Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

Predictive Modelling of Employee Attrition Using Deep Learning

Dino Michael Quinteros*

Professional Academic School of Systems Engineering, Cesar Vallejo University, 15314 Lima, Peru

* Correspondence: Dino Michael Quinteros (dquinterosna@ucv.edu.pe)

Received: 09-13-2023 **Revised:** 11-05-2023 **Accepted:** 11-10-2023

Citation: D. M. Quinteros, "Predictive modelling of employee attrition using deep learning," *Acadlore Trans. Mach. Learn.*, vol. 2, no. 4, pp. 212–225, 2023. https://doi.org/10.56578/ataiml020404.

Abstract: This investigation delineates an optimised predictive model for employee attrition within a substantial workforce, identifying pertinent models tailored to the specific context of employee and organisational variables. The selection and refinement of the appropriate predictive model serve as cornerstones for enhancements and updates, which are integral to honing the model's precision in prognosticating potential departures. Through meticulous optimisation, the model demonstrates proficiency in pinpointing the pivotal factors contributing to employee turnover and elucidating the interdependencies among salient variables. A suite of 27 general and eight critical variables were scrutinised. Pertinent correlations were unearthed, notably between monthly income and job satisfaction, home-to-work distance and job satisfaction, as well as age with both job satisfaction and performance metrics. Drawing from prior studies in analogous domains, a three-stage analytical methodology encompassing data exploration, model selection, and implementation was employed. The rigorous training of the optimised model encompassed both attrition factors and variable correlations, culminating in predictive outcomes with a precision of 90% and an accuracy of 87%. Implementing the refined model projected that 113 out of 709 employees, equating to 15.93%, were at a heightened risk of exiting the organisation. This quantitative foresight equips stakeholders with a strategic tool for preemptive interventions to mitigate turnover and sustain organisational vitality.

Keywords: Deep learning; Employee attrition prediction; Predictive modelling

1 Introduction

Employee attrition presents a significant challenge within corporate structures, leading to considerable delays, costs, and elongated recruitment processes. The complexities that accompany staff resignations exacerbate risks associated with the attainment of annual corporate objectives. In the quest for an efficacious resolution, an optimised deep learning model has been developed to predict attrition with commendable accuracy and reliability. Comparative analyses demonstrate the model's superior performance against extant literature.

Notably, emerging economies in Latin America exhibit a lack of adequate policies and procedures to gauge work environment dynamics and potential attrition, affecting 45% of businesses. Contrarily, in the United States and Europe, although 60% of firms reportedly possess robust protocols to anticipate staff departures, the application of such measures remains suboptimal. Human resources management, while central to organizational function, frequently overlooks the pre-resignation phase, thereby complicating the transition period up to the onboarding of new staff. The challenges are magnified when the roles in question demand specialised skills [1].

Attrition, defined as the withdrawal from employment for varied reasons, both intrinsic and extrinsic, results in substantial operational and strategic gaps. The rate of attrition emerges as a crucial metric, albeit one often neglected in corporate analyses [2]. The pertinence of human resources to corporate health is undisputed, yet the mechanisms to effectively predict and mitigate employee turnover require further elucidation. Mansor et al. [3] advocate for the identification of employee-aligned characteristics to enhance predictive accuracy regarding attrition. Maximizing profitability through the utilization of technical personnel skilled in specialized tasks is recognized as a critical component for sustaining business operations. Accordingly, the recruitment and retention of such specialized staff are acknowledged as decisive factors in the continuance of commercial enterprises [1].

The withdrawal of specialised personnel from an organisation not only undermines the achievement of institutional goals but also precipitates a loss of skills, experience, and potential business opportunities, underscoring the significance of retention strategies [4]. Within this context, the efficacy of machine learning models for customer churn prediction is well-documented. Nevertheless, such traditional models falter when tasked with addressing the nuances of employee-related processes, thereby necessitating more tailored approaches [5].

Customised deep learning architectures are increasingly recognised for their capacity to refine forecasts and bolster decision-making frameworks, thus circumventing inefficiencies and resource depletion. It is imperative to identify the catalysts for employee departure [6]. Deep learning, a cornerstone of data science, utilises an array of techniques to achieve precise outcomes that inform strategic decisions [2]. Machine learning models, as delineated by Alsheref et al. [7], undergo a tripartite development process: initial data exploration, subsequent model selection, and final implementation.

In other studies, it was identified that the highest dropout rates in higher education students are associated with issues of personal and social cost. For this reason, the authors specify that it is necessary to know the students at risk and understand the main dropout factors [8]. Likewise, the factors that influence dropout are associated with background, family, academic records and socioeconomic status. These input factors could be associated with various data collected [9].

Tran et al. [10] posit that the identification of pivotal variables, contingent upon company-specific contexts, is crucial for model integration. They further utilised a K-means clustering algorithm, uncovering its significance in predicting employee turnover. It was found that accuracy peaked at 97% for smaller, homogeneous groups, while larger groups yielded 80% accuracy, underscoring the efficacy of grouping specialised employees in crucial operational processes.

Contrastive analysis of classification models revealed varying levels of predictive accuracy and interpretability [6]. Logistic regression, for instance, achieved an accuracy of 88% and a receiver operating characteristic (ROC) curve of 85%, underscoring the need to elucidate job abandonment factors. Pratt et al. [11] employed an assortment of models, including decision trees, binomial and logistic regression, support vector machines, and random forests, which collectively reached an average accuracy of 85%. Dake and Buabeng-Andoh [12] implemented a classification model aimed at predicting student attrition, considering a suite of 24 variables predominantly related to personal context and environmental factors. The model achieved 72% accuracy and highlighted the imperative for further optimisation to enhance predictive performance. Yahia et al. [4] deployed an attrition prediction model that mapped influential characteristics within a dataset of 450 employees, achieving 84% accuracy.

This study unveils an optimised deep learning model tailored for attrition prediction, necessitating diversified models replete with pertinent indicators, reflective of the categorical and institutional segments. The analysis method pertaining to the optimised model, alongside the resultant findings and deduced conclusions, are comprehensively delineated.

2 Methodology

2.1 Method

The methodology employed in this research incorporated an exploratory data analysis, adhering to the tripartite framework delineated by Alsheref et al. [7]. This framework comprises data exploration, model selection, and model implementation. Within the scope of this study, 27 variables were scrutinised to ascertain their impact on the model's predictive capability. It was identified that the most salient variables included: age of the employee, attrition, commute distance from house to workplace (Distance_House), monthly income (Income_M), job satisfaction regarding the employee's monthly salary (Satisfaction_I), job satisfaction with respect to the employee's age (Satisfaction_A), performance evaluations and job satisfaction regarding distance from the employee's home (Satisfaction_DH).

2.2 Data Exploration

Data description is a relevant process in data analysis, including collection, determination, cleaning, consistency and verifications. Therefore, consistent, complete and accurate data will provide results more efficiently [13]. The dataset under scrutiny comprised 709 records pertaining to employees within the life insurance sector, encompassing a diverse array of roles and attributes.

Table 1 presents the variables under consideration and their respective data types.

