Eduvest – Journal of Universal Studies
Volume 5 Number 5, May, 2025
p- ISSN 2775-3735- e-ISSN 2775-3727

IMPLEMENTATION OF ORANGE DATA MINING FOR EMPLOYEE TURNOVER PREDICTION OF COMPANY X

Thomas Wigung Aji Prayitna*, Imam Yuadi
Universitas Airlangga, Surabaya, Indonesia
Email: thomas.wigung.aji-2024@pasca.unair.ac.id, imam.yuadi@fisip.unair.ac.id

ABSTRACT

Employee turnover presents a significant challenge in Human Resources, particularly for companies operating across broad geographic areas such as Indonesia. High turnover rates can disrupt organizational continuity, increase recruitment costs, and affect overall performance. To mitigate these impacts, companies need to predict employee turnover likelihood accurately. This study uses the Orange Data Mining platform to compare the effectiveness of various machine learning models in predicting employee turnover. The models evaluated in this research include Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (K-NN), Neural Network, Decision Tree, and Logistic Regression. Model performance was assessed using cross-validation, Receiver Operating Characteristic (ROC) analysis, and confusion matrix metrics such as precision, recall, and false positives. The findings reveal that the Naive Bayes model outperforms the other models, demonstrating the highest precision rate and the lowest false positive rate. These results suggest that Naive Bayes offers a reliable and efficient approach to turnover prediction, enabling Human Resource departments to implement proactive retention strategies. This study implies that data-driven decision-making in HR analytics can substantially improve workforce planning and reduce the operational costs associated with high turnover.

KEYWORDS *Turnover, Orange Data Mining, Data Analytics, Prediction.*

Article Info:
Submitted: 17-04-2025 Final Revised: 13-05-2025 Accepted: 15-05-2025 Published: 22-05-2025

INTRODUCTION

Company X is a national company whose working area covers the entire territory of Indonesia and makes its employees must be ready to be placed or assigned throughout Indonesia from Sabang to Merauke (Obinna et al., 2022). In this condition, many employees can survive, but not a few who apply for transfer and even resign, thus increasing the turnover rate. Turnover is the tendency of

How to cite: Prayitna, T. W. A., & Yuadi, I. (2025). Implementation of Orange Data Mining for Employee Turnover Prediction of Company X. Journal Eduvest. *5*(5): 5203-5214.
E-ISSN: 2775-3727
Published by: https://greenpublisher.id/

employees to leave their current job and seek employment elsewhere (Mardiani et al., 2023). Turnover can be permanent when employees leave the institution where they work, or it can be characterized by horizontal mobility when employees seek or receive transfers to other departments (Maulidah, Supriyadi, et al., 2021).

Turnover can cause various problems in the company if not addressed immediately (Schreiber-Gregory & N, 2018). Losing highly skilled employees can have disruptive implications for the organization, such as disruption of organizational functions, service delivery, and administration (Safitri et al., 2023). This is particularly detrimental to organizations as they are challenged to find suitable replacements (Irawan et al., 2020). In the process of employee turnover, resources are needed in the form of costs for recruiting new employees and costs for education and training until the employee can have the appropriate competencies (Ike et al., 2023). In addition to requiring costs, the employee turnover process also requires time that is not instantaneous (Bothma et al., 2013). Given its detrimental impact on the company, turnover needs to be addressed by predicting the possibility of turnover of each employee with historical data (Rangiwai et al., 2022). The results of predicting employee turnover can then be the basis for human resource management to develop programs or provide treatment to related employees, so that the possibility of turnover can be reduced (Mukhopadhyay et al., 2020).

Predictive analysis of employee turnover can be done with Orange Data Mining tools (Setyohadi et al., 2018). Orange is a comprehensive component-based framework for machine learning and data mining (Sari, Safitri, et al., 2023). Orange has been used in science, industry, and learning (Ghazi et al., 2021). So far, predictive models with Orange Data Mining tools are widely used in the fields of health, education and economics (Matondang et al., 2018). While predictive models in the Human Resources field, especially turnover prediction, are still minimal (Santoso & C, 2023).

