Data-driven talent management: predictive modeling for human capital decisions in the BPO industry

Gestión del talento basada en datos: modelado predictivo para decisiones sobre capital humano en la industria BPO

Gestão de talentos orientada por dados: modelagem preditiva para decisões de capital humano no setor de BPO (Business Process Outsourcing)

Diego Alejandro Díaz Fonseca[1]
Oswaldo Alberto Romero Villalobos[2]
Julio Barón Velandia[3]

Received: December 15th, 2025
Accepted: March 13th, 2026
Available: May 5th, 2026

How to cite this article:
D. A. Díaz Fonseca, O. A. Romero Villalobos, and J. Barón Velandia, "Data-Driven Talent Management: Predictive Modeling for Human Capital Decisions in the BPO Industry," *Revista Ingeniería Solidaria,* vol. 22, no. 2, 2026.
doi: https://doi.org/10.16925/2357-6014.2026.02.09

Research article. https://doi.org/10.16925/2357-6014.2026.02.09

[1] Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
Email: dadiazf@udistrital.edu.co
ORCID: https://orcid.org/0009-0000-3517-0382
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000082407

[2] Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
Email: oromerov@udistrital.edu.co
ORCID: https://orcid.org/0000-0001-8308-3128
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000607428

[3] Universidad Distrital Francisco José de Caldas, Bogotá, Colombia.
Email: jbaron@udistrital.edu.co
ORCID: https://orcid.org/0000-0002-9491-5564
CvLAC: https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0000292931

Abstract

Introduction: This article is the result of the research project "Model for Profiling and Personnel Selection Applied to the BPO Sector in Colombia Based on Predictive Analytics Techniques and Genetic Algorithms", developed at Universidad Distrital Francisco José de Caldas in 2025.

Problem: The organization under analysis exhibited early attrition of 62.82%, resulting in high costs associated with recruitment, training, and the learning curve, which revealed limitations in traditional recruitment processes.

Objective: To design and evaluate a predictive model that identifies candidates with the highest probability of at least 3 months of tenure and determines the most influential attributes for such retention.

Methodology: A mixed approach was adopted. Relevant variables were initially identified through consultation with recruitment professionals; subsequently, machine learning models were trained on 560 hiring records. Algorithms including SVC, Random Forest, and KNN were applied. Class imbalance was addressed using SMOTE, and performance was evaluated using metrics such as accuracy, sensitivity, specificity, and F1-score.

Results: The proposed model reduced early attrition, through simulation, from 62.82% to 4.11%, representing a relative reduction of 93.6%, along with estimated savings of up to COP $1B in one quarter. The SVC model stood out with a specificity of 94.44%, minimizing high-risk hires.

Conclusions: Predictive models in recruitment significantly improve retention and resource optimization.

Originality: The study links predictive accuracy with measurable financial impact in the Colombian BPO sector.

Limitations: Sensitive attributes were not collected to prevent discriminatory bias.

Keywords: BPO, Employee Selection, Human Resources, Machine Learning, Predictive Models.

Resumen

Introducción: Este artículo es producto de la investigación "Modelo de perfilación y selección de personal aplicado al sector BPO en Colombia basado en técnicas de predictive analytics y algoritmos genéticos", desarrollada en la Universidad Distrital Francisco José de Caldas en el año 2025.

Problema: La organización bajo análisis presentó una deserción temprana de 62.82%, evidenciando altos costos asociados con el reclutamiento, la capacitación y la curva de aprendizaje, lo cual implicaba limitaciones en los procesos tradicionales de reclutamiento.

Objetivo: Diseñar y evaluar un modelo predictivo que identifique a los candidatos con mayor probabilidad de al menos 3 meses de permanencia y determine los atributos más influyentes para dicha retención.

Metodología: Se adoptó un enfoque mixto. Las variables relevantes fueron inicialmente identificadas mediante consulta con profesionales de reclutamiento; posteriormente, se entrenaron modelos de aprendizaje automático sobre 560 registros de contratación. Se aplicaron algoritmos como SVC, Random Forest y KNN. El desbalance de clases fue abordado utilizando SMOTE y el desempeño fue evaluado mediante métricas como exactitud, sensibilidad, especificidad y F1-score.

Resultados: El modelo propuesto redujo la deserción temprana mediante simulación de 62.82% a 4.11%, representando una reducción relativa de 93.6%, junto con ahorros estimados de hasta COP $1B en un trimestre. El modelo SVC se destacó con una especificidad de 94.44%, minimizando contrataciones de alto riesgo.

Conclusiones: Los modelos predictivos en el reclutamiento mejoran significativamente la retención y la optimización de recursos.

Originalidad: El estudio vincula la precisión predictiva con un impacto financiero medible en el sector BPO colombiano.

Limitaciones: No se recolectaron atributos sensibles para prevenir sesgo discriminatorio.

Palabras clave: Aprendizaje automático, BPO, modelos predictivos, recursos humanos, selección de personal.

Resumo

Introdução: Este artigo é produto do projeto de pesquisa "Modelo de Perfil e Seleção Aplicado ao Setor de BPO na Colômbia com Base em Técnicas de Análise Preditiva e Algoritmos Genéticos", realizado na Universidade Distrital Francisco José de Caldas em 2025.

Problema: A organização analisada apresentou uma taxa de rotatividade inicial de 62,82%, revelando altos custos associados a recrutamento, treinamento e curva de aprendizado, o que implicou limitações nos processos tradicionais de recrutamento.

Objetivo: Desenvolver e avaliar um modelo preditivo que identifique os candidatos com maior probabilidade de permanecer na empresa por pelo menos três meses e determine os atributos mais influentes para essa retenção.

Metodologia: Adotou-se uma abordagem de métodos mistos. As variáveis relevantes foram inicialmente identificadas por meio de consulta a profissionais de recrutamento; posteriormente, modelos de aprendizado de máquina foram treinados em 560 registros de contratação. Algoritmos como SVC, Random Forest e KNN foram aplicados. O desequilíbrio de classes foi tratado utilizando SMOTE, e o desempenho foi avaliado por meio de métricas como acurácia, sensibilidade, especificidade e pontuação F1. Resultados: O modelo proposto reduziu a rotatividade inicial por meio de simulação de 62,82% para 4,11%, representando uma redução relativa de 93,6%, juntamente com uma economia estimada de até COP 1 bilhão em um trimestre. O modelo SVC destacou-se com uma especificidade de 94,44%, minimizando contratações de alto risco.

Conclusões: Modelos preditivos em recrutamento melhoram significativamente a retenção e a otimização de recursos.

Originalidade: O estudo relaciona a precisão preditiva com um impacto financeiro mensurável no setor de BPO colombiano.

Limitações: Atributos sensíveis não foram coletados para evitar viés discriminatório.

Palavras-chave: Aprendizado de máquina, BPO, modelos preditivos, recursos humanos, seleção de pessoal.

1. INTRODUCTION

In today's business environment, recruitment is a critical function to ensure organizational success. The ability to predict employee performance prior to hiring can not only reduce costs but also significantly improve talent retention. However, inadequate selection can limit the efficient use of resources and distort organizational outcomes. This study presents a data-driven approach to integrate predictive analytics into the personnel selection process, thereby improving its efficiency and effectiveness. One of

the main challenges faced by companies in the BPO (Business Process Outsourcing) sector is early turnover, which occurs when hired employees resign within the first 90 days. This phenomenon generates significant economic impacts in three main areas: [1]–[6]

- Administrative costs derived from the search, selection, and hiring process.
- Penalties associated with non-compliance with key early turnover indicators.
- Costs related to the learning curve, since employees do not generate revenue equivalent to their salary during this initial period.

According to the data collected in the company under study, the average administrative cost for hiring and training an employee amounts to COP $1,236,361, while the costs associated with the learning curve represent COP $1,960,550. This results in a total cost of COP $3,196,550 per employee who does not complete three months of employment. Of the 560 employees hired in a typical month, 64.82% (363 employees) resign before reaching three months, generating an estimated economic loss of COP $1,160,347,650.

As a solution to this problem, this study proposes the development of a predictive model designed to identify applicants with the highest probability of completing at least three months of tenure. Additionally, the most relevant factors for future hiring are analyzed, integrating a quantitative approach that complements existing empirical knowledge [7].

This document is structured as follows:

- Theoretical framework, which compiles the conceptual foundations.
- Methodology, which describes the data collection and analysis process, as well as the results obtained.
- Analysis and discussion, where the findings are interpreted, key attributes are highlighted, and the performance of the applied algorithms is evaluated.
- Conclusions, which synthesize the main implications of the study and propose future research directions.

