 40% lower risk of attrition, those in services have a 45% lower risk, and those in technology have a 61% lower risk (p < 0.005).

Mode of transportation also impacts attrition. Employees who commute by car have an 18% lower risk, while those who walk have a 29% lower risk compared to bus commuters (p = 0.05 and 0.04, respectively).

Professionally, consultants have a 68% higher risk of attrition and engineers a 75% higher risk compared to those in business. Technology professionals, however, show a 33% lower risk of attrition (p = 0.07 and 0.10, respectively).

The source of traffic significantly affects attrition. Employees contacted through job sites (Empjs) have a 98% higher risk of attrition, while those who applied through job sites (Youjs) have a 48% higher risk (p < 0.005 and 0.05, respectively).

4.2.2. Reduced Model

The reduced CPH model includes covariates that were significant at the 90% significance level in the full model. The significance level for evaluating the reduced model remains at 95%. The table below highlights the reduced model predictors with their hazard ratios and p-values:

Table 2 The Reduced Model of Cox Proportional Hazards Model upon Employee Atttrition

Predictor	*Coef*	*AHR*	*Coef 95% CI*	*AHR 95% CI*	*P value*
Age	0.02	1.02	0.01~0.03	1.01~1.03	<0.005***
Greywage	0.55	1.62	0.23~0.73	1.26~2.08	<0.005***
Industry					
Finance (R.C)					
Industry	-0.34	0.71	-0.58~-0.09	0.56~0.91	0.01***
Services	-0.43	0.65	-0.63~-0.23	0.53~0.79	<0.005***
Technology	-0.84	0.43	-1.21~-0.47	0.30~0.63	<0.005***
Traffic					
KA (R.C)					
Empjs	0.45	1.56	0.23~0.66	1.26~1.94	<0.005***
Youjs	0.17	1.18	-0.04~0.37	0.96~1.45	0.11

Way					
Bus (R.C)					
Car	-0.23	0.79	-0.42~0.05	0.66~0.96	0.01***
Foot	-0.46	0.63	-0.78~0.14	0.46~0.87	0.01***

Table 2 shows the reduced Cox Proportional Hazards (CPH) model for employee attrition, highlighting significant predictors and their impact on the risk of leaving the organization. Each predictor's coefficient (Coef), adjusted hazard ratio (AHR), confidence intervals (95% CI), and p-values are provided.

Age is a significant predictor, with each additional year increasing the risk of attrition by 2% (AHR = 1.02, p < 0.005).

Employees with grey wages face a 72% higher risk of attrition compared to those with formal wages (AHR = 1.72, p < 0.005).

Higher self-control and anxiety reduce the risk of attrition by 6% each (AHR = 0.94, p < 0.005 and 0.01, respectively).

Industry type significantly affects attrition. Employees in the industry sector have a 28% lower risk, those in services have a 34% lower risk, and those in technology have a 55% lower risk compared to finance (p < 0.005).

Consultants have an 83% higher risk of attrition compared to those in business (AHR = 1.83, p = 0.02).

Employees contacted through job sites have a 66% higher risk of attrition, while those who applied through job sites have a 26% higher risk (AHR = 1.66 and 1.26, p < 0.005 and 0.03, respectively).

Mode of transportation is also significant. Commuting by car lowers the risk by 17%, and walking lowers it by 35% compared to taking the bus (p = 0.05 and 0.01, respectively).

Conclusion

This study used Cox Proportional Hazards (CPH) models to identify key factors influencing employee attrition. Significant predictors included age, grey wage, psychological factors, industry type, professional role, traffic source, and mode of transportation. The reduced model streamlined these to the most impactful predictors. These insights highlight the importance of targeted retention strategies addressing specific employee needs and industry challenges. By implementing these strategies, organizations can improve stability, employee satisfaction, and overall performance.

Acknowledgement

I would like to express my deepest gratitude to my thesis supervisor, Assoc. Prof. Dr. Zarina Mohd Khalid, for her invaluable guidance and support throughout this journey. Special thanks to Universiti Teknologi Malaysia (UTM) for providing financial support and necessary resources. I also acknowledge the camaraderie of fellow postgraduate students and the unwavering support of my family, whose love and encouragement have been my constant motivation. Lastly, I extend my thanks to everyone who contributed to this research, directly or indirectly. Your support has been greatly appreciated.

References

[1] Abedi, F. (2022). Survival analysis for user disengagement prediction: question-and-answering communities' case. Social Network Analysis and Mining, 12(1).

