Techniques of Machine Learning Applied to Reduce Employee Turnover in a Company Cleaning and Disinfection

Erika Noemi Romero Rojas
Facultad de Ingeniería, Carrera de Ingeniería Industrial
Universidad de Lima
Lima, Perú
20162488@aloe.ulima.edu.pe

Yvan Jesús García López
Facultad de Ingeniería, Carrera de Ingeniería Industrial
Universidad de Lima
Lima, Perú
ygarcia@ulima.edu.pe

Abstract

Staff turnover in large Peruvian manufacturing industries has been increasing in recent years. While job rotation is a natural effect in organizations, it generates higher training costs for new staff and impacts work performance and climate when unwanted. Given this problem arises the need to identify the possible causes of rotation of operational personnel and predict these events through data analysis at an early stage to avoid and reduce its impact on the company. This article of quantitative approach and exploratory scope-explanatory aims to identify the propensity of rotation of the operation of a company manufacturing cleaning and disinfection through a model of forecast by collecting data using Machine Learning and encourage proposals that enable solutions to be found to the factors influencing staff turnover. MS Excel and Orange software were used for data analysis, where the data were trained with different intelligence models such as Random Forest, Logistic Regression, Decision Tree, and SVM, and Python to run the model and get numerical indicators like the Area under the curve (AUC) and the analysis of the ROC curve. The proposed study shows that the models perform well in classification, with high accuracy and recall rates, 96% and 97%, respectively, and an overall accuracy of 96%.

Keywords
Machine Learning – Staff turnover – Manufacturing – Human Resources – Operational staff.

Introduction
Omondiagbe et al. (2019). Machine Learning is a type of artificial intelligence that focuses on developing software that can change according to exposure to new data; its development uses past data that helps the classification process, prediction, and detection. According to Ifft and Kuhnsb, not all Machine Learning models are ideal for prediction. However, most have greater predictive power than standard models, improving performance when analysis variables are well selected (Armendariz and Zuñiga 2018).

King (2016) Data analysis in Human Resources Development is increasingly common since it favors comprehensive data organization and helps decision-making.

Nocker and Sena (2019). Predictive analysis is the most common type of analysis used by HR departments. It has been implemented in several contexts, such as modeling staff turnover and employee engagement. The main objective of applying Predictive Analysis in Human Resources is to optimize performance and produce a better return on investment for organizations through decision-making based on data collection, HR metrics, and predictive models. (Sujeet et al. 2016).

However, despite the benefits of applying predictive analysis in organizations, directors still have skepticism about using this tool in Human Resources. Ekawati (2019) according to an investigation conducted by IBM when interviewing 700 human resources managers showed that less than 25% use sophisticated analysis to predict future results and decision-making (p. 387). Editorial Ecoprensa (2019) presented a study indicating that Peru is one of the countries with an average turnover rate of 20.7%, while Latin America has 10.9%, reducing the outlook to a voluntary turnover rate. Latin America has 5.4% while Peru reaches 9.8% based on studies by Saratoga de PricewaterhouseCoopers (para. 4). Certain findings show that, in most cases, the undesired job rotation is related to fortuitous events, that is to say, not controllable by the company, whereas the departures due to job dissatisfaction are due to the unpaid in conformity, little recognition and absence of career lines. While all organizations are constantly changing and developing, leading to job rotation, when this is not foreseen, or the causes that motivate the event are unknown, it can cause a loss of control affecting their activities and impacting economically when unwanted.

Composition (2019). "The high turnover of employees entails a high cost to the company, not only because of the loss of that talent, which already knew the operation and provided value but also involves a new disbursement of money to attract, hire, train and retain a new partner" (para. 3). According to Gallup, replacing an employee costs 150% of their annual salary (as cited in Computrabajo 2019, p. 5).

Many studies have been done on how to make machines learn by themselves without needing to be programmed, and many mathematicians and programmers have applied various approaches to finding a solution when you have lots of data (Mahesh B., 2020). According to Adadi and Berrada, the concept of "machine learning," a branch of Artificial Intelligence (AI), is applied in the technology industry. It allows such learning without the need to program them previously and now provides systematization to identify and predict data. (As quoted in Garcia & Panduro, 2022, p.1).

Therefore, it is necessary to implement techniques to control and reduce such events; Holwerda (2021), organizations that maintain a firm foundation in management theory and welcome new analytical methods can win the race to generate business value from big data when applied to HR. In this sense, this article seeks, through the use of Machine Learning techniques, to identify What factors influence the rotation of operating personnel of a cleaning and disinfection manufacturing company. So that this problem, which has persisted for many years in the organization, can be countered through talent retention.

Methodology

Russo et al. (2016) points out: "Machine Learning is an area of artificial intelligence that encompasses a set of techniques that make machine learning possible through training with large volumes of data" (p. 131). For the present study, it is sought that the algorithm reviews the data and can predict future behaviors automatically.