It is important to note that this research excluded attributes that could lead to discrimination, such as gender or race. However, one of the main limitations was data availability, since the success of the model depends on the quantity and quality of the information collected regarding applicant characteristics [8].

e-ISSN 2357-6014 / Vol. 22, no. 2 / may-august 2026 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

1.1 Literature Review or Research Background

a. Business Process Outsourcing

Business Process Outsourcing (BPO) emerged during the 1980s, when companies identified the need to outsource functions not directly related to their core business, commonly referred to as "non-core" activities. This strategy aimed to improve efficiency and reduce operational costs, allowing organizations to focus on their primary activities [9], [10].

As this approach enabled organizations to delegate secondary tasks to specialized providers, combined with the effects of globalization, both the supply and demand for outsourcing services have increased significantly in recent years, particularly in regions where operational costs are lower [11], [12].

The BPO industry is generally divided into two main categories: Front Office processes, which include customer service, sales, and technical support, and Back Office processes, such as case management and information analysis [13].

b. BPO in Colombia

The BPO sector in Colombia has emerged as a significant driver of economic growth and employment generation. According to a study conducted by the Bogotá Chamber of Commerce, Colombia has positioned itself as one of the leading destinations for outsourcing services in Latin America, due to its highly skilled workforce, competitive costs, and political and economic stability. By the end of 2023, the sector comprised approximately 600 companies, representing 2.8% of the national GDP and employing around 700,000 people. Of these employees, 80% are under 30 years old, 12% hold a professional degree, and only 8% have postgraduate qualifications.

Despite these achievements, the BPO sector in Colombia faces several challenges, including the need to improve service quality and specialization, as well as to address concerns related to data security and job stability. Currently, there is a strong emphasis on continuous training in both technical and soft skills, in addition to the implementation of effective cybersecurity measures to protect sensitive customer information [14]–[17].

The Colombian BPO Association (BPro), established in 2001, is a non-profit organization that represents companies within the BPO sector and its value chain. It includes more than 70 affiliated companies and contributes to the development of strategies aimed at enhancing the growth and competitiveness of the sector at both national and international levels, as well as promoting best practices [18].

a. Recruitment and Selection Processes

Given that, in today's business environment, personnel selection is a critical function to ensure organizational success, all companies have a human resources department, within which a selection area is typically established. These processes are decisive for proper organizational functioning, as they promote productivity through informed decision-making. This is based on the premise that the ability to predict employee performance prior to hiring can not only reduce costs but also significantly improve talent retention. The primary function of the selection area is to identify and select the most suitable candidates to fill available vacancies. Such vacancies arise from new positions, retirements, or employee replacements. These processes are mainly associated with HR departments and include psychological, technical (specific skills), and assessment center evaluations. Inadequate selection can limit the efficient use of resources and bias organizational outcomes [1], [3], [5], [6], [19], [20].

In this regard, the article "La selección de personal basada en competencias y su relación con la eficacia organizacional" defines, in general terms, the steps required to conduct an effective selection process, which are described below [21]:

- Writing the announcement: This step defines the guidelines of the selection process. A well-structured announcement can determine the quality and quantity of applicants for the vacancy.
- Resume analysis: The submitted resumes are reviewed to identify profiles aligned with the vacancy. At this stage, aspects such as writing quality are considered, and candidates to be invited for interviews and tests are selected.
- The interview: This involves a structured dialogue aimed at observing applicants' expressions, gestures, posture, and behavior. It is one of the most important factors in the final decision.
- Competency-based interview: Applicants are encouraged to respond based on factual experiences rather than opinions, ensuring objectivity. This requires a strong interaction between interviewer and interviewee. The interviewer must possess analytical skills and agility to identify behavioral patterns from the applicant's narrative, based on the premise that individuals tend to repeat past behaviors. Four key components are evaluated: technical competencies, functional experience, professional competencies, and social skills. According to the purpose, specific questions, tests, and evaluations are defined. This type of interview is generally more extensive

e-ISSN 2357-6014 / Vol. 22, no. 2 / may-august 2026 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

than a conventional one and enables a more comprehensive evaluation of candidates [21].

- Interview record: At this stage, the entire evaluation process must be documented, as it is essential for decision-making. Records should include information such as experience, knowledge, academic level, current remuneration, presentation, communication skills, and competencies.

Once all stages of the selection process have been completed, the final decision is made by the direct manager. Although the human resources department conducts the process and provides recommendations, it is not responsible for the final decision; however, it must provide sufficient information to support it.

It is important to note that job applicants are increasingly attracted to digital selection processes, as they perceive them as innovative. However, the absence of human interaction may have the opposite effect, leading to negative perceptions. To mitigate this, hybrid or "augmented" selection models are often implemented [22], [23].

b. Machine Learning in Predictive Analytics

Artificial intelligence and computational intelligence processes are characterized by replicating human cognitive functions through machines, including capabilities ranging from voice and image analysis to automated learning processes, such as Machine Learning (ML). Unlike human cognition, these tools enable the rapid and objective analysis of large volumes of data, facilitating the identification of patterns and the extraction of useful information to support more efficient decision-making. ML integrates algorithms and computational intelligence techniques that allow systems to learn and improve performance based on experience [24]–[27].

Within computational intelligence techniques, Data Analytics processes include a subfield known as Predictive Analytics (PA), which extracts information from data to forecast future outcomes. Within this framework, Machine Learning methods automatically learn relationships between descriptive attributes and a target variable based on historical data, enabling predictions about future behavior. Machine Learning models in Predictive Analytics are generally classified into two categories according to their learning nature [28], [29]:

- Supervised learning: The model is trained using a target variable that guides the learning process and addresses the research objective.

- Unsupervised learning: There is no target variable, and the objective is to identify patterns or structures within the data, which may not necessarily produce directly applicable results.

These models require two types of variables:

- Independent variables: These correspond to the input data available at the beginning of the process and are used to generate predictions (e.g., age, education level).
- Dependent (behavioral) variable: This is the target variable to be predicted. In supervised learning, the objective is to identify relationships between this variable and the independent variables, enabling valuable insights for decision-making.

c. Predictive Analytics Algorithms

The following section describes several machine learning algorithms applicable to this study:

- Logistic Regression: Commonly used in medical and social sciences, it models the probability of an event as a function of other variables, establishing relationships between a categorical target variable and predictor variables [30], [31].
- Decision Trees: These models are easily interpretable and allow evaluation and adjustment of accuracy. They map input attributes to a target variable by recursively partitioning the data [32].
- Random Forest: An ensemble method that constructs multiple decision trees and combines their outputs to improve predictive performance [33].
- K-Nearest Neighbors (KNN): This algorithm classifies data points based on proximity to neighboring instances in the feature space, using distance metrics for inference [32], [34].
- Support Vector Classifier (SVC): A classifier that separates classes by maximizing the margin between them [35].
- Gradient Boosting: An ensemble method that combines multiple weak learners (typically decision trees) to improve accuracy [36].
- AdaBoost: An algorithm that integrates weak classifiers to enhance classification performance [37].

e-ISSN 2357-6014 / Vol. 22, no. 2 / may-august 2026 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

- Naïve Bayes: Based on Bayes' theorem, it assumes independence among predictors [38].
- Multi-Layer Perceptron (MLP): A neural network model capable of learning complex representations through multiple layers [39].
- Bagging Classifier: Generates multiple subsets of training data to improve model accuracy and reduce variance [40].
- Extra Trees: A variant of decision trees that randomly selects features and thresholds to enhance generalization [41].

d. Performance Measures

To evaluate the classifiers, the dataset is divided into two subsets: a training set, used to build the model, and a validation set, used to assess its performance. In the validation phase, predictions generated by the trained model are compared with actual outcomes.

The performance of each classifier is evaluated using confusion matrices, which enable the calculation of metrics such as true positive rate, false positive rate, and harmonic mean (F1-score). The confusion matrix provides a graphical representation of results, where columns represent actual classes and rows represent predicted classifications [42].

The performance measures that are broken down from the confounding matrix are described in the Equation 1, 2, 3 and 4, along with its objective.

Equation 1: Accuracy - Measures the percentage of correct positive predictions

$$Precisión = \frac{Verdaderos\ Positivos}{\text{Verdaderos Positivos} + \text{Verdaderos Negativos}} \quad (1)$$

Equation 2: Accuracy - Measures the percentage of individuals classified correctly.

$$Accuracy = \frac{V + N}{\text{N} + \text{FN} + \text{FP} + \text{V}} \quad (2)$$

e-ISSN 2357-6014 / Vol. 22, no. 2 / may-august 2026 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

Equation 3: Recall - Measures the percentage of positive individuals detected.

$$Recall = \frac{Verdaderos\ Positivos}{\text{Verdaderos Positivos} + \text{Falsos Negativos}} \quad (3)$$

Equation 4: Specificity - Measures the percentage of negative individuals detected

$$Especifidad = \frac{Verdaderos\ Negativos}{\text{Verdaderos Negativos} + \text{Falsos Positivos}} \quad (4)$$

This stage is critical for the development of this research, as, following validation testing, at least one algorithm was selected for implementation in each problem scenario. This selection process was incorporated into the proposed model, and in each case, the chosen algorithm contributed to addressing the research question defined at the outset.