Exploratory data analysis was conducted to elucidate factors influencing employee attrition and to pinpoint the principal causes of turnover. The selection of variables was informed by a survey instrument whose validity was established through expert appraisal, focusing on relevance and consistency in the instrument's implementation. Various analytical techniques, such as graphical representations, time-series analyses, and correlation matrices, were employed. These methods facilitated the identification of data patterns instrumental in forecasting the ramifications of employee attrition and underscored the significance of factor identification. A notable correlation was observed between the variables of 'Distance_House' and 'Age', and indices of satisfaction and performance, thus warranting a detailed examination of these relationships due to their potential impact.

Figure 1 delineates the time series and correlations among 'Distance_House' and 'Age' with 'Satisfaction_DH', 'Satisfaction_A', and 'Performance'.

Table 1. Variables and types

No.	Variable	Type	No.	Variable	Type
0	Age	ext. 64	14	Job_Level	ext. 64
1	Desertion	object	15	Role	object
2	Business_Trip	object	16	Satisfaction_I	ext. 64
3	Daily_Rate	ext. 64	17	Status_P	object
4	Area	object	18	Income_M	ext. 64
5	Distance_House	ext. 64	19	Monthly_Rate	ext. 64
6	Education	ext. 64	20	Companies_W	ext. 64
7	Profession	object	21	Over_18	object
8	Employee	ext. 64	22	Over_E	object
9	Employee_N	ext. 64	23	Salary_Por	ext. 64
10	Satisfaction_A	ext. 64	24	Performance art	ext. 64
11	Gender	object	25	Satisfaction_DH	ext. 64
12	Hourly_Rate	ext. 64	26	StandardHours	ext. 64
13	Job_Part	ext. 64			

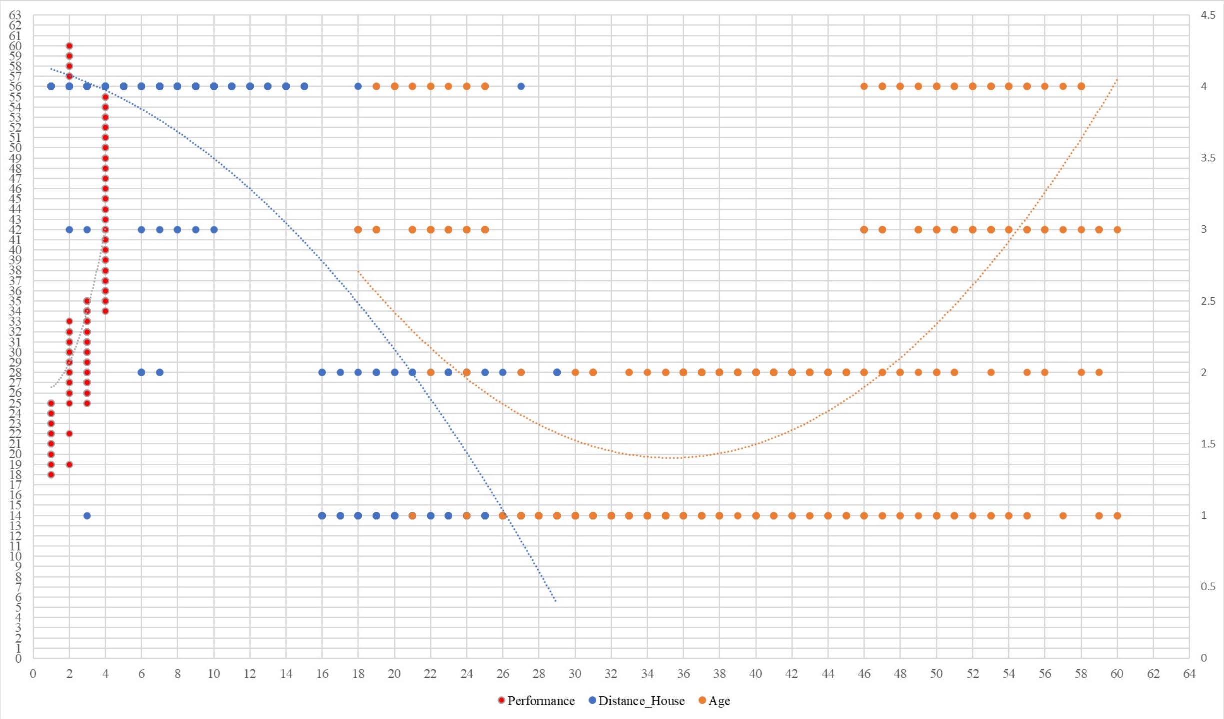


Figure 1. Time series and correlations (Distance_House and Age with Satisfaction_DH, Satisfaction_A and Performance)

Data distributions and line graphs were utilized for a suite of variables, including 'Age', 'Attrition', 'Daily_Rate', 'Distance_House_Education', 'Employee', 'Employee_N', 'Satisfaction_E', 'Hourly_Rate', 'Job_Part', 'Job_Level', 'Job_Satisfaction', 'Income_M', 'Monthly_Rate', 'Companies_W', 'Over_18', 'Over_E', 'Salary_Por', 'Performance', 'Satisfaction', and 'StandardHours'. Predominant levels identified within these variables necessitated correlation to each corresponding factor. The variables exhibited significant attributes that were imperative for discerning specific employee characteristics, thereby enhancing the precision of attrition predictions. Such predictions are informed not solely by the general corporate context but also by the individual contexts of employees. The dataset was thus required to embody a degree of granularity that would allow for substantive relational analyses among the variables.

Figure 2 illustrates the dataset characteristics.

For the efficacy of the predictive model, two critical criteria were considered essential: the nature of the dataset and the inter-variable correlations.