Supervised Machine Learning Prediction Models

Machine Learning is the field of study of statistical models that gives computer systems the ability to perform specific tasks without having been explicitly programmed (Mahesh B. 2020). Its operation is based on learning "what to do with the data" and performing its work automatically. Its use applies to everyday life, from simple tasks such as searching for online information to forecasting complex activities. Each Machine Learning model can solve a specific problem effectively depending on the type of problem, number of variables, and data types. Supervised learning is a machine learning technique that assigns an input and an output function, in the form of input-output pairs. Models using this type of technique need external assistance. Figure 1 shows the Supervised Learning Flow.

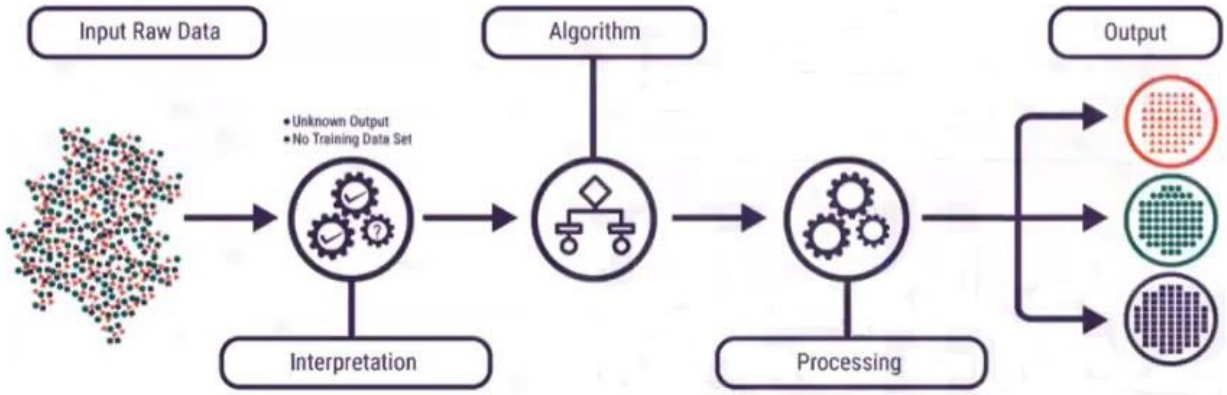

Figure 1. The Supervised Learning Flow

Nota. De "Machine Learning Algorithms – A Review," by B. Mahesh, 2020, *International Journal of Science and Research*, 9, p. 381-386 (https://doi.org/10.21275/ART20203995)

Among the algorithms used in the development of the article it is detailed:

a. Random Forest

Aracena (2022) points out: "It refers to an algorithm that generates a model of several decision trees, and each has a vote in the final prediction. A subset of randomly selected variables is used in tree generation to obtain variability between them" (p. 3). The model starts with many boot samples, which are extracted with the replacement of the original training data set; a regression tree is adjusted to each of the boot samples, distributing for each node of the tree a small group of input data and the subsequent selection of the variable is given randomly for the realization of the binary partition which is based on the choice of the input variable (Siddharth and Hao 2020). A Random Forest classification diagram is shown in Figure 2.

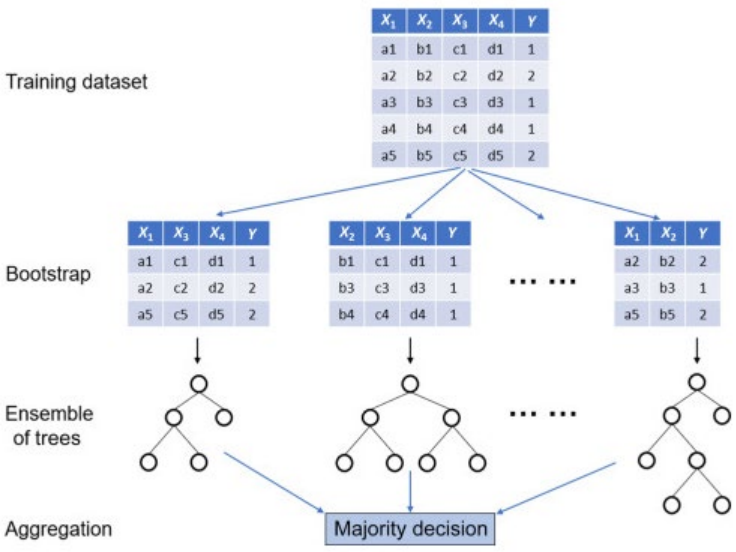

Figure 2. Random Forest Model

Nota. De "Chapter 9 – Noninvasive fracture characterization based on the classification of sonic wave travel times", by Siddharth, M. y Hao L., 2020, *El Sevier*, 1, p. 243-287 (https://doi.org/10.1016/B978-0-12-817736-5.00009-0)

a. Logistic Regression

Amazon (s.f.) notes: "It is a data analysis technique that uses mathematics to find the relationships between two data factors. Then, use this ratio to predict the value of one of those factors based on the other". Logistic regression is a standard statistical method for modeling the probability of a positive result (i.e., Yi = 1) as a function of the co-variables. (Blanchard, Clark, et al., 2023). The logistic Regression diagram is shown in Figure 3.