2. MATERIALS AND METHODS

To address the inherent complexity of the personnel selection process in companies within the BPO sector in Colombia, this research adopted a methodology that combines qualitative and quantitative approaches. Initially, a qualitative approach was used to identify the key characteristics of successful employees, based on the knowledge and experience of the selection psychologists in the company under study. Subsequently, a quantitative analysis was conducted, applying computational intelligence tools and genetic algorithms to formalize these findings and develop a robust predictive model. This approach is based on the integration of qualitative and quantitative techniques, which together enable a more accurate assessment of candidates. Figure 1 presents the step-by-step methodological process [43].

The mixed methodology employed in this study enabled a more comprehensive understanding of the selection process, combining the rigor of quantitative analysis with the depth of qualitative insights. This approach is particularly relevant in a business context where hiring decisions have a significant impact on organizational culture and performance. The results indicate that the integration of predictive techniques into

personnel selection is not only feasible but also offers advantages in terms of process simplicity and improved predictive accuracy [44]–[47].

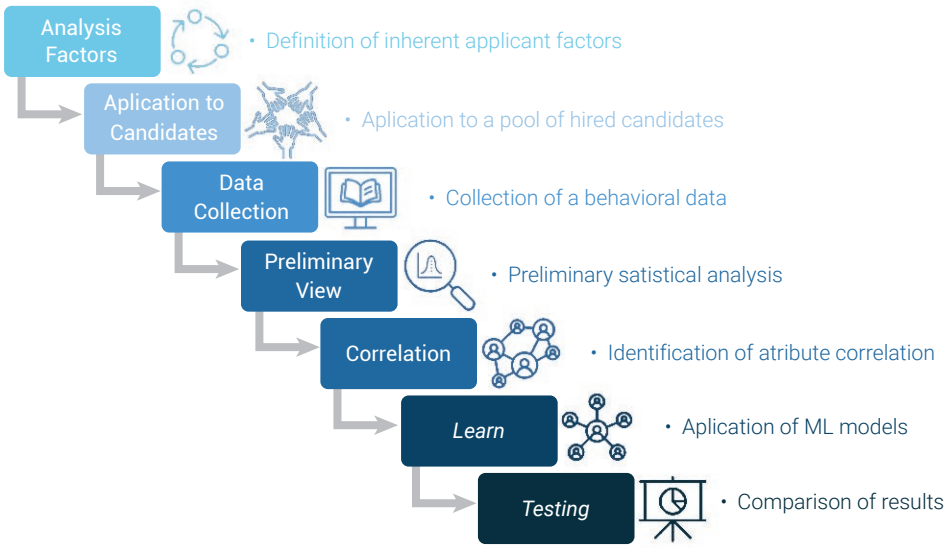

Figure 1. Methodological process.
Source: Own work

a. Definition of Applicants' Inherent Factors

Along with the selection area, those attributes were defined considering that they were relevant to the selection process, as well as that they were not discriminatory to avoid biases in the results and legal restrictions, as shown in the [48]Figure 2.

The following is a description of the attributes considered for each factor

- **Personal factor:** Age, number of children, marital status.
- **Family factor:** Number of people you live with, number of dependents, type of housing, place of residence, spouse, spouse works, lives with father, lives with mother, lives with siblings, lives with children, lives with others.
- **Academic factor:** No. Languages, Level of education, Are currently studying, branch of study, study center, speak English, French, Portuguese, other.
- **Professional Factor:** Months of experience, months of experience in the BPO sector, Number of jobs in the last 5 years, months without employment, duration of the last job, salary aspiration, previous salary, availability of income, previous job is BPO.

e-ISSN 2357-6014 / Vol. 22, no. 2 / may-august 2026 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

- **Business Factor:** Salary contracted, type of service for which it was contracted, contracted working hours.
- **Objective Factor:** The employee did not resign until at least three months after their contract start date. In the collection form, they were labeled a 'Good Employee,' referring solely to the fact that they met the minimum required period of stay, regardless of performance.

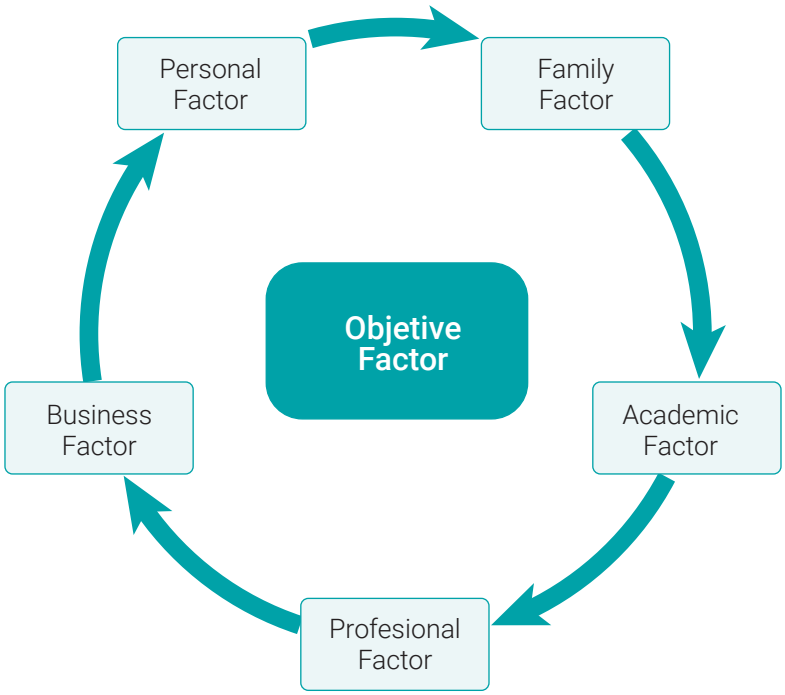

Figure 2. Based on psychologists selection of study company.
Source: Own work.

The selection of variables for the exploratory analysis was based on their relevance to the study objective. These variables were chosen because the head of the selection area—responsible for data collection together with their team—indicated, based on professional experience, that these factors could significantly influence job performance. Age, for instance, is considered relevant due to its association with work experience and professional maturity. The number of dependents may reflect additional family responsibilities that could affect an individual's availability and performance. Marital status can also influence work–life balance, potentially impacting productivity and engagement. Similarly, the number of previous jobs and the duration of unemployment provide insights into job stability and prior experience, both of which are key in evaluating performance and adaptability. In summary, the selected variables—such

as age, marital status, number of dependents, and work experience—were chosen for their practical relevance and their potential impact on performance and adaptation, as assessed by the selection area.

b. Application to a Pool of Hired Candidates and Collection of Behavioral Data

The data collection period was established as one month, selecting a representative period without significant seasonal fluctuations, such as December (holiday season) or April (Easter period). Psychologists responsible for each hiring process administered a survey to candidates at the time of onboarding. Subsequently, information related to the previously defined "company" factor—such as salary, type of service, and working hours—was incorporated. Three months after hiring, the Human Resources department completed the records by indicating which candidates met the minimum retention threshold of three months.

To ensure standardized response categories and minimize missing data, a Google Form with multiple-choice options was used. All individual datasets were consolidated into a single database containing records for 560 employees hired during the study period.

c. Preliminary Statistical Analysis

Since many data analysis algorithms require numerical input, converting categorical variables into numerical format is essential. This process, known as categorical variable encoding, consists of assigning numerical values to the different categories within each variable. This facilitates correlation analysis, the application of predictive models, and the overall interpretation of the data. The procedures illustrated in Figures 3 and 4 perform this transformation by identifying categorical variables in the dataset and assigning numerical values to their corresponding categories. This approach ensures the effective integration of categorical variables into subsequent analyses, enabling a more comprehensive and accurate exploration of the data.

e-ISSN 2357-6014 / Vol. 22, no. 2 / may-august 2026 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

```
# Show categories per attribute
for i in data:
    if data[i].dtypes == "object":
        print("%s has: " % i, np.size(data[i].unique()), " categories")
```

```
marital_status has:  4  categories
housing_type has:  3  categories
district has:  23  categories
spouse has:  2  categories
spouse_works has:  3  categories
lives_with_father has:  2  categories
lives_with_mother has:  2  categories
lives_with_siblings has:  2  categories
lives_with_children has:  2  categories
lives_with_others has:  2  categories
education_level has:  6  categories
currently_studying has:  2  categories
major has:  26  categories
study_center has:  3  categories
english has:  2  categories
french has:  2  categories
portuguese has:  2  categories
other_language has:  1  categories
previous_position has:  3  categories
```

Figure 3. Identification of categories to categorical variables.
Source: Own work.

```
b = "good_employee"
i = 0
datat = pd.DataFrame(np.random.randint(low=0, high=1, size=(len(data[b].unique()), 2)))

for c in sorted(data[b].unique()):
    datat.loc[i, b] = c
    datat.loc[i, "Number"] = i
    i = i + 1

display(datat)

for idx, row in data.iterrows():
    for j in range(len(datat[b])):
        if data.loc[idx, b] == datat.loc[j, b]:
            data.loc[idx, b] = datat.loc[j, "Number"]
            break
```

	0	1	good_employee	Number
0	0	0	No	0.0
1	0	0	Yes	1.0

Figure 4. Example of assigning numbers to categorical variables.
Source: Own work.