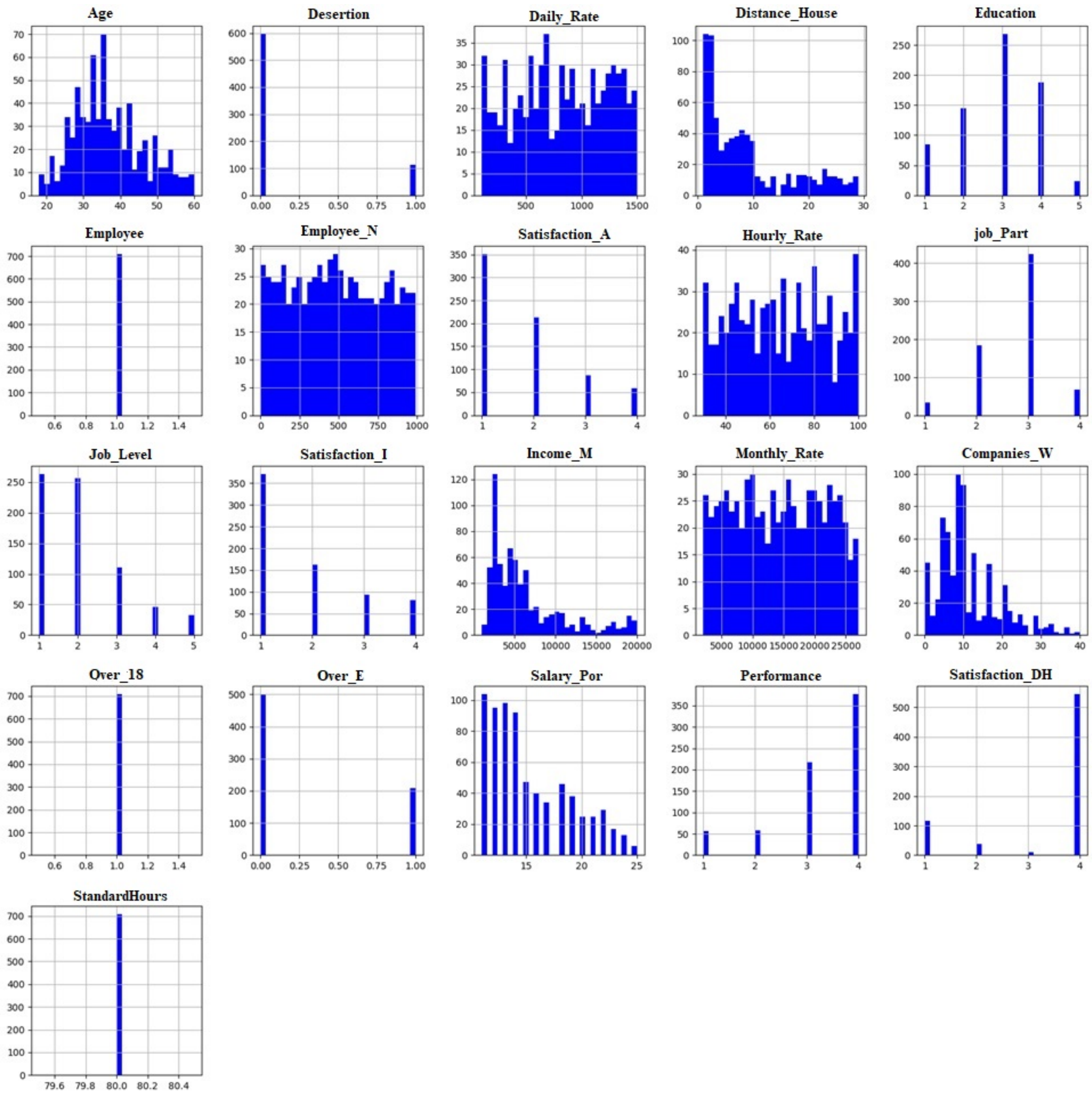


Figure 2. Characteristics of the dataset

2.3 Model Selection

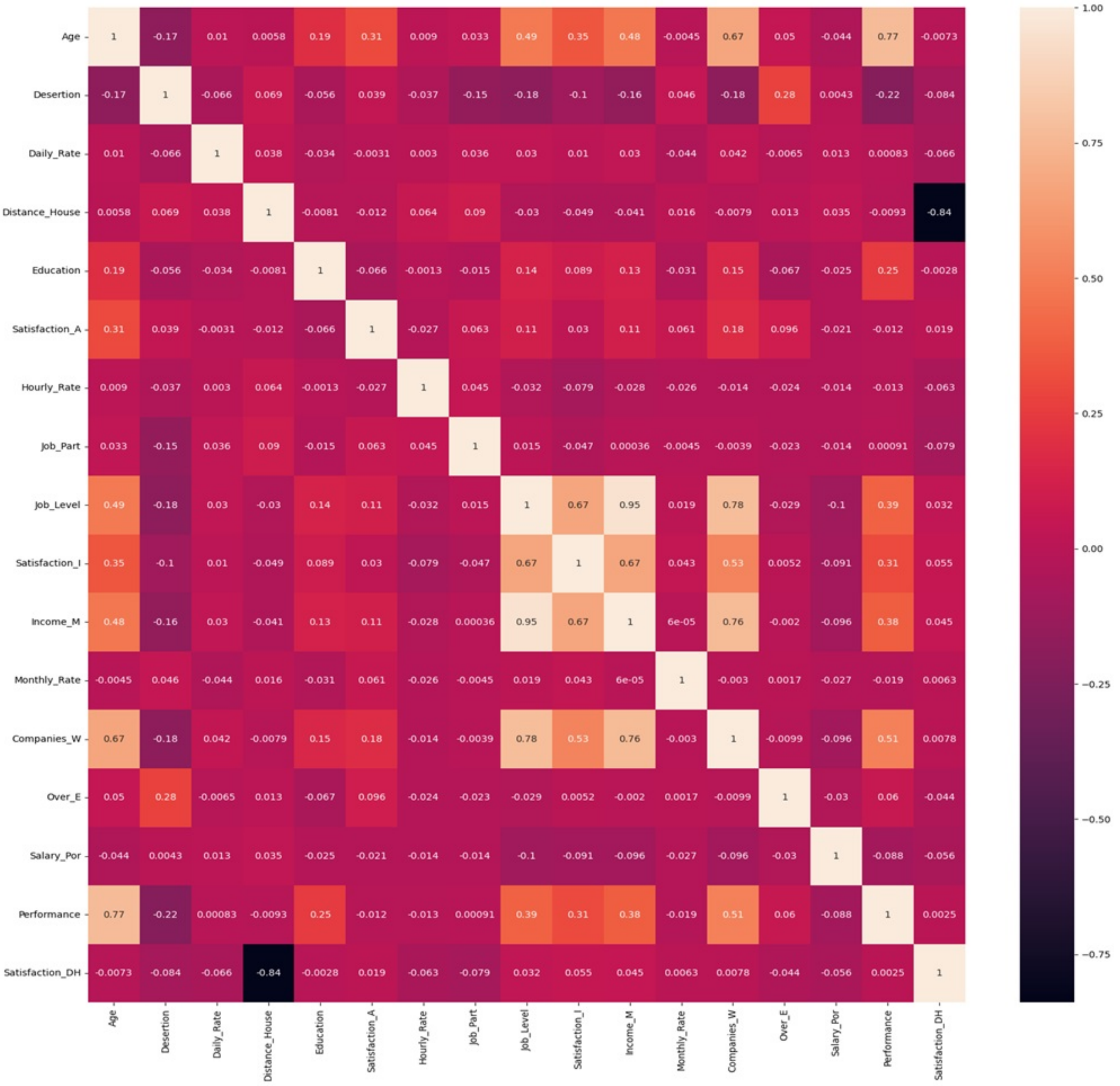
The alignment of the study's objectives with the selection of an appropriate deep learning model necessitated a meticulous review of literature and subsequent empirical evaluation. For this purpose, general and specific contextual variables were meticulously considered. From the range of models identified in prior research, three were subjected to rigorous testing against the dataset to determine their congruence with the study's objectives: logistic regression, random forest, and neural networks.

2.3.1 Logistic regression model

In instances where recurrent application and effective classification are paramount, logistic regression models have been favoured. These models necessitate a substantial volume of data records, interlinked with relevant variables tailored for specific contexts [14]. The logistic regression model was applied to the dataset, yielding a precision of 85% and an accuracy of 83%.