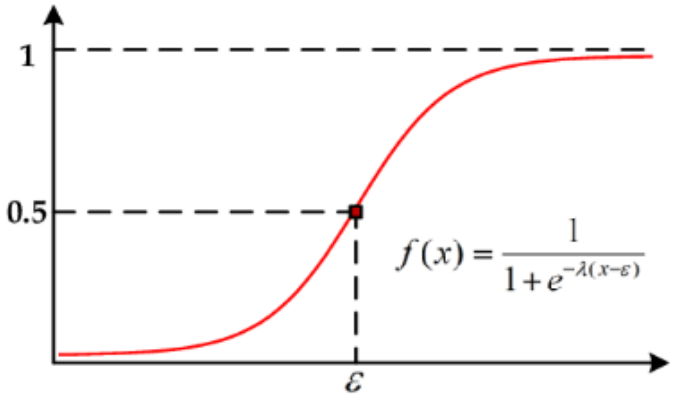

Figure 3. Logistic Regression Model

Nota. **De "Machine Learning Algorithms – A Review," by B. Mahesh, 2020,** ***International Journal of Science and Research*****, 9, p. 381-386 (** **https://doi.org/10.21275/ART20203995)**

b. **Decision Tree**

Mahesh (2020) is a graph to represent options and their results as a tree. The nodes in the chart represent an event or choice, and the chart's edges represent the decision rules or conditions. Each tree consists of nodes and branches. Each node represents attributes in a classified group, and each branch represents a value the node can take. Figure 4 shows an example diagram of decision-making under the Decision Tree.

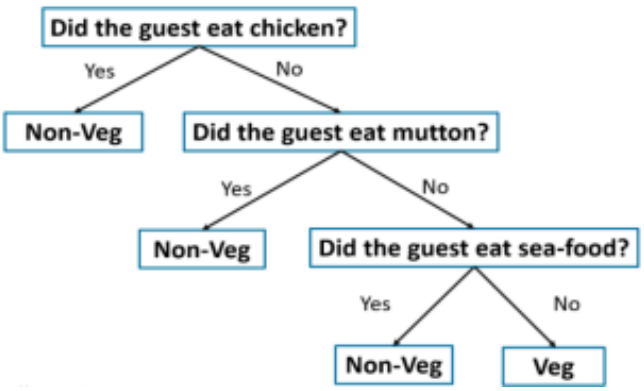

Figure 4. Decision Tree Model

Nota. De "Machine Learning Algorithms – A Review," by B. Mahesh, 2020, *International Journal of Science and Research*, 9, p. 381-386 (https://doi.org/10.21275/ART20203995)

SVM

Aracena (2022) points out: "It is an algorithm that generates a separator hyperplane between the data with their respective labels. The objective is to find a hyperplane with maximum separation and to make the least number of errors" (p. 3). According to Zhan, this algorithm analyzes binary variables in a two-dimensional plane, seeking the

maximum separation between observations of a small or medium data set but having a high level of complexity (as cited in Panduro, 2022, p. 2). Figure 5 shows the range between classes of an SVM model.

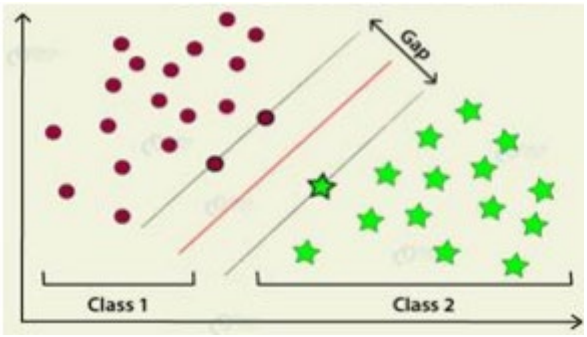

Figure 5. SVM Model

Nota. De "Machine Learning Algorithms – A Review," by B. Mahesh, 2020, *International Journal of Science and Research*, 9, p. 381-386 (https://doi.org/10.21275/ART20203995)

Collection of data

For the study, we used the data collection of the payroll of operational personnel that ceased between 2018 and 2021 (147 workers dismissed); these data were provided by a cleaning and disinfection company based in Lima, Peru. The "exit surveys" applied informally, that is to say, orally (since they did not have an exit survey) to the workers who stopped. Allow us to identify the causes that would have motivated the resignation of their operating personnel.