After converting the categorical variables into numerical ones, an exploratory approach was conducted to better understand the relationship between different characteristics of the data. This initial approach involves a visual exploration of the data, which provides an intuitive and insightful perspective on the trends and patterns present in the dataset. In this context, a specific visualization was used to examine the relationship between the target category "Good Employee" and several other key characteristics of the dataset. These characteristics include age, number of dependents,

marital status, number of previous jobs, and number of months without employment. By visually analyzing these relationships, potential significant associations and trends can be identified that are crucial to understanding employees' job performance and circumstances, for example in the Figure 5 the behavior of the target variable with respect to age indicates a higher tendency to resign between the ages of 20 to 28 years old, as well as the tendency of retention between 32 and 40 years old.

```
good_employee_data = data[data.good_employee == 1].age
bad_employee_data = data[data.good_employee == 0].age

plt.figure(figsize=(8, 5))

bins = np.linspace(min(data.age), max(data.age), 10)
plt.hist( good_employee_data, bins=bins, color='blue', alpha=0.6,label='Good Employee', edgecolor='black',
    weights=np.ones(len(good_employee_data)) / len(good_employee_data)
plt.hist(bad_employee_data, bins=bins, color='orange', alpha=0.6, label='Bad Employee', edgecolor='black',
         weights=np.ones(len(bad_employee_data)) / len(bad_employee_data))
plt.title("Age Distribution by Job Performance (Normalized)", fontsize=12)
plt.xlabel("Age", fontsize=10)
plt.ylabel("Proportion", fontsize=10)
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()
```

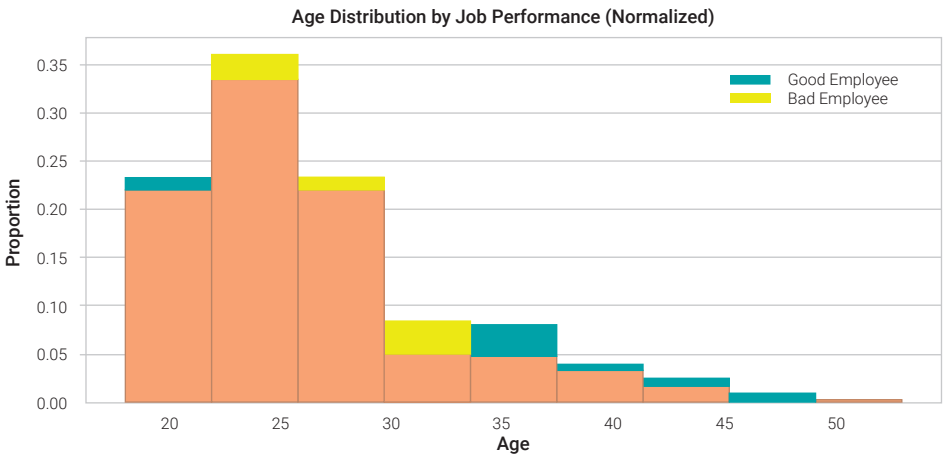

Figure 5. Distribution of the category "Good Employee" according to age.
Source: Own work.

According to the data collected, it can be observed that there is a greater number of resignations when employees do not have children or when they have one compared to when they have two or more, as evidenced in Figure 6.

```
good_employee_data = data[data.good_employee == 1].dependents
bad_employee_data = data[data.good_employee == 0].dependents

plt.figure(figsize=(8, 5))

bins = np.arange(min(data.dependents), max(data.dependents) + 1, 1)  # Adjust if values are integers
plt.hist( good_employee_data, bins=bins, color='blue', alpha=0.6, label='Good Employee', edgecolor='black',
    weights=np.ones(len(good_employee_data)) / len(good_employee_data))
plt.hist( bad_employee_data, bins=bins, color='orange', alpha=0.6, label='Bad Employee', edgecolor='black',
    weights=np.ones(len(bad_employee_data)) / len(bad_employee_data))
plt.title("Normalized Distribution of Dependents vs. Employee Performance", fontsize=12)
plt.xlabel("Number of Dependents", fontsize=10)
plt.ylabel("Proportion", fontsize=10)
plt.legend()
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()
```

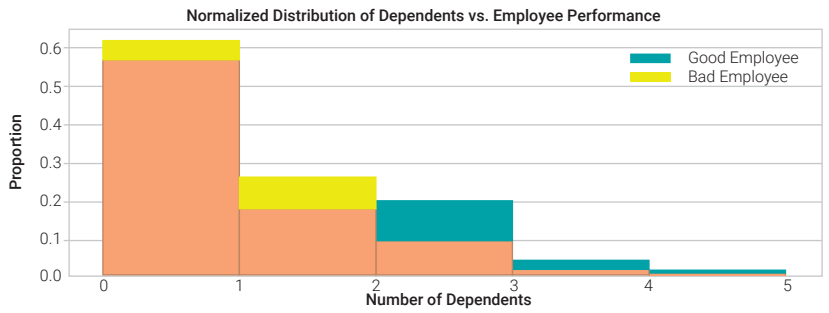

Figure 6. Distribution of the "Good Employee" category according to dependents.
Source: Own work.

When an applicant has been unemployed for 6 months or more, the tendency to quit is lower compared to when they come from another job directly, as can be seen in the Figure 7.

```
good_employee_data = data[data.good_employee == 1].months_unemployed
bad_employee_data = data[data.good_employee == 0].months_unemployed

plt.figure(figsize=(8, 5))
bins = np.arange(min(data.months_unemployed), max(data.months_unemployed) + 1, 1)
plt.hist( good_employee_data, bins=bins, color='blue', alpha=0.6, label='Good Employee', edgecolor='black',
    weights=np.ones(len(good_employee_data)) / len(good_employee_data))
plt.hist( bad_employee_data, bins=bins, color='orange', alpha=0.6, label='Bad Employee', edgecolor='black',
    weights=np.ones(len(bad_employee_data)) / len(bad_employee_data))
plt.title("Normalized Distribution of Months Unemployed by Employee Performance", fontsize=12)
plt.xlabel("Months Unemployed", fontsize=10)
plt.ylabel("Proportion", fontsize=10)
plt.legend(loc='upper right')
plt.grid(axis='y', linestyle='--', alpha=0.6)
plt.show()
```

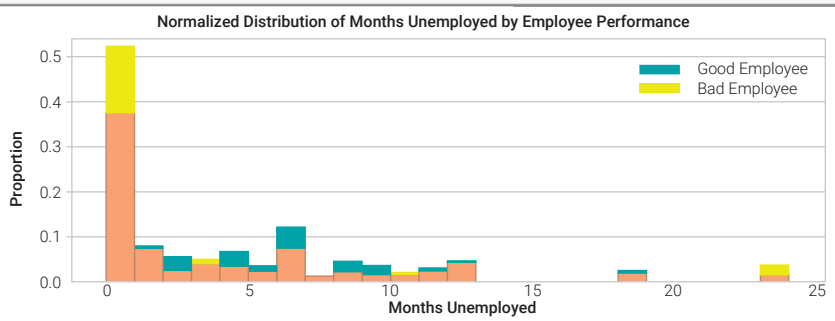

Figure 7. Distribution of the "Good Employee" category according to the number of months without an employee.
Source: Own work.