Previous research conducted by Maulidah, N, et al. (2021) used Support Vector Machine (SVM) and Naive Bayes to predict diabetes mellitus; the results showed that the SVM model could provide better accuracy. Another study was conducted by Dinda Safitri et al in 2023 which used the Naive Bayes, K-NN and Neural Network models to predict student graduation, and the results showed that K-NN was able (Muharrom, 2023) to provide the best accuracy. The last research that became a reference was conducted by Tri Kartika Sari and Imam Yuadi in 2023, which used four models, namely K-NN, Decision Tree, Naive Bayes, and Logistic Regression, to predict the nutritional status of toddlers and showed that the K-NN model had the best performance (Wahyuni, 2022). Based on previous research, this study will compare the performance of six methods, namely SVM, Naive Bayes, K-NN, Neural Network, Decision Tree, and Logistic Regression, and make predictions on testing data using the best method. Cross-validation, ROC analysis, and confusion matrix methods were used to test the performance of each prediction model (Sari, D., et al., 2023).

The results of this research are expected to make a practical contribution to Company Management, especially HR Managers, by predicting employee turnover

and then developing the most appropriate retention program to retain the best employees and reduce turnover rates.

RESEARCH METHOD

The research process and methods are shown in Figure 1 below:

Figure 1. Research Methods

Identification

This is the first stage carried out and is used to explore existing problems as a basis for setting the objectives of data analysis and subsequent processes. At this stage, it was found that high employee turnover rates were a problem, so a model was needed to predict employee turnover in the future as a management consideration in developing employee retention programs.

Data Collection

Data collection was conducted in one of the Company X units whose working area covers West Nusa Tenggara (NTB) Province (Yang, Thu Hue, et al., 2024). This unit was chosen because it manages all of Company X's business processes from upstream to downstream, so it represents Company X as a whole. The Company X unit in NTB has 867 employees so that the data will be divided into training and testing data. Training data is taken from as many as 80% or 694 employees. In this training data, there is a turnover column, a prediction target with records 0 (no turn) and 1 (turn). The testing data is 20%, or 173 employees. In this testing data, there is no turnover column.

To perform data analysis, data is needed in the form of basic employee data and other attribute data commonly used as a reason for turnover. Basic employee data in the form of (1) NIP; (2) Name; (3) Position; (4) Unit; (5) Grade; (6) Position Level; (7) Status / History of Structural Position; (8) Education Level; (9) Period of Service; (10) Last Position Period; (11) Gender; (12) Marital Status and (13) Number of family members who are still covered. Basic employee data is secondary data that can be obtained from the HR department, while other data needed are attributes that are often used as reasons for turnover, namely: (1) Whether living in homebase; (2) Whether living with nuclear family; (3) Spouse's employment status and (4) Health condition. The following additional data is primary data collected through a questionnaire distributed to all employees online through Google Forms.

Preprocessing Data

The collected data is merged into one dataset, and data cleansing and visualization are carried out. This process ensures that no data is missing or empty and that the data is consistent.

The results of data preprocessing showed no empty data and a consistent data format, so the dataset can be processed further in the data mining process and model testing.

The data preprocessing stage is also used to separate the dataset into training data and testing data.

Data Mining and Data Testing

Data mining and model testing were carried out with Orange Data Mining using six models based on previous research as references, namely as follows: (1) SVM, (2) Naive Bayes, (3) K-NN, (4) Neural Network, (5) Decision Tree and (6) Logistic Regression.

This process begins by importing training data files in the Orange Data Mining application through the File widget, then selecting attributes through the Select Columns widget. Select Columns determines attributes that will become Features (independent variables) and Targets (dependent variables).

The feature attributes are Position, Grade, Position Level, Structural, Education Level, Period of Service, Last Position Period, Gender, Marital Status, Number of Dependents, Living in Homebase, Living with Immediate Family, Spouse Employment Status, and Health Condition. The attribute that becomes the Target is Turnover. The NIP and Name attributes become meta or additional information that has no effect on the decision or target.

The column selection process is shown in Figure 2 below:

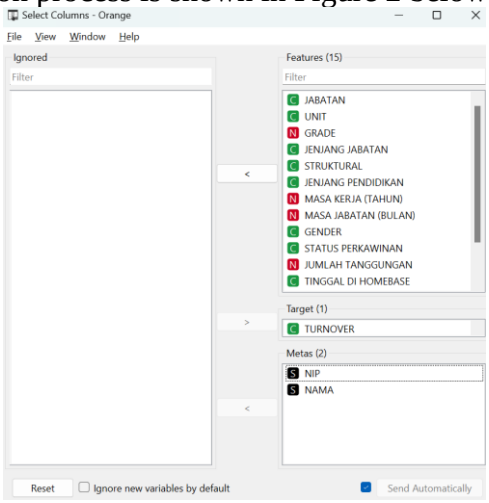

Figure 2. Widget Select Columns

The following process creates a widget design to correlate the dataset with the model and prediction functions. The widget design is shown in Figure 3.