2.3 Tabulation and Analysis of Data

The information obtained through the payroll and exit surveys of operating personnel has been ordered and processed in MS Excel based on the following steps:

Step 1: Group the data based on the reasons for termination and assign them a code for their accounting.

Table 1. Reason for termination of employment

Reason for termination	**Code**	**Quantity**
Completion of contract	0	41
voluntary resignation	1	97
Dismissal	2	8
Mutual Dissent	3	1
Total		147

Step 2: Break down the factors that motivated the termination of the operating staff based on voluntary and involuntary events.

Table 2. Factors that motivated the termination of employment

Factors	Code	%
Other job offer	OTOFE	54%
Non-renewal (Low production)	NOBPR	10%
Non-renewal (Low performance)	NOBDE	9%
Non-renewal (Did not pass trial period)	NOSPP	5%
Non-renewal (accumulation of absences)	NORAF	5%
Does not renew to continue studies at the university	ESTUD	5%
Dismissal due to abandonment of work	DEABA	3%
Resignation due to workload	RECAR	2%
Does not continue in employment due to unforeseen trips	VIINE	2%
Non-renewal of contract (The operator resigned when he discovered that his contract would not be renewed.)	NOREN1	2%
Non-renewal (The person entered to fill a vacancy temporarily)	NOCVA	1%
Mutual dissent	MUDES	1%
Health problems	PROSA	1%
Total		100.00%

Step 3: Train the data in the Orange Software and then process it in Python.

2.4 Data Processing with Machine Learning

This research is classified as Practical-Experimental with a quantitative approach, Exploratory-Explanatory scope, and pure experimental design.

- Feature Reduction

In this section, the columns that will not be used by 'drop' from the pandas library were eliminated: *'DNI,' 'NOMBRES Y APELLIDOS,' 'CECO,' 'MOTI,' 'CARGO,' 'FECHA DE INICIO,' 'FECHA FIN.'*

- Data Splitting

The training and test data were created with the following variables: “X_train,” “y_train,” “X_Test,” and “y_test” using train_test_split from the scikit-learn library.

- Step 1: The independent variables were separated into the variable “x”.
- Step 2: The objective or dependent variable was separated into the “and” variable.
- Step 3: Created the variables X_train, y_train, X_test and y_test using train_test_split from scikit learn.

2.5 Fine tune Fine-Hyperparameters

For the present study, four models were used: Random Forest, Logistic Regression, Decision Tree, and SVM, to which hyperparameter fine-tuning was performed to find the best hyperparameters using GridSeachCV from the sci-kit-learn library. Table 1 shows the details of each value for the possible hyperparameters used in each model.

Table 2. Date to fine-tune hyperparameters

Modelos	Detalle
Random Forest	For this model, the following values were used as possible hyperparameters: max_depth: [10, 50, 100, 150] min_samples_split: [2, 4, 8, 10] n_estimators: [10, 12, 100, 150]
Logistic Regression	For this model, the following values were used as possible hyperparameters: penalty: [‘l1’, ‘l2’, ‘elasticnet’] solver: ['lbfgs', 'liblinear', 'newton-cg,' 'sag,' 'saga']
Decision Tree	For this model, the following values were used as possible hyperparameters: criterion: [‘gini,’ ‘entropy,’ ‘log_loss’] max_depth: [10, 50, 100, 150] min_samples_split: [2, 4, 8, 10]
SCV	For this model, the following values were used as possible hyperparameters: C: [0.1, 1] gamma: [‘scale,’ ‘auto’] kernel: ['rbf,' 'poly,' 'sigmoid,' ‘linear"]

2.6 Machine Learning

In this section, the prediction was made using each machine-learning model. Thanks to the information from the FINE TUNE HYPERPARAMETER, it was possible to obtain the best hyperparameters for each model. Table 2 shows us the values assigned to each hyperparameter for each model.

Table 3. Algorithms Detail

Models	Details
Random Forest	For this model, the following values were used as possible hyperparameters: max_depth: 10 n_estimators: 10
Logistic Regression	For this model, the default hyperparameters were used.
Decision Tree	For this model, the following values were used as possible hyperparameters: max_depth: 10
SCV	For this model, the following values were used as possible hyperparameters: C:1

Results and Discussion

The precision, F1-score, recall, and accuracy metrics were obtained using precision_score, f1_score, recall_score and accuracy_score from the scikit-learn library. In addition, the ROC CURVE graph was made using roc_curve and AUC from the sci-kit-learn library for each machine-learning model.

The results obtained for the four machine learning models applied to a three-class classification problem indicate promising performance on the ROC curve. In this section, the results of each model used will be shown.