Based on the graphs in this chapter, we can observe that the relationship between the target category "Good Employee" and factors such as age, the number of dependents and months without employment shows clear patterns. For example, the greatest tendency to quit occurs in young employees (20 to 28 years old), while retention rates increase between 32 and 40 years old. In addition, employees with more family responsibilities (2 or more dependents) tend to keep their jobs longer, and those with long periods of unemployment (6 months or more) are less likely to quit.

d. Identifying Attribute Correlation

To represent the dependence between variables, a correlation matrix was used, this statistical tool measures the intensity and direction of the linear relationships between variables, using values ranging from -1 to 1, a value of 1 indicates a perfect positive correlation, -1 a perfect negative correlation, and 0 denotes no correlation. In this matrix, the strongest correlations are visualized on red scales, indicating significant positive relationships, while blue ones represent negative relationships. It is worth noting that the main diagonal line that is completely in red reflects the perfect correlation of each variable with itself and therefore does not provide relevant information for the analysis. Additionally, a white line corresponding to the variable 'otro_idioma' indicates a lack of correlation, suggesting that this attribute is independent. In contrast, the variable 'English' shows a strong correlation with both 'salary aspiration' and 'hired salary', as illustrated in Figure 8 [15].

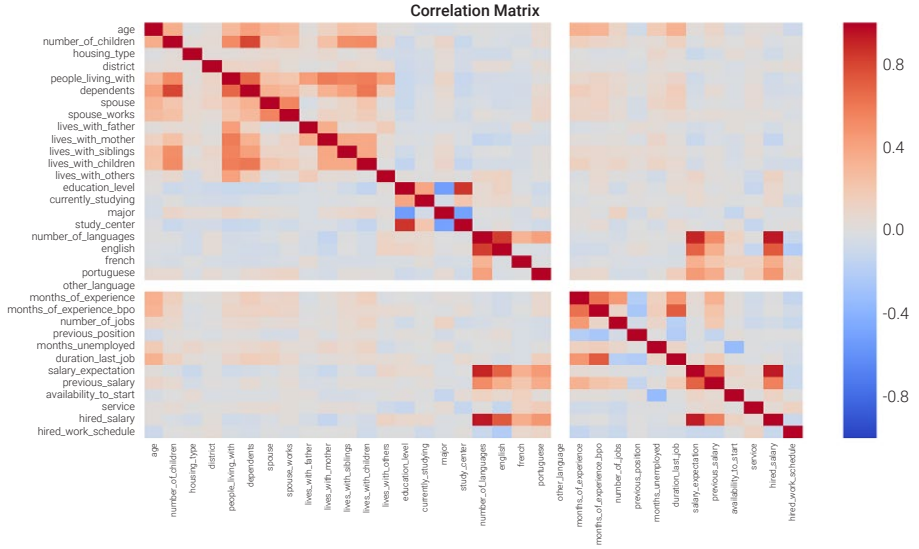

Figure 8. Correlation matrix.
Source: Own work.

e-ISSN 2357-6014 / Vol. 22, no. 2 / may-august 2026 / Bogotá D.C., Colombia
Universidad Cooperativa de Colombia

Since there is no significant correlation with the objective variable "Good employee" visually, we proceeded to quantify its relationship with each variable, where it was evidenced that the highest correlation is 0.16; Due to its proximity to 0, it is inferred that the correlations are not significant, therefore there is no direct relationship with any of the variables collected, as presented in the Figure 9.

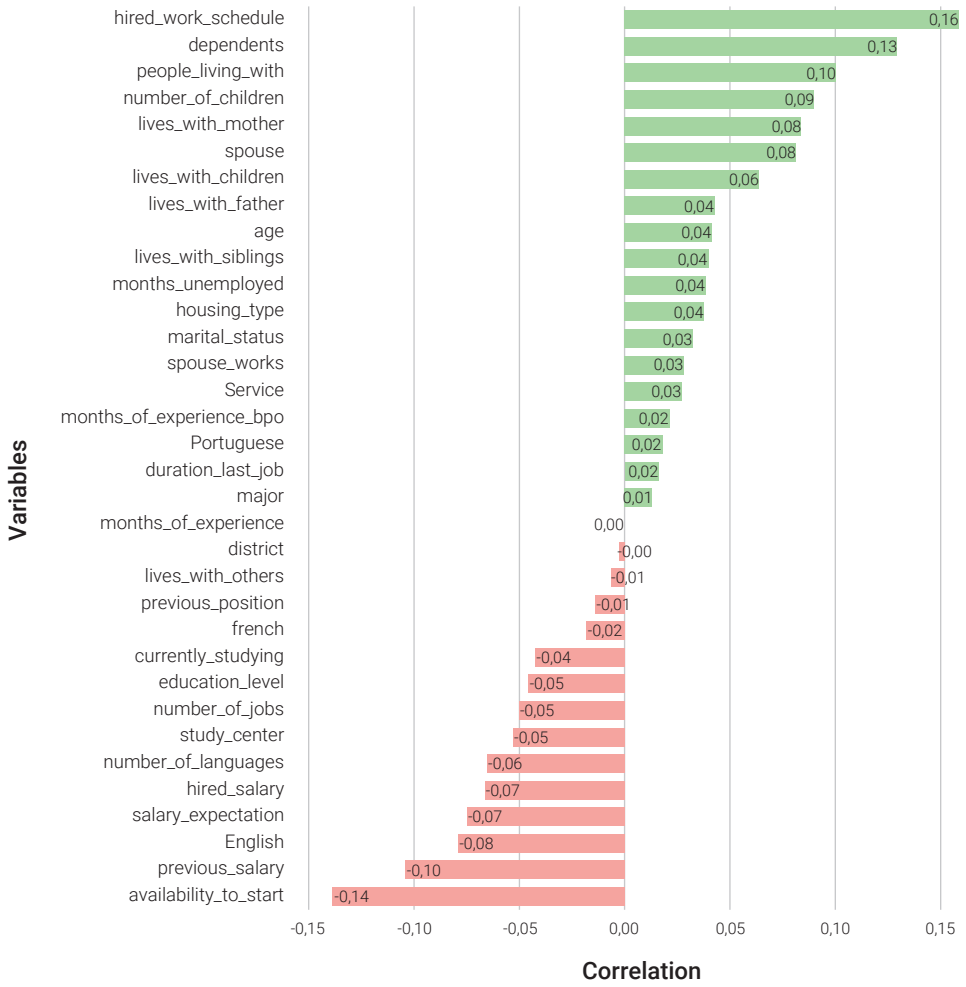

Figure 9. Correlation graph of variables vs. variable "Good Employee".
Source: Own work.

To evaluate the relationship between the target variable "Good employee" and the other variables, a **chi-square** test was performed. This analysis allowed us to explore possible significant associations, considering as a reference a p-value < 0.05 to identify relevant relationships. In the Figure 10 The results obtained for each variable are presented.

```
names2 = ['age', 'marital_status', 'number_of_children', 'housing_type',
    'district', 'people_living_with', 'dependents', 'spouse',
    'spouse_works', 'lives_with_father', 'lives_with_mother', 'lives_with_siblings',
    'lives_with_children', 'lives_with_others', 'education_level', 'currently_studying',
    'major', 'study_center', 'number_of_languages', 'english', 'french', 'portuguese',
    'other_language', 'months_of_experience', 'months_of_bpo_experience',
    'number_of_jobs', 'previous_position', 'months_unemployed',
    'duration_of_last_job', 'salary_expectation', 'previous_salary',
    'availability_to_start', 'service', 'contracted_salary', 'contracted_schedule']

FunctionChisq( inpData=data, TargetVariable='good_employee', CategoricalVariablesList=names2)
```

```
number_of_children is correlated with good_employee | P-Value: 0.02637073572586591
housing_type is NOT correlated with good_employee | P-Value: 0.4889710767176483
district is NOT correlated with good_employee | P-Value: 0.8885266367324636
people_living_with is NOT correlated with good_employee | P-Value: 0.45362782293575465
dependents is correlated with good_employee | P-Value: 0.0014721350445338498
spouse is NOT correlated with good_employee | P-Value: 0.066101283215613
spouse_works is NOT correlated with good_employee | P-Value: 0.13640326890853235
lives_with_father is NOT correlated with good_employee | P-Value: 0.3550856234989572
lives_with_mother is NOT correlated with good_employee | P-Value: 0.056209899840388565
lives_with_siblings is NOT correlated with good_employee | P-Value: 0.39688512684599386
lives_with_children is NOT correlated with good_employee | P-Value: 0.1597933528360904
lives_with_others is NOT correlated with good_employee | P-Value: 0.9968296612544694
education_level is NOT correlated with good_employee | P-Value: 0.597049808360804
currently_studying is NOT correlated with good_employee | P-Value: 0.40065284431096704
major is NOT correlated with good_employee | P-Value: 0.12152941998477963
study_center is NOT correlated with good_employee | P-Value: 0.3677785321914389
number_of_languages is NOT correlated with good_employee | P-Value: 0.17353263992496976
english is NOT correlated with good_employee | P-Value: 0.09860733803985978
french is NOT correlated with good_employee | P-Value: 0.9222688883398322
```

Figure 10. Chi Square test results graph.
Source: Own work.