[2] Dave, V., Hasan, M., Zhang, B., & Reddy, C. (2018). Predicting interval time for reciprocal link creation using survival analysis. Social Network Analysis and Mining, 8(1).

[3] Sheng, J., Qian, X., & Ruan, T. (2018). Analysis of influencing factors on survival time of patients with heart failure. Open Journal of Statistics, 08(04).

[4] Chen, S. (2021). A novel estimator of survival without assumption of censored survival time.

[5] Chen, G. (2020). Deep kernel survival analysis and subject-specific survival time prediction intervals.

[6] Wang, P., Li, Y., & Reddy, C. (2019). Machine learning for survival analysis. ACM Computing Surveys, 51(6), 1-36.

[7] Nemati, M., Ansary, J., & Nemati, N. (2020). Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. Patterns, 1(5), 100074.

[8] Wu, Z. 2010. A Hidden Markov Model for Earthquake Declustering. *Journal of Geophysical Research*, 115

[9] Galili, B., Samohi, A., & Yakhini, Z. (2021). On the stability of log-rank test under labeling errors. Bioinformatics, 37(23), 4451-4459.

[10] Kerschke, L., Faldum, A., & Schmidt, R. (2020). An improved one-sample log-rank test. Statistical Methods in Medical Research, 29(10), 2814-2829.

[11] Mukhopadhyay, P., Ye, J., Anderson, K., Roychoudhury, S., Rubin, E., Halabi, S., … & Chappell, R. (2022). Log-rank test vs maxcombo and difference in restricted mean survival time tests for comparing survival under nonproportional hazards in immuno-oncology trials. Jama Oncology, 8(9), 1294.

[12] Yang, S. (2018). Interim monitoring using the adaptively weighted log-rank test in clinical trials for survival outcomes. Statistics in Medicine, 38(4), 601-612.

[13] Jiang, R. (2021). Two bias-corrected Kaplan-Meier estimators. Quality and Reliability Engineering International, 38(6), 2939-2952.

[14] Hess, A., and Hess, J. (2020). Kaplan–Meier survival curves. Transfusion, 60(4), 670-672.

[15] Gu, J., Fan, Y., & Yin, G. (2021). Reconstructing the Kaplan–Meier estimator as an M-estimator. The American Statistician, 76(1), 37-43.

[16] Yasin, R. (2021). Responsible leadership and employees' turnover intention: explore the mediating roles of ethical climate and corporate image. Journal of Knowledge Management, 25(7), 1760-1781.

[17] Wulansari, P., Meilita, B., & Ganesan, Y. (2020). The effect of employee retention company to turnover intention employee—case study on head office lampung bank.

[18] Arishi, M., Elsaid, A., Dawi, S., & Elsaid, E. (2018). Impact of socially responsible leadership on employee leave intention: exploratory study on it companies in egypt. Business and Management Research, 7(2), 17..

[19] Raza, A., Munir, K., Almutairi, M., Younas, F., & Fareed, M. (2022). Predicting employee attrition using machine learning approaches. Applied Sciences, 12(13), 6424.

[20] Goerdten, J., Carrière, I., & Muniz-Terrera, G. (2020). Comparison of Cox proportional hazards regression and generalized Cox regression models applied in dementia risk prediction. Alzheimer S & Dementia Translational Research & Clinical Interventions, 6(1).

[21] El-Rayes, N., Fang, M., Smith, M., & Taylor, S. (2020). Predicting employee attrition using tree-based models. International Journal of Organizational Analysis, 28(6), 1273-1291.

[22] Rennert, L., and Xie, S. (2021). Cox regression model under dependent truncation. Biometrics, 78(2), 460-473.

[23] Zucchini, W., MacDonald, I. L., Langrock, R. 2016. Hidden Markov Models for Time Series: An Introduction Using R, 2nd ed. CRC Press, Boca Raton.

[24] Bowden, T., Bricknell, I., & Preziosi, B. (2017). Comparative pathogenicity of Vibrio spp., Photobacterium damselae ssp. damselae and five isolates of Aeromonas salmonicida ssp. achromogenes in juvenile Atlantic halibut (Hippoglossus hippoglossus). Journal of Fish Diseases, 41(1), 79-86.

[25] Kwon, O., Hong, S., Ghang, B., Lee, C., & Yoo, B. (2017). The reply. The American Journal of Medicine, 130(10), e469.