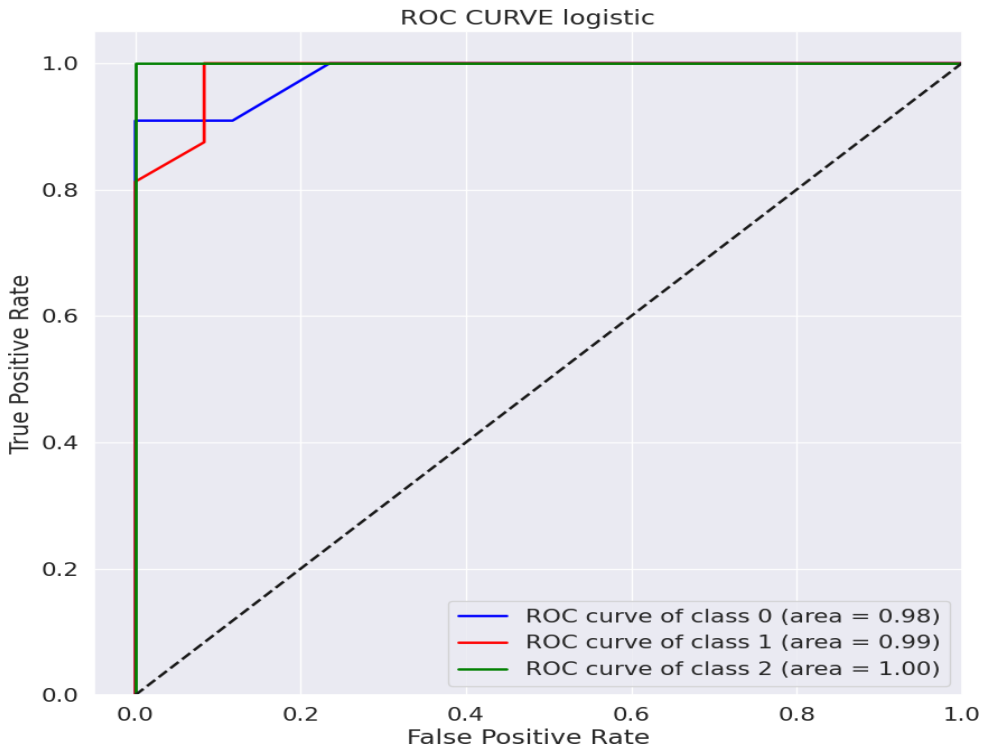

Figure 6. ROC Curve Logistic Regression

Figure 6 shows a logistic regression model, where the first class presents an area under the curve (AUC) of 0.98, the second class of 0.99, and the third class of 1.00, which suggests a high discrimination capacity in the classification.

Figure 7 shows us a decision tree model, where an AUC of 0.95 is observed for the first class, 0.93 for the second class, and 1.00 for the third class. This indicates that the model has good separation ability for most classes, although it shows lower accuracy for the second class.

Figure 8 shows us a Random Forest model, where the results are similar, with an AUC of 0.94 for the first class, 0.95 for the second class, and 1.00 for the third class. This demonstrates the ability of the model to effectively discriminate between the classes, although a slight decrease in accuracy is also observed for the first class.

Figure 9 shows us a Support Vector Machine model, where an AUC of 0.95 is obtained for the first class, 0.96 for the second class, and 1.00 for the third class, indicating good classification performance.