Thus, the variables "Number of children", "Dependents", "Previous salary", "Income availability", "Hired salary", and "Working schedule" were identified as presenting statistically significant results in the applied test, suggesting a relevant relationship with the "Good Employee" category. This finding highlights the importance of considering both personal and work-related factors in the analysis of employee performance.

For example, "Number of children" and "Dependents" may be associated with a greater sense of responsibility and stability, factors that could positively influence employee retention and engagement. On the other hand, variables such as "Previous salary" and "Hired salary" may reflect salary expectations aligned with the position, which could contribute to reduced turnover.

Additionally, "Income availability" and "Working schedule" may be related to the employee's ability to adapt to the working conditions offered, thereby influencing their decision to remain in the organization. These results not only provide valuable insights for improving the selection process but also highlight the need for further research on how these factors interact to influence job performance.

Application of Machine Learning Models

To properly apply Machine Learning models, it is necessary to define the output variable, referred to as "Good Employee", in relation to the predictor variables. The dataset is then divided into two subsets: a training set (80%) and a test set (20%). This is a standard practice in machine learning, allowing a balance between model training and evaluation. This separation is essential to objectively assess model performance. The procedure performed is illustrated in Figure 11 [49].

```
#Spliting Data into Test and Train
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, MinMaxScaler
PredictorScaler=MinMaxScaler()


x_train,x_test,y_train,y_test=train_test_split(df,y,test_size=0.2,random_state=0)

# Sanity check for the sampled data
print(x_train.shape)
print(y_train.shape)
print(x_test.shape)
print(y_test.shape)
```

```
(580, 35)
(580,)
(146, 35)
(146,)
```

Figure 11. Data separation code in training and testing.
Source: Own work.

To seek to balance the distribution between minority and majority classes, which can improve the generalizability and prediction of the model by reducing class bias, the SMOTE (Synthetic Minority Over-sampling Technique) technique is applied to the training set to generate synthetic samples of the minority class ("Good Employee") by creating close artificial data in the characteristic space as presented in the [50] Figure 12.

```
#Seperating Output variable
y=data['buen_empleado']
df=data.drop(['buen_empleado'],axis=1)
```

```
#Over-Sampling Minority Group
from imblearn.over_sampling import SMOTE

sm = SMOTE(random_state=0, k_neighbors=5, ratio='minority')
df, y = sm.fit_sample(df, y)
```

Figure 12. SMOTE technique application code.
Source: Own work.

With the pre-processed and balanced data, the various machine learning algorithms of classification, regression and clustering are implemented and evaluated. The models were selected based on the literature review and their relevance to the type of data and objectives of the analysis, evaluating each one based on its predictive accuracy and interpretability. Below is a summary of the algorithms used: [51]

- **RandomForestClassifier:** Decision tree-based model that improves accuracy with multiple trees. [33]
- **SVC:** Classifier that uses the maximum margin between classes for separation. [35]
- **GradientBoosting:** Assembly model that combines several weak trees to improve accuracy. [36]
- **LogisticRegression:** Model that estimates probabilities and classifies based on logistic regression. [30]
- **Decision Trees:** A decision tree that classifies data by dividing based on characteristics. [32]
- **Adaboost:** An assembly method that adjusts the weights of errors in successive iterations. [37]
- **K-Nearest Neighbor (KNN):** Classifier based on the proximity of the points in the feature space. [34]
- **Naive Bayes:** Classifier based on Bayesian theory and independence between characteristics. [52]
- **Multi-layer Perceptron (MLP):** Neural network that uses multiple layers to learn complex representations. [39]
- **BaggingClassifier:** Assembly method that uses multiple base classifiers to improve accuracy. [40]
- **Extra Trees:** A variant of decision trees that introduces randomness in the division of nodes. [41]

3. RESULTS

The model was trained using a portion of the dataset, while the remaining portion was used to validate its performance. Metrics such as accuracy, sensitivity, and specificity were employed to assess the effectiveness of the model in predicting candidate success. The results were analyzed using a matrix that enabled the identification of patterns and the evaluation of model performance under different scenarios. This

analysis provided valuable information for refining the model, improving its accuracy and long-term applicability.

The matrix presented in Table 1 [49] provides a comparative evaluation of the performance of the selected algorithms using several relevant metrics. "Training Accuracy" reflects the model's ability to fit the training dataset, while "Testing Accuracy" measures its performance on previously unseen data. The "Sensitivity" and "Specificity" metrics offer complementary insights into the model's ability to correctly identify both positive cases (employees who remain) and negative cases (employees who leave). The "F1 Score" balances precision and sensitivity, and the "AUC Score" summarizes the overall effectiveness of the model based on the ROC curve.

The best results are highlighted in green, indicating the algorithms with the strongest performance for each metric, while the values in red indicate the lowest performance. Additionally, the categories "TP", "FP", "FN", and "TN" correspond to true positives, false positives, false negatives, and true negatives, respectively, enabling a detailed analysis of the errors and correct classifications for each model.

Board 1. Comparison of algorithm training results.

Algorithm Name	Testing Accuracy	Sensitivity	Specifivity	F1 Score	AUC Score	HCMC	FP	FN	TN
RandomForestClassifier	0,7260	0,7313	0,7215	0,7403	0,7251	49	22	18	57
SVC	0,8013	0,7216	0,9444	0,7727	0,8064	70	3	27	51
GradientBoosting	0,8079	0,8143	0,8025	0,8176	0,8071	57	16	13	65
LogisticRegression	0,6358	0,6364	0,6353	0,6626	0,6338	42	31	24	54
Decision Trees	0,7815	0,8030	0,7647	0,7975	0,7797	53	20	13	65
Adaboost	0,8013	0,7945	0,8077	0,8077	0,8011	58	15	15	63
K-Nearest Neighbor (KNN)	0,8146	0,8462	0,7907	0,8293	0,8126	55	18	10	68
Naive Bayes	0,5629	0,7692	0,5435	0,6944	0,5493	10	63	3	75
Multi-layer Perceptron	0,4967	0,4866	0,6667	0,0952	0,5119	71	2	74	4
BaggingClassifier	0,8079	0,8056	0,8101	0,8153	0,8075	58	15	14	64
Extra Trees	0,7748	0,7671	0,7805	0,7805	0,7746	56	17	17	61

Source: Own work

4. DISCUSSION AND CONCLUSIONS

4.1 Analysis and Interpretation

The results of this research demonstrate that the developed predictive model achieves high accuracy in identifying candidates with a high probability of retention. Among the most significant attributes are "Number of children", "Dependents", "Previous salary", "Income availability", "Hired salary", and "Working schedule", which validate the initial hypothesis regarding the critical factors in the selection process. In addition, the qualitative analysis complemented the quantitative approach by capturing subjective aspects of the selection process that would not have been identified otherwise.

These findings have important implications for human resource management. The ability to predict candidate retention prior to hiring improves the efficiency of the selection process and contributes to better alignment between recruitment decisions and the strategic objectives of the organization.

In the recruitment context, the KNN algorithm demonstrates strong performance in key metrics such as Sensitivity (84.62%) and F1 Score (82.93%), indicating its effectiveness in correctly identifying both employees who are likely to remain and those who may leave. This makes it suitable for scenarios where overall predictive balance is required and both false positives and false negatives must be minimized.

However, when addressing the practical objective of minimizing costs associated with early employee turnover, the Support Vector Classifier (SVC) emerges as a more appropriate alternative. This algorithm achieved the highest Specificity (94.44%) and only three false positives, indicating a strong capacity to correctly identify employees who are unlikely to leave. This characteristic is particularly relevant in scenarios where the cost of hiring employees who subsequently resign is high, as it reduces unnecessary allocation of resources to candidates with a high risk of attrition.

From a practical perspective, especially when the primary objective is cost reduction, SVC stands out due to its effectiveness in minimizing the risk of hiring employees who will leave the organization prematurely.

4.2 Conclusions

The predictive model developed in this study enabled a substantial reduction in early turnover, decreasing from an initial 62.82% to 4.11% in the final simulation. In addition to improving retention, the model contributes to more efficient use of financial and operational resources, corresponding to estimated savings exceeding COP $1,000 million in quarterly cash flow. This estimation is based on the cost of hiring

(administrative and learning curve costs) multiplied by the number of employees who resign within the first three months.

