Discovering Key Aspects to Reduce Employee Turnover Using a Predictive Model

Paula Andrea Cárdenas López(✉) and Marta Silvia Tabares Betancur

Universidad EAFIT, Medellín, Colombia
paulacardenas_25@hotmail.com

Abstract. High employee turnover is a phenomenon that occurs in different types of companies and often leads to losses that could affect the organization's productive continuity. This situation leads to the understanding, with solid evidence, of the factors that influence employees to leave their jobs and, thus, develop talent retention strategies proactively. This article proposes a predictive model to identify the most relevant factors that could cause employee turnover, specifically in the logistics process of a food and beverage production company. To achieve the objective, the CRISP-DM methodology was applied. Initially, various types of variables were identified, such as demographic, contractual, and payroll-related factors (N = 1517, period: 2017–2022). Then, five machine learning models, namely Logistic Regression, Random Forest, XGBoost, SVM, and AdaBoost, were trained, and optimal hyperparameters were used to improve the models' performance and generalization. The performance evaluation of these models was conducted using classification metrics and the construction of confidence intervals for the accuracy metric through non-parametric Bootstrap. The results obtained demonstrate that the XGBoost and Random Forest models show the highest AUC value, with a result of 99%. This indicates that variables such as work environment, years of service, salary, workplace location, and monthly salary deductions are the most significant factors influencing the evaluated human talent to leave their job. Therefore, it is possible to conclude that the aforementioned two models are accurate and reliable for predicting employee turnover in the logistics process of the analyzed company.

Keywords: Employee turnover · Machine Learning · ROC curve · Non-parametric Bootstrap · Survival analysis

1 Introduction

Voluntary employee turnover is one of the most studied topics in academic literature on organizational management. Aguado [1] presents this challenge as one of the main hurdles for human resources managers in the era of talent competition. Jain and Nayyar [16] explain that employee turnover represents a significant

M. Tabares et al. (Eds.): CCC 2023, CCIS 1924, pp. 380–395, 2024.
https://doi.org/10.1007/978-3-031-47372-2_30

challenge due to its negative impact on productivity, goal achievement, and competitiveness for organizations. For this reason, it is important to understand the reasons that lead employees to leave their positions and, based on that knowledge, redesign their human resources strategies and policies.

Garg et al. [14] propose focusing the prediction on turnover intentions through the use of machine learning models. They suggest a wide range of factors that may be related to the probability of withdrawal, which varies based on employees' location, culture, and job profiles. Sisodia et al. [23] mention that work-related factors such as the number of promotions, salary increase, latest evaluation, tenure in the company, and working hours are fundamental attributes for the classification models used to predict turnover. They also include additional attributes created based on hypotheses, such as employees with promotions but no salary change, employees without promotions in the last 5 years, among others. Liu et al. [21] identified that factors such as age, gender, place of birth, marital status, and family responsibilities are related to the individual and impact turnover. According to the above, most of the data comes from human resource database systems that are not very efficient in prediction and modeling, resulting in less accurate models. Therefore, to achieve more accurate predictions, a machine learning model, specifically XGBoost, is proposed, known for its robustness and ability to significantly improve predictions [16].

However, the previous proposals do not consider external variables such as the unemployment rate, poverty index, and Gini coefficient as relevant factors when understanding employee turnover. In this article, several supervised learning algorithms were applied with the aim of classifying whether an employee will leave the company or not. Additionally, the most important features in the models were identified, which can serve as a basis for decision-makers to improve employee turnover rate. The study also employed the survival analysis technique with the Kaplan-Meier approach to understand how an event is impacted over time.

Several classification models were used, such as Logistic Regression, Random Forest, XGBoost, SVM, and AdaBoost, which were tuned with optimal hyperparameters to improve evaluation metrics. The model performance was assessed using metrics such as accuracy, precision, recall, F1-score, area under the curve (AUC), and 5-fold cross-validation. Additionally, confidence intervals for accuracy were generated using non-parametric Bootstrap.

The CRISP-DM methodology was employed for the development of this study, providing a solid structure for prediction and machine learning projects. Its flexibility allows for iteration and error correction in different phases, addressing challenges in feature extraction, data cleaning, and model performance, seeking continuous improvements at each step. This adaptability is crucial to achieve accurate and reliable results in model development [18].