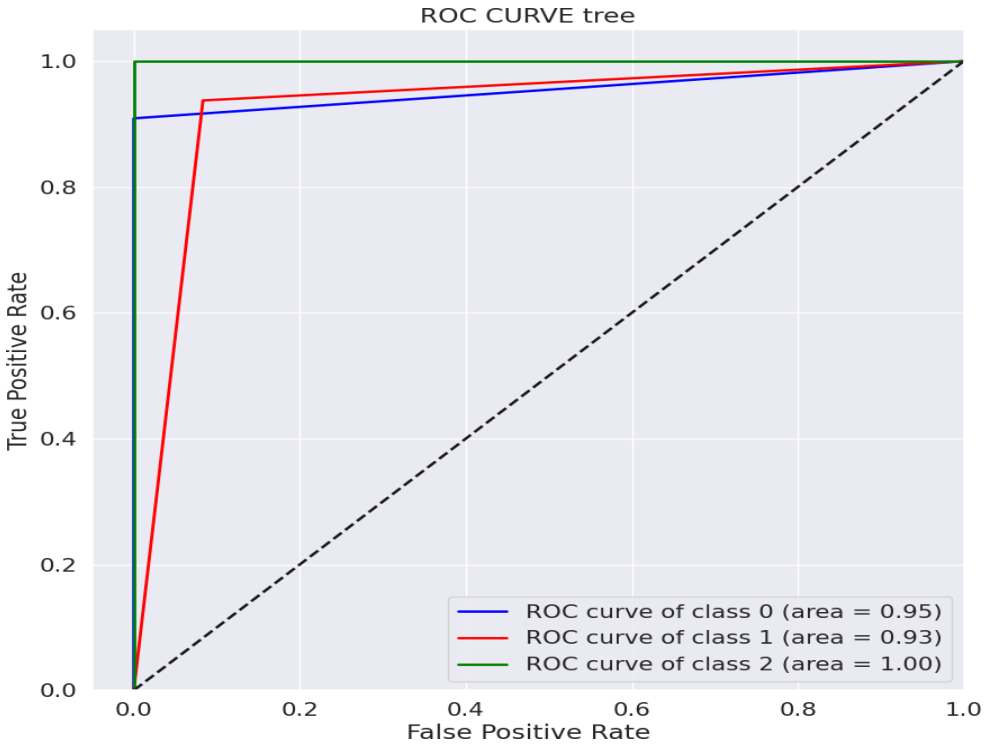

Figure 7. ROC Decision Tree Curve

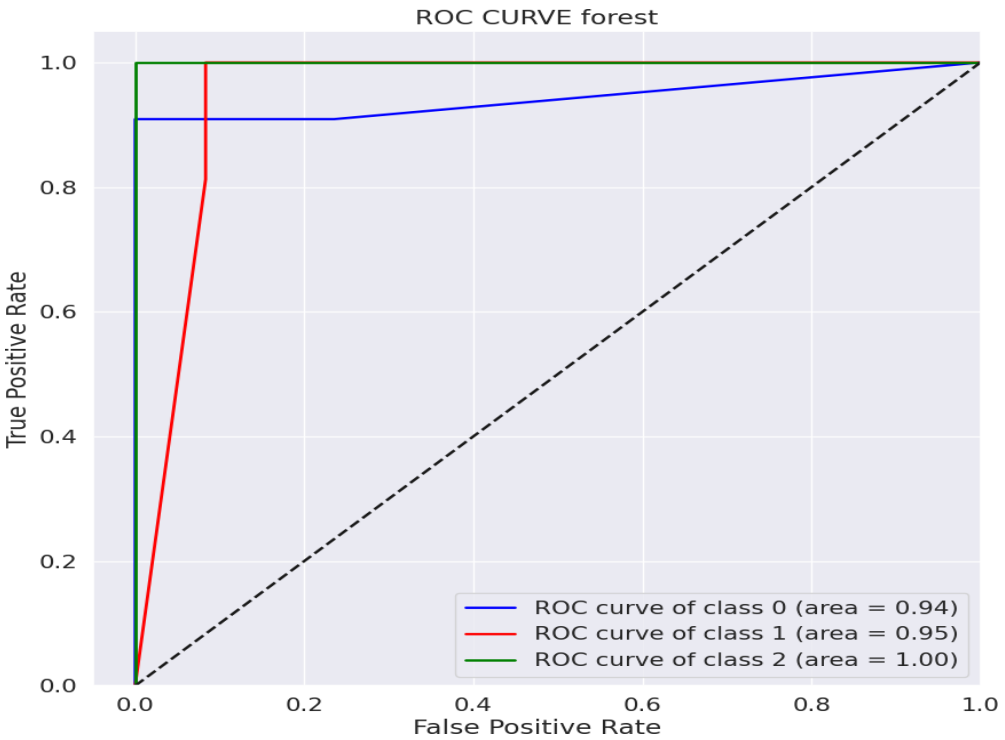

Figure 8. ROC Random Forest Curve

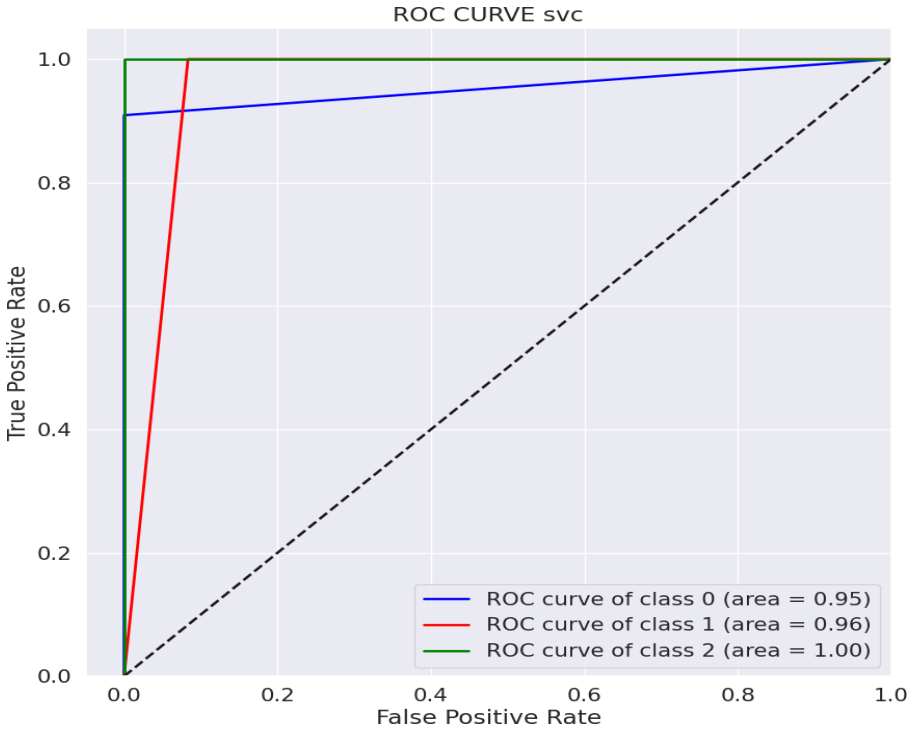

Figure 9. ROC Support Vector Machine Curve

These results suggest that all four models are promising for the three-class classification problem, with high AUC values in the ROC curve. However, it is important to note that other metrics, such as precision, recall, and F1-score, must also be considered for a complete evaluation of the performance of the models in terms of accurate classification of each class and the balance between them.

Table 5. Resultados de experimentos

Model	Precision	Fi-Score	Recall	Accuracy
Logistic Regression	0.980	0.974	0.970	0.964
Decision Tree	0.949	0.949	0.949	0.929
Radom Forest	0.980	0.974	0.970	0.964
SVM	0.980	0.974	0.970	0.964

Table 5 shows the summary results of evaluating the four machine learning models regarding precision, F1 score, recall, and accuracy. In the logistic regression model, a precision of 0.97, an F1-score of 0.96, a recall of 0.96, and an accuracy of 0.96 are obtained. In the case of the decision tree, a precision of 0.93, an F1-score of 0.93, a recall of 0.93, and an accuracy of 0.93 are observed. The Random Forest and Support Vector Machine (SVM) models present consistent results, with precision, F1-scores, recall, and accuracies of 0.97 and 0.96, respectively.
These results indicate that the models have good classification performance, high precision and recall rates, and an overall accuracy of 96%.

Conclusions

According to the results obtained, we can conclude that operational personnel's most significant number of dismissals occurs because they receive a better job offer, representing 54% of the total resignations between 2018 and 2021. This reason why It is classified as an unexpected event and as a decision of a personal nature with a negative impact on the company because it entails costs that are sometimes not noticeable at first glance; as Suzanne Lucas mentions, the decrease in productivity, since the person who leaves the position leaves the specific