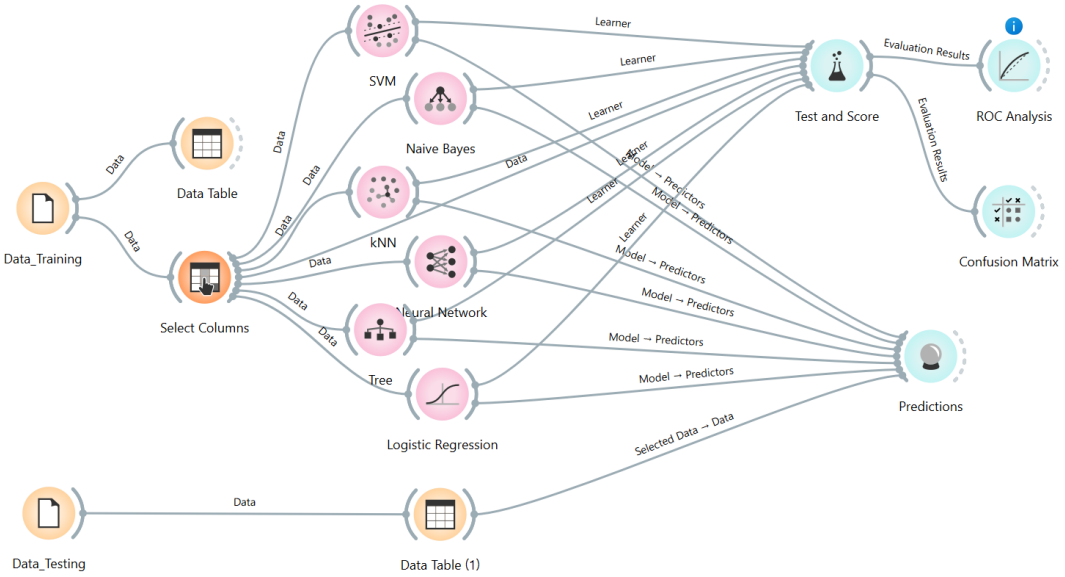

Figure 3. Widget Design

Evaluation

The performance evaluation of each data analysis model is done by comparing models through the Test and Score widget in the Orange Data Mining application.

	SVM	NAÏVE BAYES	kNN	NEURAL NETWORK	TREE	LOGISTIC REGRESSION
SVM		0.668	0.999	0.073	0.986	0.112
NAÏVE BAYES	0.332		0.999	0.111	0.992	0.166
kNN	0.001	0.001		0.001	0.309	0.001
NEURAL NETWORK	0.927	0.889	0.999		0.989	0.540
TREE	0.014	0.008	0.691	0.011		0.013
LOGISTIC REGRESSION	0.888	0.834	0.999	0.46	0.987	

Figure 4. Performance comparison between models

RESULT AND DISCUSSION

To get the best model for predicting employee turnover based on the historical data of the NTB Unit of Company X, each model's performance is measured using Cross Validation, ROC Analysis, and a Confusion Matrix. The performance evaluation between these models results in the best model, which is then recommended to the company management for use as a turnover prediction model.

Cross Validation

This cross-validation testing is done through the Test and Score Widget with the following results:

	AUC	CA	F1	Precision	Recall	MCC
SVM	0.937	0.947	0.934	0.947	0.947	0.533
NAÏVE BAYES	0.931	0.896	0.909	0.931	0.896	0.485
kNN	0.687	0.917	0.896	0.887	0.917	0.165
NEURAL NETWORK	0.955	0.951	0.946	0.946	0.951	0.603
TREE	0.733	0.942	0.94	0.938	0.942	0.556
LOGISTIC REGRESSION	0.954	0.963	0.959	0.960	0.963	0.705

Figure 5. Cross-validation test results

Based on the cross-validation test results, the Neural Network model has the best performance in terms of AUC and CA, which is 0.955 for AUC and 0.951 for CA. Meanwhile, the Logistic Regression model has the best performance in terms of F1 (0.959), Precision (0.960), Recall (0.963), and MCC (0.705).