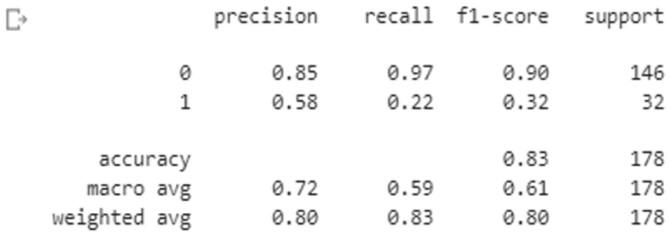
Figure 3 displays the correlations between variables.

Figure 4 depicts the outcomes of the logistic regression model.



	Age	Desertion	Daily_Rate	Distance_House	Education	Satisfaction_A	Hourly_Rate	Job_Part	Job_Level	Satisfaction_I	Income_M	Monthly_Rate	Companies_W	Over_E	Salary_Por	Performance	Satisfaction_DH
Age	1	-0.17	0.01	0.0058	0.19	0.31	0.009	0.033	0.49	0.35	0.48	-0.0045	0.67	0.05	-0.044	0.77	-0.0073
Desertion	-0.17	1	-0.066	0.069	-0.056	0.039	-0.037	-0.15	-0.18	-0.1	-0.16	0.046	-0.18	0.28	0.0043	-0.22	-0.084
Daily_Rate	0.01	-0.066	1	0.038	-0.034	-0.0031	0.003	0.036	0.03	0.01	0.03	-0.044	0.042	-0.0065	0.013	0.00083	-0.066
Distance_House	0.0058	0.069	0.038	1	-0.0081	-0.012	0.064	0.09	-0.03	-0.049	-0.041	0.016	-0.0079	0.013	0.035	-0.0093	-0.84
Education	0.19	-0.056	-0.034	-0.0081	1	-0.066	-0.0013	-0.015	0.14	0.089	0.13	-0.031	0.15	-0.067	-0.025	0.25	-0.0028
Satisfaction_A	0.31	0.039	-0.0031	-0.012	-0.066	1	-0.027	0.063	0.11	0.03	0.11	0.061	0.18	0.096	-0.021	-0.012	0.019
Hourly_Rate	0.009	-0.037	0.003	0.064	-0.0013	-0.027	1	0.045	-0.032	-0.079	-0.028	-0.026	-0.014	-0.024	-0.014	-0.013	-0.063
Job_Part	0.033	-0.15	0.036	0.09	-0.015	0.063	0.045	1	0.015	-0.047	0.00036	-0.0045	-0.0039	-0.023	-0.014	0.00091	-0.079
Job_Level	0.49	-0.18	0.03	-0.03	0.14	0.11	-0.032	0.015	1	0.67	0.95	0.019	0.78	-0.029	-0.1	0.39	0.032
Satisfaction_I	0.35	-0.1	0.01	-0.049	0.089	0.03	-0.079	-0.047	0.67	1	0.67	0.043	0.53	0.0052	-0.091	0.31	0.055
Income_M	0.48	-0.16	0.03	-0.041	0.13	0.11	-0.028	0.00036	0.95	0.67	1	6e-05	0.76	-0.002	-0.096	0.38	0.045
Monthly_Rate	-0.0045	0.046	-0.044	0.016	-0.031	0.061	-0.026	-0.0045	0.019	0.043	6e-05	1	-0.003	0.0017	-0.027	-0.019	0.0063
Companies_W	0.67	-0.18	0.042	-0.0079	0.15	0.18	-0.014	-0.0039	0.78	0.53	0.76	-0.003	1	-0.0099	-0.096	0.51	0.0078
Over_E	0.05	0.28	-0.0065	0.013	-0.067	0.096	-0.024	-0.023	-0.029	0.0052	-0.002	0.0017	-0.0099	1	-0.03	0.06	-0.044
Salary_Por	-0.044	0.0043	0.013	0.035	-0.025	-0.021	-0.014	-0.014	-0.1	-0.091	-0.096	-0.027	-0.096	-0.03	1	-0.088	-0.056
Performance	0.77	-0.22	0.00083	-0.0093	0.25	-0.012	-0.013	0.00091	0.39	0.31	0.38	-0.019	0.51	0.06	-0.088	1	0.0025
Satisfaction_DH	-0.0073	-0.084	-0.066	-0.84	-0.0028	0.019	-0.063	-0.079	0.032	0.055	0.045	0.0063	0.0078	-0.044	-0.056	0.0025	1

Figure 3. Correlations between variables



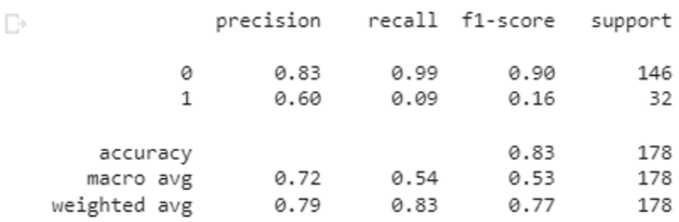
```
              precision    recall  f1-score   support

           0       0.85      0.97      0.90       146
           1       0.58      0.22      0.32        32

    accuracy                           0.83       178
   macro avg       0.72      0.59      0.61       178
weighted avg       0.80      0.83      0.80       178
```

Figure 4. Results of the logistic regression model

2.3.2 Random forest model

Random forest models employ an ensemble of decision trees to converge upon a singular objective, acclaimed for their interpretability and alignment with research aims [11]. Deployment of the dataset through a random forest model achieved a precision and accuracy of 83%.

Figure 5 illustrates the results derived from the random forest model.



```
              precision    recall  f1-score   support

           0       0.83      0.99      0.90       146
           1       0.60      0.09      0.16        32

    accuracy                           0.83       178
   macro avg       0.72      0.54      0.53       178
weighted avg       0.79      0.83      0.77       178
```

Figure 5. Results of the random forest model

2.3.3 Neural network model

Neural networks, an emergent model within the realm of deep learning, have demonstrated prowess in processing extensive variable sets expeditiously. These models are renowned for their innovative techniques in resolving business challenges. The architecture comprises an input layer, which introduces patterns into the network, and several hidden layers where data processing occurs. Connections within these layers, characterized by weights and biases, facilitate the computational processes. Upon receiving input, neurons calculate a weighted sum, inclusive of bias. The resultant value, in conjunction with an activation function, dictates neuron activation, leading to data transmission across the network. The utilisation of a non-linear activation function is chiefly for binary classification models, where the prediction of probabilities is required [11].

Within the neural network architecture, each neuron is programmed to compute a weighted sum of the received input values, to which a bias term is added. It is then subjected to an activation function, which determines the neuron's activation status based on the calculated sum. Subsequently, the activated neuron propagates the information through the network, eventually reaching the final hidden layer, which connects to the output layer. The employment of a non-linear activation function facilitates the scaling of output data, which is primarily utilized in binary classification models for the prediction of probabilities as outcomes.

Figure 6 presents the nonlinear artificial neural network (ANN) model.