functions assigned to him, which must be assumed by some of his colleagues, who do not necessarily have the specific knowledge to do so, thus also incurring other costs called work overload. And lost ability, respectively (as cited in Evaluar, parr. 1). It is the machine learning models that provide us with a demonstrably effective tool to determine the propensity for voluntary resignation of the operational staff of the cleaning and disinfection company, reaching 96% accuracy and thus reducing turnover rates.

Recommendations

It is recommended to incorporate the exit survey with previously defined questions as part of the offboarding process since the information obtained will be of great importance to continue feeding previous data and build rotation indicators and have better control of it. Likewise, it would be ideal to work on implementing a potential turnover survey of all workers who are still in the company to find the reasons that are causing discomfort or job dissatisfaction and increase the worker's desire to resign. To your workplace. Based on the results where the highest turnover percentage is due to a better job offer, reviewing and updating the existing benefits for workers and generating an approach according to the current age group is ideal. All issues related to organizational culture must also be reinforced to make each of its workers feel like a brand ambassador and build their loyalty. In this way, the best work environment will be maintained, thanks to the alert presented by the prediction model, and managers will be able to make better decisions with their immediate subordinates so that attrition does not occur.

References

Amazon, Aviable: https://aws.amazon.com/es/what-is/logistic-regression/, Accessed on June 14, 2023.

Aracena, C., Villena, F., Arias, F., and Dunstan, J., Aplicaciones de aprendizaje automático en salud. *Revista Médica Clínica Las Condes,* vol. 33, no. 6 , pp. 568 - 575, 2022.

Blanchard, W., Clark, R., Hui, F., Tian, R., and Woods, H., Dealing with complete separation and quasi-complete separation in logistic regression for linguistic data. *El Sevier,* vol. 2, no. 1, pp. 1-10, 2023.

Computrabajo, Aviable: https://recursos-empresa.computrabajo.com/el-costo-de-la-rotacion/, Accessed on August 13, 2023.

Editorial Ecoprensa, *El índice promedio de rotación laboral llega a 20,7% en las empresas peruanas, June 16, 2019,* Aviable: https://www.eleconomistaamerica.pe/empresas-eAm-peru/noticias/10257100/12/19/El-indice-promedio-de-rotacion-laboral-llega-a-207-en-las-empresas-peruanas.html, Accessed on July 21, 2023.