ROC Analysis

ROC Analysis testing compares the accuracy between models in predicting binary classification. This test is done with the ROC Analysis widget in the Orange Data Mining application with the following results:

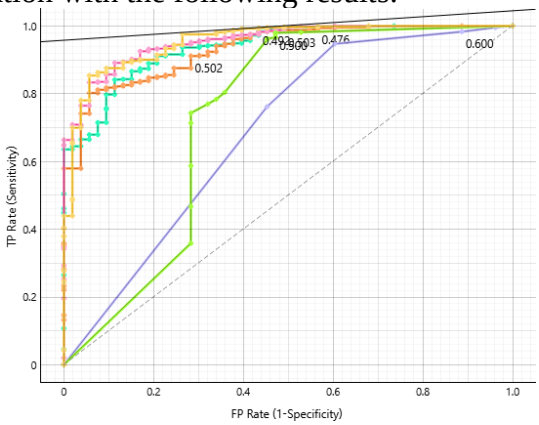

Figure 6. ROC Analysis 0 (no turn)

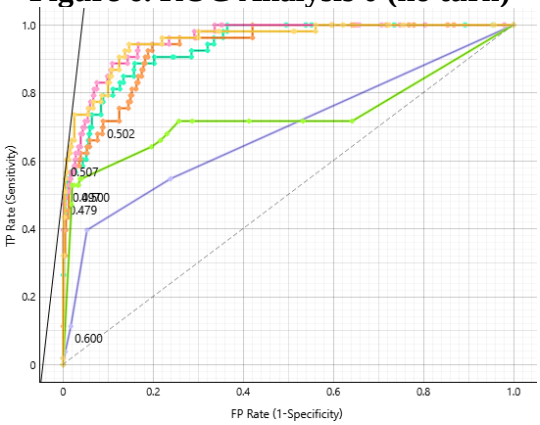

Figure 7. ROC Analysis 1 (turn)

In ROC Analysis, the curve close to 1 is the model that provides the best accuracy. The following ROC Analysis results show that there are four curves close to 1: the blue curve (SVM), pink curve (Neural Network), orange curve (Naive Bayes), and yellow curve (Logistic Regression). So, it can be concluded that the

SVM, Naive Bayes, Neural Network, and Logistic Regression models can provide the best accuracy.

Confusion Matrix

This test was conducted with the Confusion Matrix widget in the Orange Data Mining application. The following is a comparison of the Confusion Matrix test results against the six prediction models above:

	Accuracy	Precission	Recall	False Positive
SVM	92,20%	95,76%	95,79%	27,2
NAIVE BAYES	87,84%	97,44%	89,17%	15,0
kNN	88,25%	93,56%	93,73%	41,4
NEURAL NETWORK	94,50%	96,30%	97,80%	24,1
DECISSION TREE	93,30%	96,07%	96,58%	25,3
LOGISTIC REGRESSION	92,93%	96,08%	96,27%	25,2

Figure 8. Confusion Matrix test results

The Confusion Matrix test results for each prediction model are as follows:

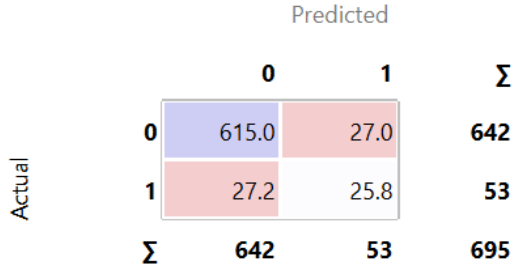

Figure 9. SVM testing results for the Confusion Matrix

$$Accuracy = \frac{615{,}0 + 25{,}8}{615{,}0 + 27{,}2 + 27{,}0 + 25{,}8} \times 100\% = 92{,}20\%$$

$$Precission = \frac{615{,}0}{615{,}0 + 27{,}2} \times 100\% = 95{,}76\%$$

$$Recall = \frac{615{,}08}{615{,}0 + 27{,}0} \times 100\% = 95{,}79\%$$

Naive Bayes

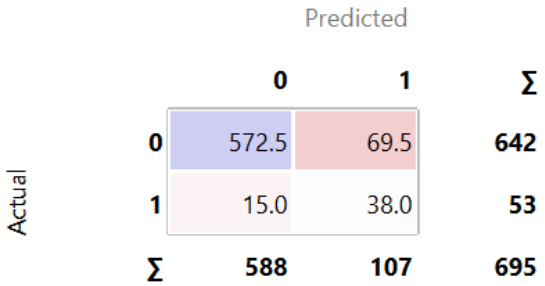

Figure 10. Naive Bayes testing results for the Confusion Matrix

$$Accuracy = \frac{572{,}5 + 38{,}0}{572{,}5 + 15{,}0 + 69{,}5 + 38{,}0} \times 100\% = 87{,}84\%$$