The Support Vector Classifier (SVC) algorithm proved to be the most suitable option for minimizing turnover-related costs, due to its high specificity (94.44%) and low number of false positives. This result is particularly relevant in business environments where employee replacement costs are high, reinforcing the value of predictive techniques in high-impact economic contexts.

The analysis identified critical variables such as working schedule ($r = 0.16$, $p = 0.0007$), dependents ($r = 0.13$, $p = 0.0015$), and number of children ($r = 0.09$, $p = 0.0264$), which showed positive and statistically significant correlations with the target variable, indicating that work-related and family-related factors directly influence employee retention. Additionally, variables such as income availability ($r = -0.14$, $p = 0.0159$) and hired salary ($r = -0.07$, $p = 0.0109$) exhibited negative correlations, suggesting that lower values in these variables may be associated with longer tenure, potentially due to economic motivations. These findings highlight the relevance of incorporating such variables into predictive models for early turnover in the Colombian BPO sector.

This study establishes a foundation for the continuous application of predictive models in human resource management, which can be adapted to other industries and improved through the incorporation of new data. This approach enables organizations to respond to evolving labor dynamics, promoting more efficient and competitive selection processes, while also providing a replicable and scalable framework for addressing common challenges in the BPO sector.

REFERENCES

[1] S. Malisetty, R. V. Archana, and K. V. Kumari, "Predictive analytics in HR management," *Indian Journal of Public Health Research & Development*, vol. 8, no. 3, pp. 115–120, 2017, doi: https://doi.org/10.5958/0976-5506.2017.00238.3.

[2] K. Gopinath and A. Balamurugan, "Human resource analytics: Leveraging machine learning for employee attrition prediction," in *Proc. Int. Working Conf. Transfer and Diffusion of IT*, Nagpur, India, Dec. 2023, pp. 137–158, doi: https://doi.org/10.1007/978-3-031-46594-9_9.

[3] R. Priyanka, K. Ravindran, B. Sankaranarayanan, and M. S. Ali, "Employee attrition prediction using analytics approaches," *Decision Analytics Journal*, vol. 6, p. 100192, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100192.

[4] B. Jabir, N. Falih, and K. Rahmani, "HR analytics: A roadmap for decision making—Case study," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 2, pp. 979–990, 2019, doi: https://doi.org/10.11591/ijeecs.v15.i2.pp979-990.

[5] A. Q. Mohammed, "HR analytics: A modern tool in HR for predictive decision making," *Journal of Management*, vol. 6, no. 3, pp. 51–63, 2019, doi: https://doi.org/10.34218/JOM.6.3.2019.007.

[6] G. K. Kaya, S. Ustebay, J. Nixon, C. Pilbeam, and M. Sujan, "Exploring the impact of safety culture on incident reporting: Lessons learned from machine learning analysis of NHS England staff survey and incident data," *Safety Science*, vol. 166, p. 106260, 2023, doi: https://doi.org/10.1016/j.ssci.2023.106260.

[7] D. Pessach, G. Singer, D. Avrahami, H. C. Ben-Gal, E. Shmueli, and I. Ben-Gal, "Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming," *Decision Support Systems*, vol. 134, p. 113290, 2020, doi: https://doi.org/10.1016/j.dss.2020.113290.

[8] J. Simons, S. A. Bhatti, and A. Weller, "Machine learning and the meaning of equal treatment," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, Virtual Event, USA, 2021, pp. 1–10, doi: https://doi.org/10.1145/3461702.3462533.

[9] J. Thakkar and S. G. Deshmukh, "A review on evolution of business process outsourcing industry," *International Journal of Advance Research and Innovative Ideas in Education*, vol. 4, no. 4, pp. 2766–2770, 2018. [Online]. Available: https://iaeme.com/MasterAdmin/Journal_uploads/JOM/VOLUME_6_ISSUE_3/JOM_06_03_007.pdf

[10] L. Aronsson and M. Abrahamsson, "BPO and outsourced software development: A systematic literature review," *Journal of Software: Evolution and Process*, vol. 29, no. 7, p. e1872, 2017, doi: https://doi.org/10.1002/smr.1872.

[11] M. C. Lacity and R. Hirschheim, "The information systems outsourcing bandwagon," *MIT Sloan Management Review*, vol. 35, no. 1, pp. 73–86, 1993. [Online]. Available: https://sloanreview.mit.edu/article/the-information-systems-outsourcing-bandwagon/

[12] R. L. Click and T. N. Duening, *Business process outsourcing: The competitive advantage*. New Jersey, NJ, USA: John Wiley & Sons, 2024. [Online]. Available: https://vdoc.pub/documents/business-process-outsourcing-the-competitive-advantage-7smgp1oe1h70

[13] N. A. Higuera and B. T. Flórez, "Propuesta de mejoramiento de tecnologías de la información para lograr impacto en el desarrollo de proyectos para el sector BPO comercial en Colombia,"

B.S. thesis, Universidad Nacional Abierta y a Distancia, Bogotá, Colombia, 2020. [Online]. Available: https://repository.unad.edu.co/bitstream/handle/10596/34262/nahiguera.pdf

[14] Cámara de Comercio de Bogotá, "Oportunidades y desafíos del sector BPO en Bogotá–Región," 2019. [Online]. Available: https://www.ccb.org.co/Documents/10212/27131/Sector_BPO_2019.pdf

[15] J. Sánchez, H. A. Santana, D. A. Restrepo, and D. A. Malagón, "Análisis descriptivo del impacto de la inteligencia artificial en el sector de contact center en Bogotá," Universidad EAN, Bogotá, Colombia, 2024. [Online]. Available: https://repository.universidadean.edu.co/server/api/core/bitstreams/6664d474-c976-476c-a60e-524f828de11b/content

[16] J. Hernández, O. Torres, and M. Molina, "The impact of Industry 4.0 on business process outsourcing: A systematic literature review," *International Journal of Innovation, Creativity and Change*, vol. 14, no. 5, pp. 1058–1075, 2020. [Online]. Available: https://repository.udistrital.edu.co/server/api/core/bitstreams/9a85b8b7-867f-4b80-8546-84fffea10354/content

[17] L. Rodríguez and E. García, "Implementation of a cybersecurity management system for BPO companies," *International Journal of Cyber-Security and Digital Forensics*, vol. 9, no. 1, pp. 51–63, 2020, doi: https://doi.org/10.17781/P002612.

[18] BPRO, "Asociación Colombiana de BPO," Feb. 25, 2025. [Online]. Available: https://www.bpro.org/

[19] J. A. Romero and B. R. I. S. A. Sánchez, "Digitalización de la gestión del personal en recursos humanos," *Revista Ciencia Administrativa*, 2018. [Online]. Available: https://openurl.ebsco.com/EPDB%3Agcd%3A2%3A33286658/detailv2?sid=ebsco%3Aplink%3Ascholar&id=ebsco%3Agcd%3A138598854

[20] Z. Kaya and G. B. Yildiz, "Prediction of employee turnover in organizations using machine learning algorithms: A decision-making perspective," in *Proc. Int. Symp. Intelligent Manufacturing and Service Systems*, Singapore, 2023, pp. 1–10, doi: https://doi.org/10.1007/978-981-99-3126-0_25.

[21] J. R. L. Gumucio, "La selección de personal basada en competencias y su relación con la eficacia organizacional," *Perspectivas*, vol. 26, pp. 129–152, 2010. [Online]. Available: http://perspectivas.ucb.edu.bo/index.php/a/article/view/186

[22] N. Folger, P. Brosi, J. Stumpf-Wollersheim, and I. M. Welpe, "Applicant reactions to digital selection methods: A signaling perspective on innovativeness and procedural justice," *Journal of Business and Psychology*, pp. 1–23, 2021, doi: https://doi.org/10.1007/s10869-021-09756-8.

[23] M. F. Gonzalez, W. Liu, L. Shirase, D. L. Tomczak, C. E. Lobbe, R. Justenhoven, and N. R. Martin, "Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes," *Computers in Human Behavior*, vol. 130, p. 107179, 2022, doi: https://doi.org/10.1016/j.chb.2022.107179.

[24] J. E. Carrillo Tubón and D. M. Cevallos Casasilla, "El rol de la inteligencia artificial en las estrategias de comunicación de marketing digital para empresas del sector de La Floresta," Universidad Politécnica Salesiana, Quito, Ecuador, 2024. [Online]. Available: https://dspace.ups.edu.ec/handle/123456789/28710

[25] J. C. Ponce Gallegos *et al.*, *Inteligencia artificial*. Iniciativa Latinoamericana de Libros de Texto Abiertos (LATIn), 2014. [Online]. Available: https://rephip.unr.edu.ar/items/8beb7e66-bc8b-48e1-bd7c-3dbc3dfcceb9

[26] M. F. Gonzalez, J. F. Capman, F. L. Oswald, E. R. Theys, and D. L. Tomczak, "Where's the IO? Artificial intelligence and machine learning in talent management systems," *Personnel Assessment and Decisions*, vol. 5, no. 3, p. 5, 2019, doi: https://doi.org/10.25035/pad.2019.03.002.