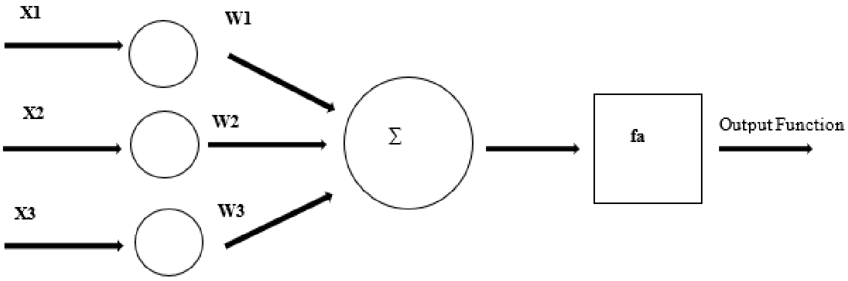
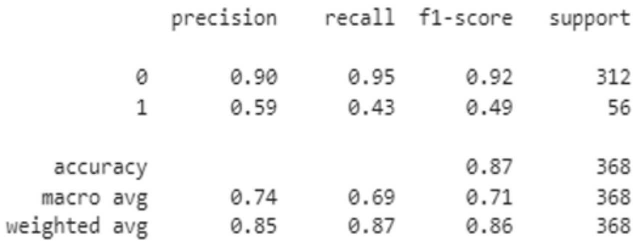


Figure 6. Nonlinear ANN model

Figure 7 displays results of the neural network model.



```
              precision    recall  f1-score   support

           0       0.90      0.95      0.92       312
           1       0.59      0.43      0.49        56

    accuracy                           0.87       368
   macro avg       0.74      0.69      0.71       368
weighted avg       0.85      0.87      0.86       368
```

Figure 7. Results of the neural network model

When subjected to the neural network model, the dataset revealed a precision of 90% and an accuracy of 87%. Therefore, the neural network model was identified as the most suitable foundation for the development of an optimised predictive model for attrition.

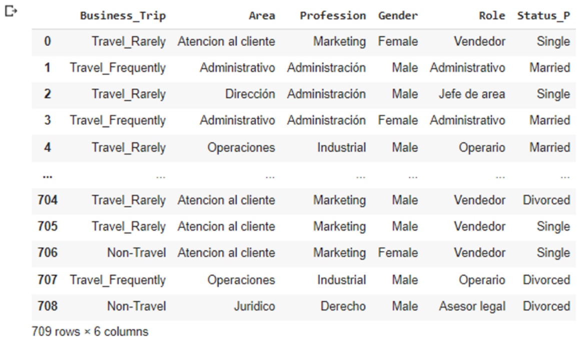
The selection criteria were twofold: first, the presence of correlated variables within the model; second, the attainment of superior precision and accuracy in predictive capability.

2.4 Implementation of the Selected Model

Within the selected model, 27 variables were identified as possessing specific characteristics pertinent to the employee's role within the internal context of the organization. Initial deployment of this model yielded a precision of 90% and an accuracy of 87%. This model served as the precursor for advancements towards an optimized predictive model. Consequently, variables intrinsic to the employee's environment were incorporated, namely: Business_Trip, Area, Profession, Gender, Role, and Status_P. These variables were imperative for initiating the data cleansing process and subsequent training phases.

For the data cleansing stage, a methodology was employed whereby variables such as Attrition, Over18, and Overtime were encoded as integers to facilitate visualization. Variables deemed non-essential to the model's predictive capacity, including Employee, StandardHours, Over_18, and Employee_N, were excised from the dataset. Training the model involved utilizing the attrition column as the target variable, influencing the component's capacity to discern patterns. Training commenced with subsets of data: an initial ten events yielded a learning accuracy of 65%, escalating to 80% with 50 events, and surpassing 95% upon the introduction of 100 events.

Figure 8 illustrates the variables relevant to the model implementation.



	Business_Trip	Area	Profession	Gender	Role	Status_P
0	Travel_Rarely	Atencion al cliente	Marketing	Female	Vendedor	Single
1	Travel_Frequently	Administrativo	Administración	Male	Administrativo	Married
2	Travel_Rarely	Dirección	Administración	Male	Jefe de area	Single
3	Travel_Frequently	Administrativo	Administración	Female	Administrativo	Married
4	Travel_Rarely	Operaciones	Industrial	Male	Operario	Married
...	...	...	...	...	...	...
704	Travel_Rarely	Atencion al cliente	Marketing	Male	Vendedor	Divorced
705	Travel_Rarely	Atencion al cliente	Marketing	Male	Vendedor	Single
706	Non-Travel	Atencion al cliente	Marketing	Female	Vendedor	Single
707	Travel_Frequently	Operaciones	Industrial	Male	Operario	Divorced
708	Non-Travel	Juridico	Derecho	Male	Asesor legal	Divorced

709 rows × 6 columns

Figure 8. Relevant variables

3 Results

3.1 Training of the Optimised Model

Subsequent to the data cleansing process, the variables retained were curated to enhance the model's predictive performance, necessitating the validation of tests and the pertinence of outcomes. The training was executed in three sequential deployments. Initially, an assemblage of ten events was processed, followed by a cohort of 50 events, and culminating in a final set of 100 events. It was observed that an accuracy threshold exceeding 95% was attained in the ultimate deployment. Training with a dataset comprising 100 events was determined to be requisite for the augmentation of the model's accuracy.

Figure 9 delineates the model's accruing accuracy throughout the training phases.

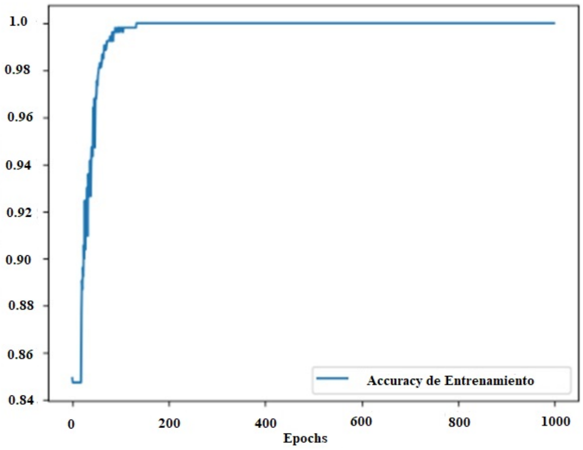
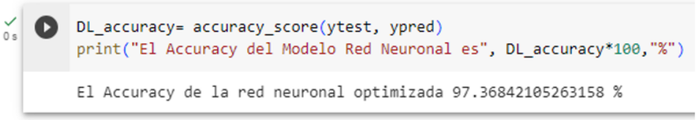


Figure 9. Accuracy of the model during training

3.2 Results of the Optimised Model

Upon the completion of the training process, the optimized model demonstrated a commendable accuracy of 97.36%.