$$Precission = \frac{572{,}5}{572{,}5 + 15{,}0} \times 100\% = 97{,}44\%$$

$$Recall = \frac{572{,}5}{572{,}5 + 69{,}5} \times 100\% = 89{,}17\%$$

kNN

Actual \ Predicted	0	1	Σ
0	601.8	40.2	642
1	41.4	11.6	53
Σ	643	52	695

Figure 11. kNN testing results for the Confusion Matrix

$$Accuracy = \frac{601{,}8 + 11{,}6}{601{,}8 + 41{,}4 + 40{,}2 + 11{,}6} \times 100\% = 88{,}25\%$$

$$Precission = \frac{601{,}8}{601{,}8 + 41{,}4} \times 100\% = 93{,}56\%$$

$$Recall = \frac{601{,}8}{601{,}8 + 40{,}2} \times 100\% = 93{,}73\%$$

Neural Network

Actual \ Predicted	0	1	Σ
0	627.9	14.1	642
1	24.1	28.9	53
Σ	652	43	695

Figure 12. Neural Network Testing Results for the Confusion Matrix

$$Accuracy = \frac{627{,}9 + 28{,}9}{627{,}9 + 24{,}1 + 14{,}1 + 28{,}9} \times 100\% = 94{,}50\%$$

$$Precission = \frac{627{,}9}{627{,}9 + 24{,}1} \times 100\% = 96{,}30\%$$

$$Recall = \frac{627{,}9}{627{,}9 + 14{,}1} \times 100\% = 97{,}80\%$$

Decision Tree

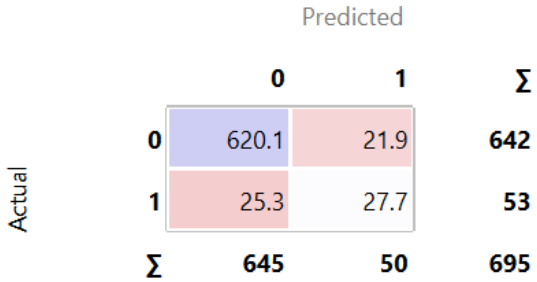

Figure 13. Decision Tree testing results for the Confusion Matrix

$$Accuracy = \frac{620{,}1 + 27{,}7}{620{,}1 + 25{,}3 + 21{,}9 + 27{,}7} \times 100\% = 93{,}20\%$$

$$Precission = \frac{620{,}1}{620{,}1 + 25{,}3} \times 100\% = 96{,}07\%$$

$$Recall = \frac{620{,}1}{620{,}1 + 21{,}9} \times 100\% = 96{,}58\%$$

Logistic Regression

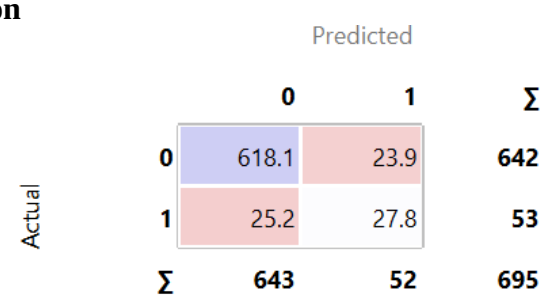

Figure 14. Logistic Regression testing results for the Confusion Matrix

$$Accuracy = \frac{618{,}1 + 27{,}8}{618{,}1 + 25{,}2 + 23{,}9 + 27{,}8} \times 100\% = 92{,}93\%$$

$$Precission = \frac{618{,}1}{618{,}1 + 25{,}2} \times 100\% = 96{,}08\%$$

$$Recall = \frac{618{,}1}{618{,}1 + 23{,}9} \times 100\% = 96{,}27\%$$

Based on the comprehensive evaluation of six machine learning models—SVM, Naive Bayes, K-Nearest Neighbors (K-NN), Neural Network, Decision Tree, and Logistic Regression—using three evaluation techniques (Cross Validation, ROC Analysis, and Confusion Matrix), this study identified the most optimal model for predicting employee turnover in the NTB Unit of Company X.