The document is organized as follows: Sect. 2 presents a brief review of the literature associated with employee turnover. Section 3 presents the development of the proposed model using the CRISP-DM methodology. Section 4 presents the discussion and conclusions.

2 Literature Review

In recent years, organizations have recognized the need to predict employee turnover and identify the key factors that contribute to retaining their workforce in the future. Alaskar et al. [9] have shown that this significantly improves the accuracy of the models. In this study, the most effective model was found to be SVM with SelectKBest feature selection, achieving an accuracy of 97%, followed by Decision Tree with RFE feature selection and an accuracy of 96.2%. Yahía et al. [3], based on the previous approach, found that the Vapnik-Chervonenkis (VC) model achieved the highest accuracy at 99%, followed by Random Forest at 98.3%. The researchers suggest that these types of models should be updated in real-time, as employee characteristics change constantly, thereby impacting the model's results.

On the other hand, Zhao et al. [27] identify that predicting employee turnover has several complexities, including the trade-off between the number of variables and the number of records. In some cases, the Hughes phenomenon, also known as the curse of dimensionality, may arise. In this case, they suggest carefully handling and understanding the dataset to deliver models and results that align with reality.

Srivastava and Eachempati [24] aim to determine the effectiveness of deep learning compared to machine learning using models such as Random Forest and Gradient Boosting. The obtained results are validated using a regression model and a Fuzzy Analytic Hierarchy Process (AHP) that considers the relative importance of the variables. The findings reveal that deep neural networks achieve an accuracy of 91.6%, whereas Random Forest and Gradient Boosting reach 82.5% and 85.4%, respectively.

Cai et al. [5] propose the approach of DBGE (Dynamic Bipartite Graph Embedding) as a method that combines employee data obtained from professional social networks with company information. In this approach, the history of employees' job experiences, including start and end dates, is modeled as a dynamic bipartite graph, which is then used to generate machine learning models. By implementing this approach, the accuracy of the models was improved by 1.5%, with Random Forest being the best-performing model with an accuracy of 89.3%.

Fan et al. [12] propose predicting turnover using unsupervised learning through a hybrid method known as Self-Organizing Map (SOM). This method clusters individual features using backpropagation neural networks, achieving an accuracy of 92%.

3 Development of the Proposed Model

For the development of the model, the CRISP-DM methodology (Cross-Industry Standard Process for Data Mining) [2] is used, specifically for the phases of business understanding, data understanding, data preparation, modeling, and evaluation. The deployment phase is presented as a proposal for future work to put the insights generated by the models into practical use within the organization.

3.1 Business Understanding

The purpose of this article is to address the problem of employee turnover in a logistics process of a food and beverage production company in Colombia, specifically focusing on three workplaces located in the cities of Bello, Bogotá, and Yumbo.

Employee turnover poses various challenges for the organization, including tangible financial costs related to the hiring and training process of new employees, as well as intangible costs associated with the loss of knowledge and skills acquired through experience. Additionally, this turnover can negatively impact productivity and corporate image.

Therefore, the objective of this study is to predict whether an employee will voluntarily leave the company and, at the same time, identify key factors influencing the decision to leave. This will provide a solid and evidence-based foundation for human resources managers to make strategic decisions regarding talent retention, based on data and analysis, enabling them to build an organizational management system with respect to human resources.

3.2 Data Understanding

The dataset comprises information collected from January 2017 to July 2022 and includes data from 1,517 unique employees. Each of these employees is associated with 32 different variables. The classification label 'Turnover' is used to indicate whether an employee leaves the company or not. Table 1 shows the collected variables along with their corresponding descriptions.

Table 1. Description of dataset variables

Variable	Description
Employee Id	Employee identification
Joining Date	Date of joining the company
Leaving Date	Date of leaving the company
Age at retirement	Employee's age at the time of retirement
Gender	Female, Male
Department	Employee's area (General Warehouse, finished goods warehouse, Dispatch, Distribution, Vehicle Maintenance)
City	Workplace location (Bello, Bogotá, Yumbo)
Job position	Job position (Operator, Driver, Warehouse Clerk, distribution Supervisor, among others)
Employment contract type	Contract type (Fixed, Indefinite)
Employment group	Employment group (Direct, Temporary)
Salary	Monthly Income
Marital status	Marital status (Married, Divorced, Separated, Single, Common-law)

(*continued*)

Table 1. (*continued*)