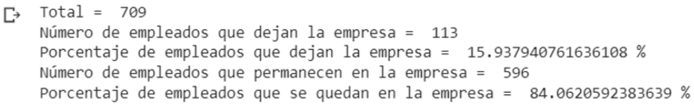
Figure 10 illustrates the results obtained from the optimized model.



```
DL_accuracy= accuracy_score(ytest, ypred)
print("El Accuracy del Modelo Red Neuronal es", DL_accuracy*100,"%")

El Accuracy de la red neuronal optimizada 97.36842105263158 %
```

Figure 10. Results of the optimized model

Concomitantly, the deployment phase allowed for the prediction of the prospective attrition among the workforce.



```
Total =  709
Número de empleados que dejan la empresa =  113
Porcentaje de empleados que dejan la empresa =  15.937940761636108 %
Número de empleados que permanecen en la empresa =  596
Porcentaje de empleados que se quedan en la empresa =  84.0620592383639 %
```

Figure 11. Forecasted number of employees to leave the company

Figure 11 presents the forecasted number of employees predicted to sever their association with the company.

The ROC curve analysis was conducted to assess the binary classification capability of the model, segregating employees into either a positive (likely to leave) or negative (likely to remain) category. The results indicated a robust model performance, characterized by a substantial area under the ROC curve, i.e. area under the curve (AUC), approaching the optimal value of 1, specifically notated as 0.97, which signifies a high level of precision.

Figure 12 depicts the test results of the ROC curve.

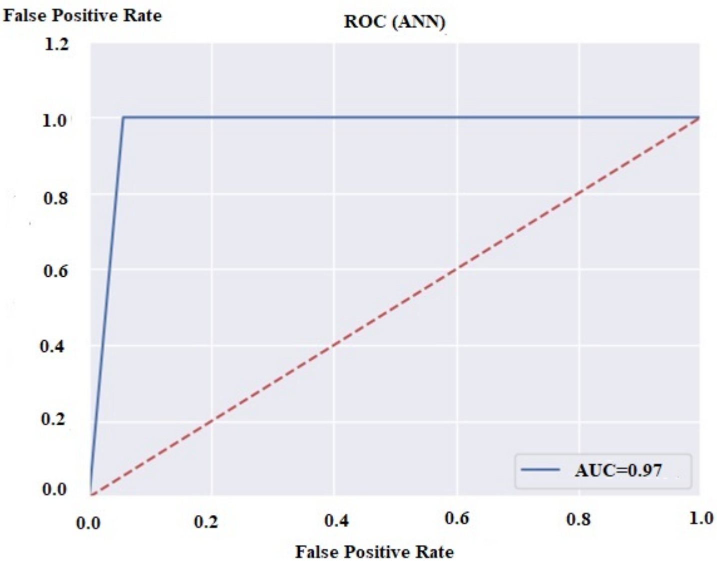


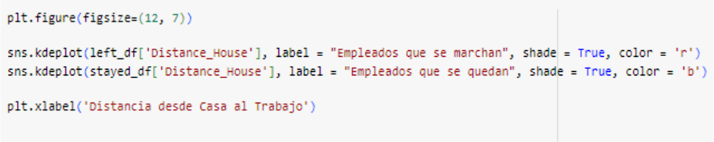
Figure 12. Results of the optimized model

3.3 Determinants of Employee Attrition

Analysis of the distance between home and workplace as a variable revealed a correlation with employee attrition; it was observed that an increased distance is associated with a higher attrition rate. Conversely, employees residing in proximity to the workplace were predominantly those who remained within the company. This suggests that proximity may play a deterrent role in the likelihood of employee turnover. The findings point to a necessity for a broader evaluation of contributory factors in this domain.

Figure 13 details the analytical code utilized for assessing the impact of commute distance.

Figure 14 provides a visual representation of the correlation between home-to-work distance and job attrition.



```
plt.figure(figsize=(12, 7))

sns.kdeplot(left_df['Distance_House'], label = "Empleados que se marchan", shade = True, color = 'r')
sns.kdeplot(stayed_df['Distance_House'], label = "Empleados que se quedan", shade = True, color = 'b')

plt.xlabel('Distancia desde Casa al Trabajo')
```

Figure 13. Code identifying the home-to-work factor

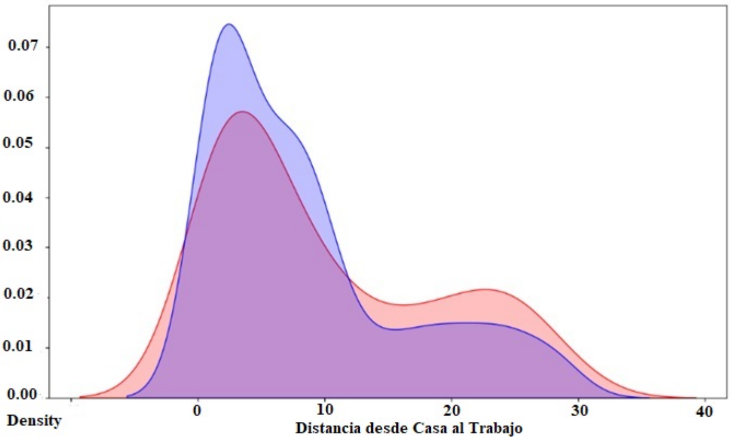
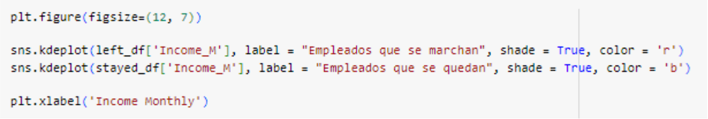


Figure 14. Impact of the home-to-work factor

Furthermore, the investigation into monthly income as a variable indicated a pattern wherein employees with lower income brackets exhibited a higher tendency to leave the company, while those earning higher wages were more inclined to stay. This underscores the significance of financial compensation in employee retention strategies.

Figure 15 illustrates the code employed to determine the influence of monthly income.



```
plt.figure(figsize=(12, 7))

sns.kdeplot(left_df['Income_M'], label = "Empleados que se marchan", shade = True, color = 'r')
sns.kdeplot(stayed_df['Income_M'], label = "Empleados que se quedan", shade = True, color = 'b')

plt.xlabel('Income Monthly')
```

Figure 15. Code identifying the monthly income factor

Figure 16 displays the relationship between monthly income levels and employee retention outcomes.

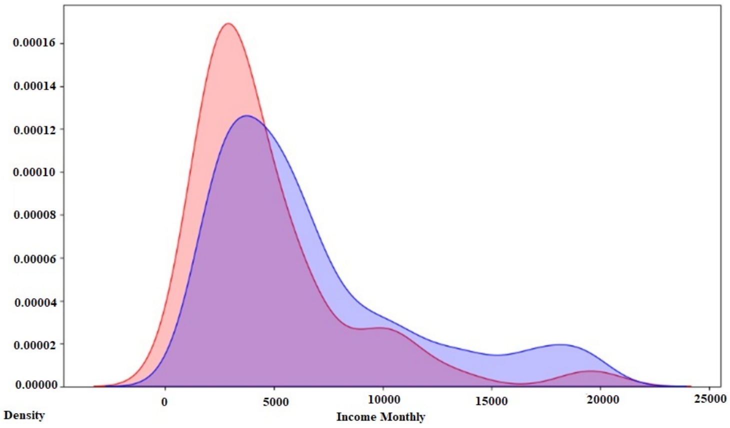
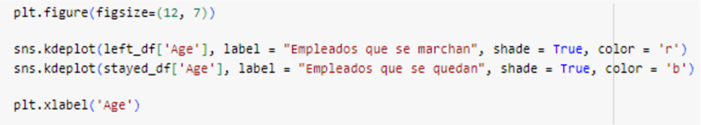


Figure 16. Impact of the monthly income factor

Age was also scrutinized as a relevant variable. The data suggest a lower attrition rate among employees aged between 30 and 40 years, while a higher departure rate was noted among younger employees, specifically those aged 20 to 30 years. This demographic distribution intimates a potential focus area for enhancing retention policies.