[27] E. Mjolsness and D. DeCoste, "Machine learning for science: State of the art and future prospects," *Science*, vol. 293, no. 5537, pp. 2051–2055, 2001, doi: https://doi.org/10.1126/science.293.5537.2051.

[28] J. D. Kelleher, B. Mac Namee, and A. D'Arcy, *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. Cambridge, MA, USA: MIT Press, 2020, doi: https://doi.org/10.7551/mitpress/11187.001.0001.

[29] D. Abbott, *Applied predictive analytics: Principles and techniques for the professional data analyst*. Hoboken, NJ, USA: John Wiley & Sons, 2014, doi: https://doi.org/10.1002/9781118729274.

[30] D. T. Larose, *Data mining and predictive analytics*. Hoboken, NJ, USA: John Wiley & Sons, 2015, doi: https://doi.org/10.1002/9781118874059.

[31] P. Putra, M. Anam, S. Defit, and A. Yunianta, "Enhancing the decision tree algorithm to improve performance across various datasets," *INTENSIF: Jurnal Ilmiah Penelitian dan Penerapan Teknologi Sistem Informasi*, vol. 8, no. 2, pp. 200–212, 2024, doi: https://doi.org/10.29407/intensif.v8i2.20814.

[32] F. Buontempo, *Genetic algorithms and machine learning for programmers: Create AI models and evolve solutions*. Raleigh, NC, USA: The Pragmatic Bookshelf, 2019, doi: https://doi.org/10.16834/pragprog.9781680503128.

[33] E. Wirsansky, *Hands-on genetic algorithms with Python: Applying genetic algorithms to solve real-world deep learning and artificial intelligence problems*. Birmingham, UK: Packt Publishing, 2020, doi: https://doi.org/10.1201/9780429434099.

[34] O. Kramer, *K-nearest neighbors*. Berlin, Germany: Springer, 2013, doi: https://doi.org/10.1007/978-3-642-38652-7.

[35] S. M. Mazarei, "Predicting employee turnover using tree-based ensemble learning algorithms," *Signal and Data Processing*, vol. 20, no. 3, pp. 73–86, 2023, doi: https://doi.org/10.1007/s42979-023-01913-1.

[36] L. S. Khati and M. Kadariya, "Prediction of health care employee turnover using gradient boosting algorithm," *Journal of Science and Technology*, vol. 3, no. 1, pp. 66–73, 2023, doi: https://doi.org/10.3126/jost.v3i1.54053.

[37] M. A. Jafor, M. A. H. Wadud, K. Nur, and M. M. Rahman, "Employee promotion prediction using improved AdaBoost machine learning approach," *AIUB Journal of Science and Engineering*, vol. 22, no. 3, pp. 258–266, 2023, doi: https://doi.org/10.53799/ajse.v22i3.681.

[38] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier—An ensemble procedure for recall and precision enrichment," *Engineering Applications of Artificial Intelligence*, vol. 136, p. 108972, 2024, doi: https://doi.org/10.1016/j.engappai.2024.108972.

[39] E. Ramalakshmi and S. R. Kamidi, "Prediction of employee attrition and analyzing reasons: Using multi-layer perceptron in Spark," in *ICICCT 2019—System Reliability, Quality Control, Safety, Maintenance and Management: Applications to Electrical, Electronics and Computer Science and Engineering*, Singapore, 2020, pp. 1–10, doi: https://doi.org/10.1007/978-981-15-7961-5_52.

[40] U. Tiwari, S. Ashwani, A. J. Tripathy, and K. D. Kumar, "A two-stage ensemble approach for analysis of optimizing customer churn with LIME interpretability," in *Proc. IEEE Int. Conf. Computing, Power and Communication Technologies*, Greater Noida, India, 2024, pp. 1–10, doi: https://doi.org/10.1109/IC2PCT60068.2024.10486392.

[41] C. Liao, "Employee turnover prediction using machine learning models," in *Proc. Int. Conf. Mechatronics Engineering and Artificial Intelligence*, Changsha, 2022, pp. 1–10, doi: https://doi.org/10.1109/ICMEAI54614.2022.00042.

[42] L. M. S. V. Mesa I., "Algoritmos genéticos: Evolución, genética y computación," *Revista Cubana de Computación GIGA*, no. 4, pp. 26–33, 1998. [Online]. Available: https://repositorio.puce.edu.ec/server/api/core/bitstreams/904f2cff-5f87-4c0e-a444-09ce4dbb3046/content

[43] U. Murugesan, P. Subramanian, S. Srivastava, and A. Dwivedi, "A study of artificial intelligence impacts on human resource digitalization in Industry 4.0," *Decision Analytics Journal*, vol. 7, p. 100249, 2023, doi: https://doi.org/10.1016/j.dajour.2023.100249.

[44] C. I. Pinkas, *Age discrimination in hiring practices: A quantitative analysis*. Northcentral University, 2018. [Online]. Available: https://www.proquest.com/openview/65e2d0b2dd7b6446d40d55a7b8b3490c/1?pq-origsite=gscholar&cbl=18750

[45] A. L. García-Izquierdo, L. D. Vilela, and S. Moscoso, *Employee recruitment, selection, and assessment*. New York, NY, USA: Psychology Press, 2015, pp. 9–26, doi: https://doi.org/10.4324/9781315694382.

[46] A. B. Cardona Ramos, *La cultura organizacional y su impacto en el desempeño laboral como factor de pertenencia y sentido de lealtad dentro de las organizaciones*. IDEA, 2023. [Online]. Available: https://repository.umng.edu.co/items/93bccc92-23d5-4070-a7b6-881c61deca1c

[47] J. Lee, S. B. Kim, Y. Kim, and S. Kim, "Predictive analytics for efficient decision making in personnel selection," *International Journal of Management and Decision Making*, vol. 22, no. 1, pp. 106–122, 2023, doi: https://doi.org/10.1504/IJMDM.2023.128742.

[48] S. R. K. Indarapu, S. Vodithala, N. Kumar, S. Kiran, S. N. Reddy, and K. Dorthi, "Exploring human resource management intelligence practices using machine learning models," *The Journal of High Technology Management Research*, vol. 34, no. 2, p. 100466, 2023, doi: https://doi.org/10.1016/j.hitech.2023.100466.

[49] I. Muraina, "Ideal dataset splitting ratios in machine learning algorithms: General concerns for data scientists and data analysts," in *Proc. Int. Mardin Artuklu Scientific Research Conf.*, Lagos, Nigeria, 2022. [Online]. Available: https://www.researchgate.net/profile/Ismail-Muraina/publication/358284895_IDEAL_DATASET_SPLITTING_RATIOS_IN_MACHINE_LEARNING_ALGORITHMS_GENERAL_CONCERNS_FOR_DATA_SCIENTISTS_AND_DATA_ANALYSTS/links/61fb97e711a1090a79cc1a8b/IDEAL-DATASET-SPLITTING-RATIOS-IN-MACHINE-LEARNING-ALGORITHMS-GENERAL-CONCERNS-FOR-DATA-SCIENTISTS-AND-DATA-ANALYSTS.pdf

[50] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 9, pp. 6390–6404, 2022, doi: https://doi.org/10.1109/TNNLS.2021.3136503.

[51] V. Kotu and B. Deshpande, *Predictive analytics and data mining: Concepts and practice with RapidMiner*. Morgan Kaufmann, 2014, doi: https://doi.org/10.1016/C2012-0-06172-1.

[52] M. Atef, S. E. D., and S. Ouf, "Early prediction of employee turnover using machine learning algorithms," *International Journal of Electrical and Computer Engineering Systems*, vol. 13, no. 2, pp. 135–144, 2022, doi: https://doi.org/10.32985/ijeces.13.2.4.

[53] V. J. Recilla, M. R. A. Enonaria, R. J. Florida, J. C. M. Bustillo, C. C. Abalorio, and J. C. Trillo, "Predicting employee turnover through genetic algorithm," in *Proc. Int. Conf. Electronics and Sustainable Communication Systems (ICESC)*, 2024, pp. 1–10, doi: https://doi.org/10.1109/ICESC60838.2024.10469463.

[54] A. Dudáš, "Graphical representation of data prediction potential: Correlation graphs and correlation chains," *The Visual Computer*, vol. 40, pp. 6969–6982, 2024, doi: https://doi.org/10.1007/s00371-024-03263-2.