Variable	Description
Children	(0 = No, 1 = Yes)
Years at company	Total years in the company
Promotions	Internal promotions during the employment contract
Salary range structure	Employee's position relative to the salary curve (1 = underpaid below 80%, 2 = paid fairly between 80% and 120%, 3 = overpaid above 120%)
Distance From Home	The Distance From Work To Home
Climate surveys	Factors evaluated in the organizational climate survey (1 = Motivation and recognition, 2 = Learning, 3 = Collaboration, 4 = Commitment and Identity, 5 = Working Conditions, 6 = Job Relationship, 7 = Leadership, 8 = Communication)
Exit Survey	Reasons for retirement marked by the employee at the time of leaving (0 = Work Environment, 1 = Personal or Professional Growth, 2 = Development and Training, 3 = Distance From Work To Home, 4 = Work Flexibility, 5 = Current Job Duties, 6 = Working Hours, Workload and Tools, 7 = Family Reasons, 8 = Not Registered, 9 = Recognition and Appreciation, 10 = Relationship with Managers, 11 = Salary or Benefits)
Workplace accident	An employee voluntarily leaves if they have absences due to work accidents exceeding 3 days. (0 = No, 1 = Yes)
Maternity/paternity leave	An employee voluntarily leaves if they return from a parental leave before 6 months. (0 = No, 1 = Yes)
Absenteeism record	An employee voluntarily leaves if they have a record of absences in the last 3 months before their retirement. (0 = No, 1 = Yes)
Return from vacation	An employee voluntarily leaves if they return from vacation the following month. (0 = No, 1 = Yes)
Sundays and holidays worked	An employee voluntarily leaves if they work more than 3 Sundays and holidays per month, one month before their retirement. (0 = No, 1 = Yes)
Monthly overtime hours limit	An employee voluntarily leaves if they generate more than 48 overtime hours per month. (0 = No, 1 = Yes)
Overtime hours increment	An employee voluntarily leaves if they experience a significant 30% increase in overtime hours. (0 = No, 1 = Yes)
Monthly deductions	Percentage of monthly deductions relative to their income.
Unused vacation days	An employee voluntarily leaves if they have more than two periods of unused vacation. (0 = No, 1 = Yes)
Benefit usage	An employee voluntarily leaves if they have not claimed any benefits from the pact or convention in the last year. (0 = No, 1 = Yes)
Job position tenure	An employee voluntarily leaves if they have remained in the same position for the last 3 years. (0 = No, 1 = Yes)

Additionally, three external variables obtained from DANE related to the labor market are available: unemployment rate [10], monetary poverty, and Gini coefficient [11]. These variables, which experience variations over time, have been integrated into the dataset by assigning to each record the annual unemployment indicator value, uniformly distributed throughout each corresponding retirement year.

It is important to mention that the salary variable was captured considering each individual's salary at the time of their retirement. Additionally, the annual increment adjustment that the company applied was used to bring the salary value of those who retired before 2022 to the current value. In this way, a fair and up-to-date comparison of salaries was ensured for employees who made the decision to retire at different times, allowing for a more accurate and consistent evaluation of their influence on the target variable.

As proposed by Alaskar [9], **Exploratory Data Analysis (EDA)** was employed to identify relevant patterns influencing both employee turnover and the relationships between variables, using graphical techniques and correlation analysis.

During the exploration, it was found that the percentage of employees who voluntarily left was higher compared to those who stayed, with 67% and 33%, respectively. This indicates a class imbalance in the output label, which will be addressed with SMOTE (Synthetic Minority Oversampling TEchnique) in the data preparation step.

Next, some relevant findings are presented:

- The finished product storage process has the largest number of employees, representing 66.5% of the total workforce, followed by the distribution processes with 15%.
- 80% of the employees hold operator positions.
- The marital status of employees shows that 45% are single, while 40% are in a common-law relationship.
- The analysis of distance between home and workplace reveals that employees in Bello have an intermediate distance of 6.42 km. On the other hand, employees in Bogotá have a longer median distance of 10.22 km, while employees in Yumbo have the longest median distance of all, with 12.16 km.
- A total of 40.23% of the retired personnel who have completed the retirement survey or interview have indicated that their main reason for leaving the company is related to long working hours, demanding work schedules, and the physical burden associated with their job. This is in addition to considering both professional and personal growth.