Figure 17 depicts the code utilized to elucidate the age factor in employee attrition.



```
plt.figure(figsize=(12, 7))

sns.kdeplot(left_df['Age'], label = "Empleados que se marchan", shade = True, color = 'r')
sns.kdeplot(stayed_df['Age'], label = "Empleados que se quedan", shade = True, color = 'b')

plt.xlabel('Age')
```

Figure 17. Code identifying the age factor

Figure 18 shows the distribution of attrition across different age groups.

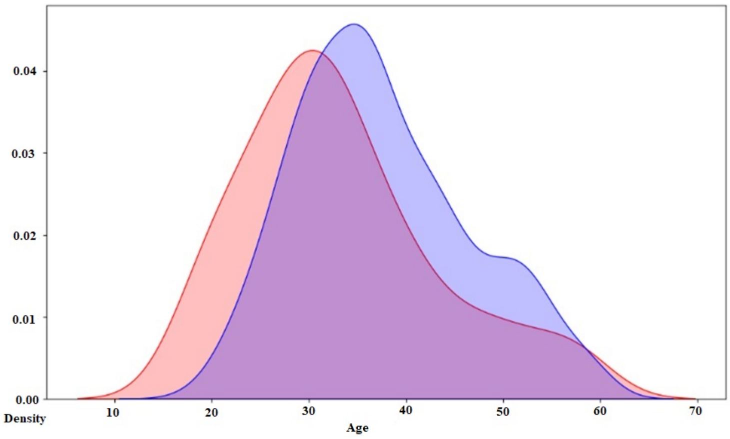


Figure 18. Impact of the age factor

3.4 Correlations

Correlational studies were conducted to ascertain the relationships between various employee attributes and job satisfaction, as well as between age and job performance. Employee remuneration emerged as a critical variable, with its influence on job satisfaction being a determining factor in the potential for job attrition. It was found that higher salaries are closely associated with elevated levels of employee satisfaction, thereby reducing the propensity for job desertion.

Figure 19 and Table 2 delineate the relationship between salary and employee satisfaction.

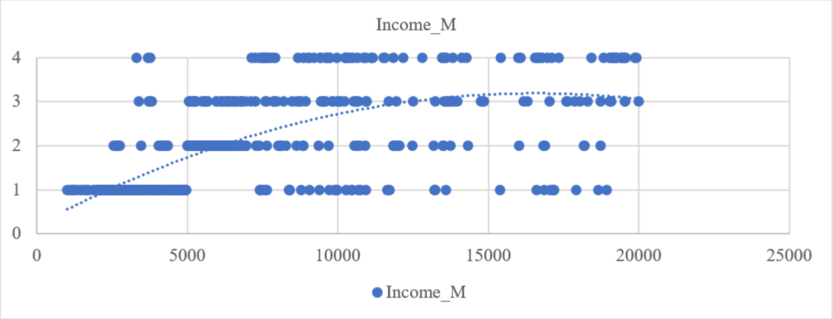


Figure 19. Relationship between monthly income and satisfaction

Table 2 presents the findings from Spearman's rho test regarding the correlation coefficient, delineating the relationship between monthly income and employee satisfaction. A significant correlation has been identified, with p=0.000$<\alpha$=0.01. This establishes a statistically significant relationship between monthly income and satisfaction levels. Specifically, a higher salary is correlated with increased satisfaction, as indicated by a Spearman's rho of 0.492, denoting a moderate positive correlation.

The commute distance was also scrutinized as a variable of interest. It was observed that greater commute distances are inversely related to employee satisfaction, which could potentially lead to increased job turnover.

Figure 20 provides an analysis of the impact of commute distance on employee satisfaction.

Table 2. Correlation between monthly income and satisfaction

Correlations		
	Monthly income	Satisfaction
Correlation coefficient	1,000	$0,492^{**}$
Follow-up (bilateral)	.	0,000
N	709	709
Correlation coefficient	$0,492^{**}$	1,000
Follow-up (bilateral)	0,000	.
N	709	709
**Correlation is significant at the 0.01 level (two-sided).		

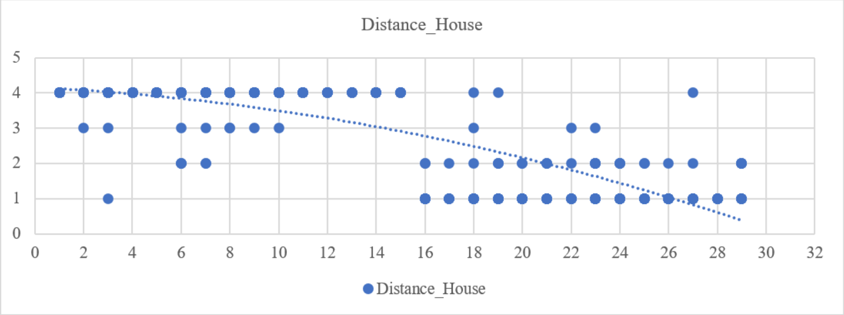


Figure 20. Impact of commute distance on satisfaction

Table 3 presents the statistical relationship between commute distance and satisfaction.

Table 3. Correlation between commute distance and satisfaction

Correlations		
	Commute distance	Satisfaction
Correlation coefficient	1,000	$0,392^{**}$
Follow-up (bilateral)	.	0,000
N	709	709
Correlation coefficient	$0,392^{**}$	1,000
Follow-up (bilateral)	0,000	.
N	709	709
**Correlation is significant at the 0.01 level (two-sided).		

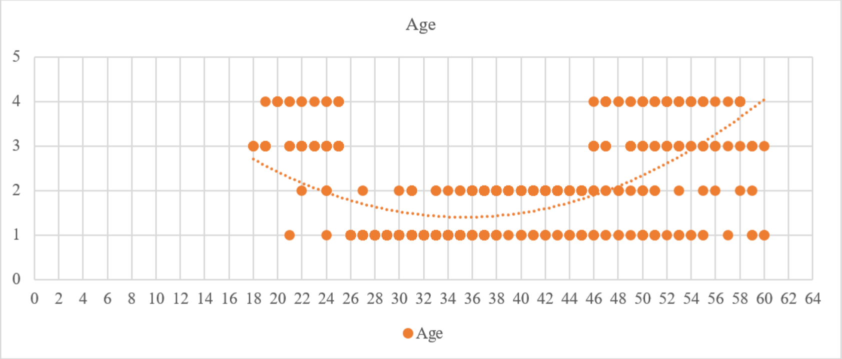


Figure 21. Impact of age on satisfaction

The results, as indicated in Table 3, were derived from Spearman's rho test, which elucidated the correlation coefficient between the distance from home to the workplace and employee satisfaction. The test yielded p=0.000 $<\alpha$=0.01. This confirms a statistically significant correlation; specifically, a shorter commute is associated with an elevated level of satisfaction among employees. The Spearman's rho of 0.392 signifies a low but positive correlation.