Ekawati, A. D., Predictive analitycs in employee churn: A systematic literature review. *Journal of Management Information and Decision Sciences,* vol. 22, no. 4, pp. 387-397, 2019.

Evaluar, Aviable: https://blogs.evaluar.com/por-que-la-rotacion-de-personal-cuesta-tanto, Accessed on August 13, 2023.

García Armendariz, J., Huertas Zúñiga, S., Lizárraga Portugal, C. A., Quiroz Flores, J. C., and Garcia Lopez, Y. J., Improving Demand Forecasting by Implementing Machine Learning in Poultry Production Company. *Repositorio de la Universidad de Lima,* vol. 8, pp. 39 - 45, 2023.

García Lopez, Y. J., Panduro, J., and Pumayari, S., Reduction of Backorders in the Cross Docking Sales Process for the Homecenter Order Service. *Repositorio de la Universidad de Lima,* vol. 12, no. 7, pp. 11, 2022.

Holwerda, J., Big data? Big deal: Searching for big data's. *El Sevier,* vol. 64, no.4, pp. 391-399, 2021.

King, K. G., Data Analytics in Human Resources: A Case Study and Critical Review. *Human Resource Development Review,* vol. 15, no. 4, pp. 487-495, 2016.

Mahesh, B., Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR), vol. 9*, no. 1, pp. 381 - 386, 2020.

Nocker, M., & Sena, V., Big Data and Human Resources Management: The Rise of Talent Analytics. *Social Sciences*, vol. 8, no. 10, pp. 273, 2019.

Omondiagbe, D., Veeraman, S., and Sidhu, A., Machine Learning Classification Techniques for Breast Cancer Diagnosis. *IOP Conference Series: Materials Science and Engineering*, vol. 495, pp. 012033, 2019.

Panduro, J. P. and García, Y.J., Use of a Machine Learning Model for the Reduction of Backorders in the Cross Docking Sales Process for the Homecenter Order Service. *Proceedings of the First Australian International Conference on Industrial Engineering and Operations Management, Sydney,* vol. 12, no. 7, pp. 2035 - 2045, 2022.

Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., and Tessore, J. P., Tratamiento masivo de datos utilizando técnicas de Machine Learning. *Repositorio Digital UNNOBA*, pp.131-134, 2016.

Siddharth, M., Hao, L. and He, J., Noninvasive fracture characterization based on the classification of sonic wave travel times. *El Sevier*, vol. 4, pp. 243 - 287, 2020.

Sujeet N., Dev R., and Yogesh P., Human Resource Predictive Analytics (HRPA) for HR Management in Organizations. *International Journal of Scientific & Technology Research,* vol. 5, no. 5, pp. 33-35, 2016.

Biographies

Erika Noemi Romero Rojas is a candidate to receive the title of industrial engineer from the Faculty of Engineering and Architecture of the University of Lima, Lima, Peru.

Garcia-Lopez Yvan Jesus is Ph.D. (c) in Engineering and Environmental Science, UNALM, "Master of Business Administration" from Maastricht School of Management, Holland, and a master's in strategic business administration from Pontificia Universidad Católica del Perú. "Master of Science" in Computer Science, Aerospace Technical Center - Technological Institute of Aeronautic, Brazil. Stage in Optimization of Processes and Technologies, University of Missouri-Rolla, USA, and Chemical Engineer from the National University of Callao. Specialization Study in Digital Transformation by Massachusetts Institute of Technology, Business Analytics, Wharton School of Management, Data Science by University of California, Berkeley, Big Data and Data Scientist by MITPro, USA Postgraduate Professor: Specialized Master from IT, MBA Centrum Católica, MBA from Calgary, Canada, and Centrum católica. Principal Consultant DSB Mobile, Executive director of Optimiza BG, advisor to the Office of Electronic Government and Information Technology (ONGEI) - PCM, Managing director of Tekconsulting LATAM, Executive director of Optimiza Business Group, Ex- Vice Dean of Information Engineering of the Universidad del Pacifico, Former Information Technology Manager of "MINERA CHINALCO PERU" Subsidiary of the Transnational Aluminum Corporation of China, Beijing, China. Former Manager of Systems and Communications of Maple Energy PLC, Director of Information Technology of Doe Run Peru SRL, Project Manager in implementation of ERP SAP, EBusiness Suite - Oracle Financial, and PeopleSoft. Process Analyst in transnational companies Fluor Daniel Corporation-USA, PETROBRAS-Brasil, Petróleos del Perú. He has over 25 years of extensive experience managing investment projects, execution, and commissioning in Peru, Colombia, USA, Brazil, and China.