The Cross Validation test indicated that the Neural Network model achieved the highest AUC (0.955) and Classification Accuracy (CA) (0.951), while Logistic Regression performed best in terms of F1-score (0.959), Precision (0.960), Recall (0.963), and MCC (0.705). Meanwhile, the ROC Analysis revealed that the SVM, Naive Bayes, Neural Network, and Logistic Regression models demonstrated ROC

curves approaching 1, indicating superior predictive accuracy for binary classification problems (Rahadi, 2021; Rozarie & Indonesia, 2017; Siswati et al., 2024; Tjendra, 2019; Yang, Y, et al., 2024).

Although each model showed varying strengths in the Confusion Matrix, Naive Bayes demonstrated balanced performance with minimal false positives. This makes it particularly valuable in HR contexts where overestimating turnover risks can lead to unnecessary interventions. The consistency of its precision and overall accuracy across metrics makes it the most practical choice for deployment.

CONCLUSION

Based on model testing using cross validation, ROC analysis and Confusion Matrix, it can be concluded that the Naive Bayes model is the best model for predicting employee turnover in the NTB Unit of Company X. This conclusion has several reasons, firstly based on the Cross Validation Test, the Naive Bayes model only has a gap of 0.029 with the Logistic Regression model. The second reason is based on ROC analysis; the Naive Bayes model is one of the models whose curve is close to 1, so it can provide good accuracy. Finally, based on confusion matrix testing, Naive Bayes shows the highest precision rate of 97.44% and the lowest False Positive rate of 15. In the case of employee turnover prediction, the best model is the one that can provide the best precision and minimize false positives.

REFERENCES

Bothma, C, C. F., Roodt, & G. (2013). The validation of the turnover intention scale. *SA Journal of Human Resource Management*, *11*(1). https://doi.org/https://doi.org/10.4102/sajhrm.v11i1.507

Ghazi, H, A., Elsayed, I, S., Khedr, & E, A. (2021). A proposed model for predicting employee turnover of information technology specialists using data mining techniques. *International Journal of Electrical and Computer Engineering Systems*, *12*(2). https://doi.org/https://doi.org/10.32985/IJECES.12.2.6

Ike, O, O., Ugwu, E, L., Enwereuzor, K, I., Eze, C, I., Omeje, O, Okonkwo, & E. (2023). Expanded-multidimensional turnover intentions: scale development and validation. *BMC Psychology*, *11*(1). https://doi.org/https://doi.org/10.1186/s40359-023-01303-2

Irawan, L., Hasibuan, L. H., & Fauzi, F. (2020). Analisa Prediksi Efek Kerusakan Gempa Dari Magnitudo (Skala Richter) Dengan Metode Algoritma Id3 Menggunakan Aplikasi Data Mining Orange. *Jurnal Teknologi Informasi: Jurnal Keilmuan Dan Aplikasi Bidang Teknik Informatika*, *14*(2). https://doi.org/10.47111/jti.v14i2.1079

Mardiani, E., Rahmansyah, N., Kurniati, I., Matondang, N., Tesalonika, T., Zanitha, D. A., & Romzy, I. (2023). Membandingkan Algoritma Data Mining Dengan Tools Orange untuk Social Economy. *Digital Transformation Technology*, *3*(2). https://doi.org/10.47709/digitech.v3i2.3256

Matondang, N, Isnainiyah, N, I., Muliawatic, & A. (2018). Analisis Manajemen Risiko Keamanan Data Sistem Informasi (Studi Kasus: RSUD XYZ). *Ikatan Ahli*

Indormatika Indonesia, *2*(1), 282–287. https://doi.org/https://doi.org/10.29207/resti.v2i1.96

Maulidah, N., Supriyadi, R., Utami, D. Y., Hasan, F. N., Fauzi, A., & Christian, A. (2021). Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes. *Indonesian Journal on Software Engineering (IJSE)*, *7*(1). https://doi.org/10.31294/ijse.v7i1.10279

Maulidah, N, Supriyadi, R, Utami, Y, D., Hasan, N, F., Fauzi, A, Christian, & A. (2021). Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes. *Indonesian Journal on Software Engineering (IJSE)*, *7*(1). https://doi.org/https://doi.org/10.31294/ijse.v7i1.10279

Muharrom, M. (2023). Analisis Penggunaan Orange Data Mining untuk Prediksi Harga USDT/BIDR Binance. *Bulletin of Information Technology (BIT)*, *4*(2). https://doi.org/10.47065/bit.v4i2.654

Mukhopadhyay, A., Singh, P., & Thenmalar, S. (2020). Analysis of employee performance and prediction of potential attrition- A survey. *International Journal of Advanced Science and Technology*, *29*(6 Special Issue).