When reviewing the graphical representation of variables to identify outliers in the dataset, it can be observed from Fig. 1 that the median age of retired employees is 28 years, while for active employees, it is 35 years.

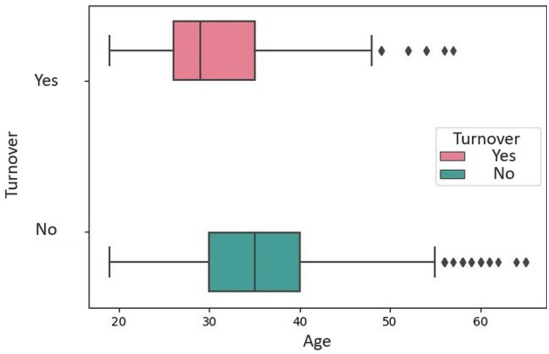

Fig. 1. Boxplot turnover vs age

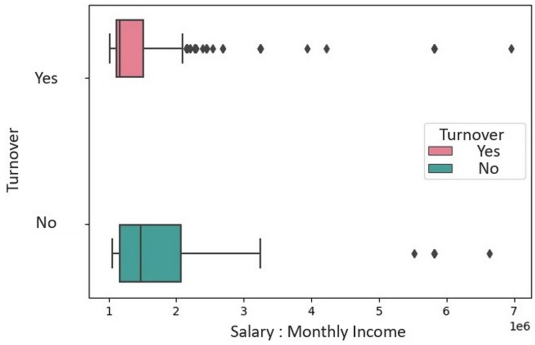

Fig. 2. BoxPlot turnover vs Salary: monthly income

The analysis of Fig. 2 reveals significant differences in salaries between retired and active employees. While the median salary is lower for retired employees (1,161,000), active employees show a higher median (1,471,000) and a wider interquartile range (1,161,000–2,064,000).

In this phase, the statistical technique used to study the time it takes for an event to occur is also included, known as **Survival Analysis**. Survival analysis methods are mainly used to analyze prospective data collected over time, where the time to the occurrence of an event of interest is recorded, allowing for the evaluation of the relationship between explanatory variables and the occurrence of the event, as addressed by Kartsonaki [20].

According to the Kaplan-Meier (KM) or product-limit approach [19], the survival analysis will be applied in this article, as its non-parametric method does not make assumptions about the distribution of the data but rather performs the estimation that best fits the data. The estimation is defined by:

$$\hat{P}(t) = \prod_{r} \left[\frac{(N-r)}{N-r+1} \right] \tag{1}$$

where N is the number of observations (either censored or not), in ascending order of magnitude $0 \leq t_1{}' \leq t_2{}' \leq ... \leq t_N{}'$, where r assumes those values $t_r{}' \leq t$, for which $t_r{}'$ measures the time to event occurrence. This estimation is the distribution that maximizes the likelihood of the observations.

The main objective of this survival analysis is to understand the factors that influence the duration of time that employees remain in the company before leaving. In Fig. 3, it can be observed that employees hired through temporary agencies have a 0.6 probability of surviving for 1 year, while directly hired employees by the company have a 50% probability of lasting for 6 years.

The KM curve in Fig. 4 displays the survival curves by marital status, where it can be seen that Single employees have a survival curve that declines faster over time compared to those who are in a Common-Law Relationship or Married. Therefore, Married employees have a higher probability of survival over time.

3.3 Data Preparation

Preparing the data before carrying out the modeling increases the chances of obtaining accurate results. In this case, a preprocessing phase was executed, which included detecting missing values, outliers, analyzing correlations between features, as well as data transformation and normalization.

It's important to highlight the results of the correlation analysis that reveal significant associations between the following variables:

- There is a strong correlation between the Gini Index and the Monetary Poverty Indicator.
- A notable correlation is observed between the Unemployment Rate and the Monetary Poverty Indicator.
- There is also a significant relationship between Retirement Age and Years Worked.

Some variables mentioned in the data understanding phase, including workplace accidents, maternity leave, absenteeism history, vacation returns, monthly holidays, monthly overtime limit, overtime increase, unused vacations, benefit utilization, and tenure, were generated by constructing new features from existing data. These new features not only enriched the dataset but were also developed with the target variable in mind, which is whether a person decides to retire from the company or not. As a result, these created variables not only enhance our understanding of the data but also directly contribute to the quality and relevance of the subsequent analysis and modeling.

Significant to note is that the dataset exhibited class imbalance in the output label, as only 33% of employees remained in the company. To address this issue, the SMOTE (Synthetic Minority Oversampling TEchnique) technique was implemented, as introduced in the study by Chawla [6]. This technique generates synthetic samples by exploring the feature space instead of the data space, thus overcoming the problem of overfitting. It was the first technique that introduced new samples into the training set to improve the data.

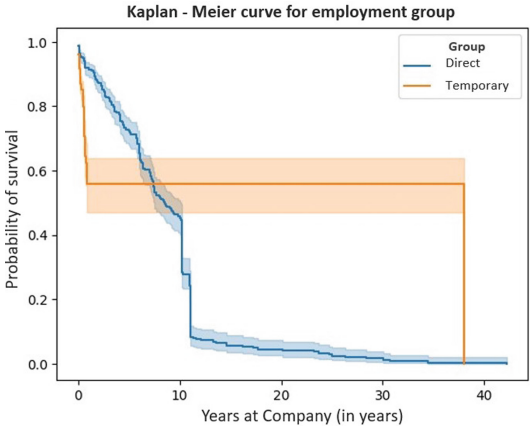

Fig. 3. A Survival Curve Based on Employment Group

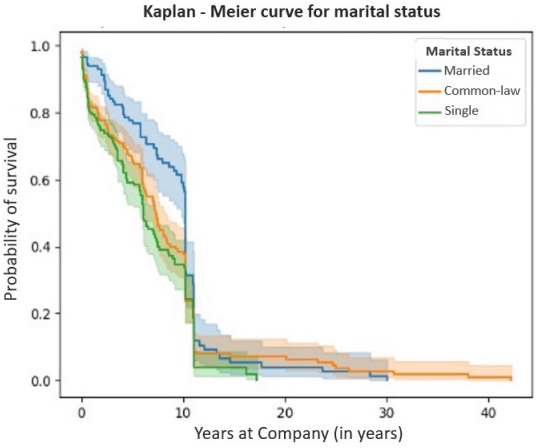

Fig. 4. A Survival Curve Based on Marital Status

3.4 Modeling

In this stage, a set of classification models are executed to predict employee turnover, such as Logistic Regression [17], SVM [15], Random Forest [4], XGBoost [7], and Adaboost [13]. The optimization of each model was carried out using GridSearchCV, aiming to find the optimal combination of hyperparameters that maximizes the model's performance [22]. The best values and combinations of parameters for each model, based on their characteristics, are presented in Table 2:

Table 2. Optimal hyperparameters

Classifiers	Hyperparameters
Logistic Regression	'C': 1.0, 'penalty': 'l1'
Random Forest	'criterion': 'gini', 'max_depth': 22, 'n_estimators': 100
XGBoost	'learning_rate': 0.1, 'max_depth': 15, 'n_estimators': 200
SVM	'C': 10, 'kernel': 'rbf'
AdaBoost	'base_estimator_max_depth': 15, 'learning_rate': 0.1, 'n_estimators': 50

The dataset was partitioned into training and testing sets, where 70% of the total records, equivalent to 1,402, were used for training the models using the **SMOTE** technique. The remaining 30%, corresponding to 456 records, were reserved for the testing phase. This division allowed for a significant portion of data to be used for training the models and properly evaluating their performance on an independent sample during the testing phase.

Additionally, the main features of each model were detected, revealing their contribution in terms of weight or influence on the predictions, using their relative importance. For this purpose, the feature_importances_ function was employed, as mentioned by Liu [21].

Table 3 presents the ten most important features for each model. A high contribution of organizational climate is observed, as well as the variable Absenteeism record, which refers to the hypothesis about the incidence of increased absenteeism 3 months before voluntary departure, among others. These findings highlight the importance of these features in predicting turnover and provide valuable insights into the key factors that influence employees' decision to leave the organization.