Obinna, I, O., Ugwu, E, L., Omeje, O, Okonkwo, E, Enwereuzor, & K, I. (2022). Expanded-Multidimensional turnover intentions: Scale development and validation Chukwudi Eze Ifeanyichukwu Police Academy Wudil Kano. *Research Square*.

Rahadi, D. R. (2021). Strategi organisasi penanganan turnover melalui pemberdayaan karyawan. *Solusi*, *19*(1), 102–116.

Rangiwai, B., Aliioaiga, E., Cook, M., Latu, F., & Tukutau, M. (2022). The effects of the 2021 Delta lockdown in Aotearoa New Zealand: Some preliminary material to inform a future research question. *Te Kaharoa*, *15*(1). https://doi.org/10.24135/tekaharoa.v15i1.394

Rozarie, C. V. R. A. De, & Indonesia, J. T. K. R. (2017). *Manajemen Sumber Daya Manusia*.

Safitri, D., Hilabi, S. S., & Nurapriani, F. (2023). Analisis Penggunaan Algoritma Klasifikasi Dalam Prediksi Kelulusan Menggunakan Orange Data Mining. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *8*(1). https://doi.org/10.36341/rabit.v8i1.3009

Santoso, & C. (2023). IMPLEMENTASI ORANGE DATA MINING UNTUK PREDIKSI BIAYA ASURANSI. *Aisyah Journal Of Informatics and Electrical Engineering (A.J.I.E.E)*, *5*(1). https://doi.org/https://doi.org/10.30604/jti.v5i1.180

Sari, D, W., Safitri, E, R., Hidayat, A, Bahar, O, R., Lavenia, & C. (2023). Overview Potential Predictor of Turnover Intention in Mining Company. *MANAZHIM*, *5*(1). https://doi.org/https://doi.org/10.36088/manazhim.v5i1.2884

Sari, W. D., Safitri, R. E., Hidayat, A., Bahar, R. O., & Lavenia, C. (2023). Overview Potential Predictor of Turnover Intention in Mining Company. *MANAZHIM*, *5*(1). https://doi.org/10.36088/manazhim.v5i1.2884

Schreiber-Gregory, & N, D. (2018). Ridge Regression and multicollinearity: An in-depth review. *Model Assisted Statistics and Applications*, *13*(4), 359–365.

Setyohadi, B, D., Purnawati, & W, N. (2018). An investigation of external factors for technological acceptance model of nurses in Indonesia. *IOP Publishing*, *403*, 012064. https://doi.org/https://doi.org/10.1088/1757-899x/403/1/012064

Siswati, S, Ernawati, T, Khairunnisa, & M. (2024). Analisis Tantangan Kesiapan Implementasi Rekam Medis Elektronik di Puskesmas Kota Padang. *Jurnal Kesehatan Vokasional*, *9*(1), 1–15.

Tjendra, I. W. (2019). Pengaruh motivasi kerja dan komitmen organisasional terhadap turnover intention pada karyawan UFO Elektronika Surabaya. *Agora*, *7*(1), 287099.

Wahyuni, M. I. (2022). Pengaruh Supportive Educative Berbasis Caring Terhadap Self Management Penderita Asma Di Puskesmas Kalianget. *Skripsi*. *Fakultas Ilmu Kesehatan*. *Universitas Wiraraja*, *14*(1).

Yang, Y, Hue, T., M, H., Takeda, & S. (2024). Turnover intention among Vietnamese millennials in the workplace. *Evidence-Based HRM*, *12*(3). https://doi.org/https://doi.org/10.1108/EBHRM-12-2022-0302

Yang, Y., Thu Hue, H. M., & Takeda, S. (2024). Turnover intention among Vietnamese millennials in the workplace. *Evidence-Based HRM*, *12*(3). https://doi.org/10.1108/EBHRM-12-2022-0302