Table 3. Top 10 key features by model

Classifiers	Feature Importances
Logistic Regression	Organizational Climate, Expired Vacation Days, Position: 'Operator', Reason for Departure: 'Current Job Responsibilities', Position: 'Sales Driver', City: 'Bello', Employment Group: 'Directs', Employment Group: 'Temporaries', Department: 'Vehicle Maintenance', Tenure in Position
Random Forest	Organizational Climate, Years at Company, Monthly Deductions, Retirement Age, Monthly Sunday Premiums, Distance from Home, Absenteeism Record, Salary, Employment Group: 'Directs', Monthly overtime hours limit
XGBoost	Organizational Climate, Years at Company, Salary, City: 'Bogotá', Monthly Deductions, City: 'Yumbo', Expired Vacation Days, Absenteeism Record, Reason for Departure: 'Development and Training', Children
SVM	Organizational Climate, Tenure in Position, Employment Group: 'Temporaries', Absenteeism Record, Department: 'Distribution', Expired Vacation Days, Monthly overtime hours limit, Position: 'Operator I', Department: 'Finished goods warehouse', Position: 'Sales Driver'
AdaBoost	Organizational Climate, Years at Company, Salary, Monthly Deductions, Expired Vacation Days, City: 'Yumbo', Distance from Home, Absenteeism Record, Position: 'Distribution Analyst', Gender: 'Male'

3.5 Evaluation

To evaluate the performance of each model, appropriate classification metrics such as accuracy, precision, recall, F1-score, and AUC were employed. These metrics provided a comprehensive view of the predictive model's performance, allowing for the assessment of its ability to make accurate predictions and distinguish between positive and negative classes, as proposed by Wang and Zhi [25].

The validation of the model training results was carried out using the test dataset, where XGBoost achieved the highest F1-score of 97.38%, followed by Random Forest with 96.77%. Table 4 presents the evaluation metric results on the test dataset, specifically for the "Turnover" target label.

Table 4. Prediction analysis result

Classifiers	Accuracy	Precision	Recall	F1-score	AUC
Logistic Regression	0.8552	0.8722	0.9130	0.8921	0.9396
Random Forest	0.9556	0.9584	0.9720	0.9677	0.9891
XGBoost	0.9649	0.9520	0.9966	0.9738	0.9880
SVM	0.8793	0.8785	0.9417	0.9090	0.9518
AdaBoost	0.9517	0.9361	0.9932	0.9638	0.9529

In order to assess whether the model could improve its performance, k-fold cross-validation with $k = 5$ was employed. In this step, the dataset was divided into five equal parts, with four parts used for model training and the remaining part as the validation set. This process was repeated five times, each time using a different part as the validation set. Finally, the accuracy results from the five iterations were averaged to obtain a final estimation of the model's performance (Table 5).

XGBoost is the model with the highest average accuracy in the cross-validation, followed by AdaBoost. See Table 6 for details.

Table 5. Cross-validation Results - Accuracy

Classifiers	Accuracy
Logistic Regression	0.8630
Random Forest	0.9550
XGBoost	0.9658
SVM	0.8994
AdaBoost	0.9657

As part of the validation process for the evaluation metrics of the models, the **Non-Parametric Bootstrap** method was used to construct confidence

intervals for the accuracy of each model. Wasserman [26] introduces the Bootstrap as a method used to estimate the variance and distribution of a statistic $T_n = g(X_1, ..., X_n)$, also used in constructing confidence intervals. In this case, the interval was constructed using percentile intervals, defined by:

$$C_n = \left(T^*_{(B\alpha/2)}, T^*_{(B(1-\alpha/2)}\right) \tag{2}$$

C_n: the interval is called the confidence interval, with n denoting the sample size, and C signifying that it is an interval.

($T^*_{(B\alpha/2)}$: It is the $\alpha/2$ percentile of the bootstrap statistics. It represents the value corresponding to the lower limit of the confidence interval.

$T^*_{(B(1-\alpha/2)}$: It is the $(1-\alpha/2)$ percentile of the bootstrap statistics. It represents the value corresponding to the upper limit of the confidence interval.

By employing this method, 1,000 random Bootstrap samples were generated with a 95% confidence level to calculate the confidence intervals for accuracy. The outcomes reveal that the accuracy estimation for XGBoost is more precise in comparison to the other classifiers mentioned, as its confidence interval is narrower than theirs. Additionally, it can be observed that the confidence intervals of XGBoost, Random Forest, and AdaBoost overlap, indicating that there is no statistically significant difference between the metric estimations. This places XGBoost and Random Forest as the top-performing models. See Table 6.

Table 6. Confidence Intervals of Accuracy with Bootstrap

Classifiers	Accuracy
Logistic Regression	[0.8531, 0.8882]
Random Forest	[0.9298, 0.9583]
XGBoost	[0.9408, 0.9649]
SVM	[0.8377, 0.8750]
AdaBoost	[0.9057, 0.9605]

Lastly, the ROC curve is analyzed to assess the balance between true positives and false positives in the test set. In Fig. 5, it can be observed that XGBoost yields the best results, as it has the highest area under the curve (AUC) with a value of 99%, followed by Random Forest. However, XGBoost is chosen due to its cutoff point being closer to the upper-left vertex of the graph, indicating a high true positive rate (TPR) and a low false positive rate (FPR) [8].

In Fig. 6, the most significant variables in predicting employee turnover with the XGBoost model are shown. It is evident that Leadership, as measured through organizational climate, years at company, and Salary as monthly income are the variables with the highest importance when predicting turnover. This implies that these variables have a positive or negative influence on the decision to leave the company.

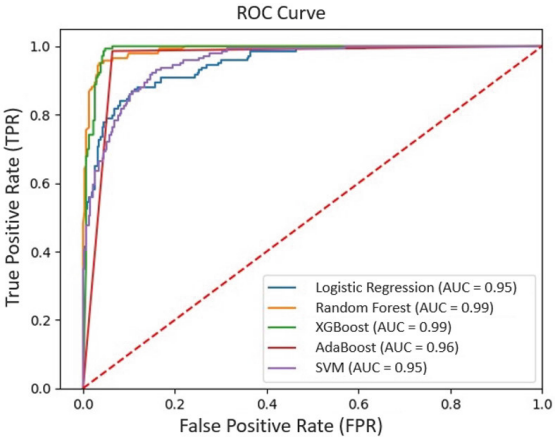

Fig. 5. ROC Curve by model

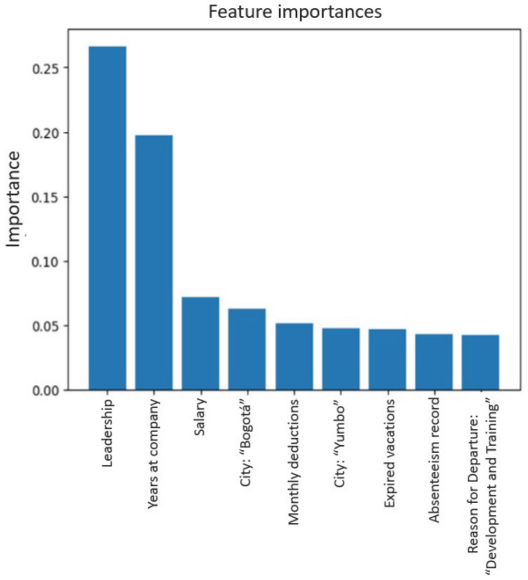

Fig. 6. Weight of variables based on XGBoost

4 Discussion and Conclusions

This article presents a method for predicting whether an employee will leave the company or not, along with the most influential factors in their decision. The method involved the execution of five predictive classification models, namely Logistic Regression, Random Forest, XGBoost, SVM, and AdaBoost. Various important classification metrics were examined to find the best model, and cross-

validation and non-parametric confidence intervals were applied using the percentile Bootstrap method.

Based on the findings, it is concluded that the XGBoost model is superior in this type of problem, achieving an accuracy of 96.58% in cross-validation, proving to be efficient and high-performing. Precisely identifying employees with a higher probability of leaving the company is essential for making strategic decisions. By recognizing these potential employees, decision-makers can implement proactive measures, such as retention plans, as early as possible.

An initial focus for these strategies should be on the five most important characteristics identified in all models: work climate, absenteeism record, unused vacation days, years of service, and salary. Additionally, the insights obtained during the exploratory data analysis can serve as valuable inputs for understanding the workforce.

The project also includes a non-parametric survival analysis using the Kaplan Meier approach, which highlights the importance of employee contract type and marital status concerning their tenure in the company. This can further inform and tailor human resource strategies.

In the future, it is expected that this method will be used in all organizations, creating a data-driven decision-making system regarding human resources, enabling the focus on retention strategies and even attracting new talent.

References

1. Aguado, D.: HR Analytics: Teoría y práctica para una analítica de recursos humanos con impacto, vol. 1. ESIC Editorial, primera edición edn. (2018)
2. Ayele, W.: Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset. Int. J. Adv. Comput. Sci. Appl. **11**, 20–32 (2020). https://doi.org/10.14569/IJACSA.2020.0110603
3. Ben Yahia, N., Jihen, H., Colomo-Palacios, R.: From big data to deep data to support people analytics for employee attrition prediction. IEEE Access **9**, 60447–60458 (2021). https://doi.org/10.1109/ACCESS.2021.3074559
4. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
5. Cai, X., et al.: DBGE: employee turnover prediction based on dynamic bipartite graph embedding. IEEE Access **8**, 10390–10402 (2020). https://doi.org/10.1109/ACCESS.2020.2965544
6. Chawla, N.V.: Data mining for imbalanced datasets: an overview. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 875–886. Springer, Boston (2009). https://doi.org/10.1007/978-0-387-09823-4_45
7. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system, vol. 13–17-Augu, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785
8. Choudhury, P., Allen, R., Endres, M.: Machine learning for pattern discovery in management research. Strateg. Manag. J. **42**, 30–57 (2021). https://doi.org/10.1002/smj.3215
9. Alaskar, L., Crane, M., Alduailij, M.: Employee turnover prediction using machine learning. In: Alfaries, A., Mengash, H., Yasar, A., Shakshuki, E. (eds.) ICC 2019. CCIS, vol. 1097, pp. 301–316. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-36365-9_25

10. DANE: Empleo y desempleo (2022). https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo
11. DANE: Pobreza monetaria y pobreza monetaria extrema (2022). https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-monetaria
12. Fan, C.Y., Fan, P.S., Chan, T.Y., Chang, S.H.: Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Syst. Appl. **39**, 8844–8851 (2012). https://doi.org/10.1016/j.eswa.2012.02.005
13. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997). https://doi.org/10.1006/jcss.1997.1504. https://www.sciencedirect.com/science/article/pii/S002200009791504X
14. Garg, S., Sinha, S., Kar, A.K., Mani, M.: A review of machine learning applications in human resource management. Int. J. Product. Perform. Manag. **71**(5), 1590–1610 (2022)
15. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2009). https://doi.org/10.1007/978-0-387-84858-7
16. Jain, R., Nayyar, A.: Predicting employee attrition using XGBoost machine learning approach, pp. 113–120 (2018). https://doi.org/10.1109/SYSMART.2018.8746940
17. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning, vol. 103. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-7138-7
18. Kannengiesser, U., Gero, J.: Modelling the design of models: an example using CRISP-DM. Proc. Des. Soc. **3**, 2705–2714 (2023). https://doi.org/10.1017/pds.2023.271
19. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **53**(282), 457–481 (1958). http://www.jstor.org/stable/2281868
20. Kartsonaki, C.: Survival analysis. Diagnostic Histopathol. **22**(7), 263–270 (2016). https://doi.org/10.1016/j.mpdhp.2016.06.005. https://www.sciencedirect.com/science/article/pii/S1756231716300639, mini-Symposium: Medical Statistics
21. Liu, J., Long, Y., Fang, M., He, R., Wang, T., Chen, G.: Analyzing employee turnover based on job skills, pp. 16–21 (2018). https://doi.org/10.1145/3224207.3224209
22. Probst, P., Wright, M.N., Boulesteix, A.L.: Hyperparameters and tuning strategies for random forest. Wiley Interdisc. Rev. Data Min. Knowl. Discov. **9**(3), e1301 (2019)
23. Sisodia, D., Vishwakarma, S., Pujahari, A.: Evaluation of machine learning models for employee churn prediction, pp. 1016–1020 (2018). https://doi.org/10.1109/ICICI.2017.8365293
24. Srivastava, D.P., Eachempati, P.: Intelligent employee retention system for attrition rate analysis and churn prediction: an ensemble machine learning and multi-criteria decision-making approach. J. Glob. Inf. Manag. **29**, 1–29 (2021). https://doi.org/10.4018/JGIM.20211101.oa23
25. Wang, X., Zhi, J.: A machine learning-based analytical framework for employee turnover prediction. J. Manag. Anal. **8**, 351–370 (2021). https://doi.org/10.1080/23270012.2021.1961318

26. Wasserman, L.: All of Nonparametric Statistics. Springer, New York (2006). https://doi.org/10.1007/0-387-30623-4
27. Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B., Zhu, X.: Employee turnover prediction with machine learning: a reliable approach. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) IntelliSys 2018. AISC, vol. 869, pp. 737–758. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-01057-7_56