Analysis of age dynamics was also conducted, which revealed distinct satisfaction levels across age brackets. Employees aged between 18 and 25, as well as those over 45, exhibit higher degrees of satisfaction, potentially attributable to the distinct career stages represented by these age groups. Conversely, those aged between 25 and 45 showed lower satisfaction levels, which may predispose them to higher turnover.

Figure 21 and Table 4 illustrate the correlation between age groups and employee satisfaction.

As presented in Table 4, the results obtained from Spearman's rho test indicate the correlation coefficient between the variable of age and employee satisfaction. With p=0.006$<\alpha$=0.01, a statistically significant correlation is established. It has been observed that an increase in age is correlated with higher levels of satisfaction within the workplace. Nevertheless, with a rho value of 0.216, this correlation is classified as low positive.

Regarding age and performance, the data indicated that optimal performance is achieved between the ages of 35 and 55. Meanwhile, younger employees (aged 18-25) and those over 55 exhibit lower performance metrics.

Table 4. Correlation between age and satisfaction

Correlations		
	Age	Satisfaction
Correlation coefficient	1,000	$0,216^{**}$
Follow-up (bilateral)	.	0,006
N	159	159
Correlation coefficient	$0,216^{**}$	1,000
Follow-up (bilateral)	0,006	.
N	709	709
**Correlation is significant at the 0.01 level (two-sided).		

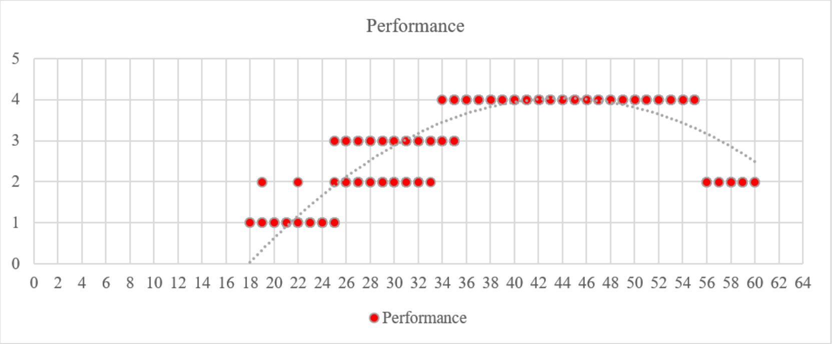


Figure 22. Impact of age on satisfaction

Figure 22 and Table 5 showcase the correlation between age and job performance.

Table 5 delineates the correlation between employee age and performance, as evidenced by the Spearman's rho test. With p=0.000$<\alpha$=0.01 indicates a highly significant correlation between the two variables. It is inferred from this result that an increase in employee age is associated with an enhancement in performance. However, the correlation coefficient, rho, valued at 0.210, signifies a modestly positive correlation.

Table 5. Correlation between age and performance

Correlations		
	Age	Performance
Correlation coefficient	1,000	$0,210^{**}$
Follow-up (bilateral)	.	0,000
N	709	709
Correlation coefficient	$0,318^{**}$	1,000
Follow-up (bilateral)	0,000	.
N	709	709
**Correlation is significant at the 0.01 level (two-sided).		

4 Discussion

The logistic regression model was examined, with its direct application to specialized company processes yielding precision and accuracy rates exceeding 83% (precision: 85%, accuracy: 83%). However, it was noted that this model does not account for variable correlation across diverse contexts, which limits its capacity for optimization based on the variables presented. The random forest model, designed to generate assertiveness without necessitating contextual variable correlations, consistently demonstrated results with an 83% success rate in both precision and accuracy. This uniformity in performance denotes its robustness. In contrast, the neural network model was capable of integrating and correlating multiple variables within employee-specific contexts as defined by the company. This model showed a superior performance with results surpassing 87% (90% precision and 87% accuracy). Subsequently, the neural network was selected based on two criteria: the inter-correlation of variables and its superior precision and accuracy. It encompassed 27 variables reflecting employee characteristics pertinent to the internal company context.

An optimized deep learning model was further considered, incorporating techniques such as identification of contextually relevant variables. This model achieved a noteworthy accuracy of 97.36% following training. The optimized model demonstrated enhanced performance relative to models proposed by Guerranti and Dimitri [6], who reported an 88% accuracy, and Pratt et al. [11] who documented an 85% accuracy average. The selection of variables representative of the company's context was instrumental in these improvements. The findings from the optimized model provide insights into predicting labor attrition based on specific and representative variables within the company's context. Future research should focus on identifying new variables that can refine the training process, thus enhancing accuracy and precision, and reducing the time required for model training. Finally, the significant monetary and labor losses incurred by companies, particularly small and medium-sized enterprises, due to employee attrition within the first six to 12 months post-hiring, are noteworthy. The optimized model has proven effective in identifying variables that may lead to attrition, thereby enabling strategies for improved hiring processes and employee retention.

5 Conclusions

Through the implementation of the optimized model, it was discerned that an estimated 113 (15.93%) of 709 employees are predicted to leave the organization. Predominant variables influencing attrition were identified as salary, distance from home, age, and performance metrics. It is imperative for organizations to address these variables to enhance employee retention strategies. Comparative analyses of logistic regression, random forest, and neural network models revealed constraints in pinpointing variables of significance that align with a company's specific context. Traditional models typically factor in one or two variables, whereas the optimized model incorporates a broader spectrum of seven pertinent variables: Age, Attrition, Distance_House, Income_M, Satisfaction_I, Satisfaction_A, Performance, and Satisfaction_DH, along with four inter-variable correlations. A robust correlation between monthly income and job satisfaction was observed, where a moderate positive correlation (rho=0.492 and p=0.000) was reported, suggesting that higher salaries are commensurate with increased employee satisfaction.

A correlation between the employee's distance from home to the workplace and their level of satisfaction was identified, with a lower positive correlation (rho=0.392 and p=0.000) indicating that closer proximity correlates with heightened satisfaction. The relationship between employee age and job satisfaction was also significant, albeit displaying a lower positive correlation (rho=0.216 and p=0.006), implying that satisfaction tends to augment with age. Additionally, a significant yet low positive correlation was noted between age and performance (rho=0.210 and p=0.000), indicating that employee performance incrementally improves with age. Future research should prioritize the integration of variables reflecting significant personal attributes, labor market trends, and entrepreneurial initiatives. This approach is anticipated to refine the model's predictive accuracy towards an optimal 100% threshold. Furthermore, a thorough examination of the identified correlations is warranted to elucidate the dynamics influencing variable interrelationships and their implications for continuous improvement strategies.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] F. Fallucchi, F. M. Coladangelo, R. Juliano, and E. W. D. Luca, "Predicting employee attrition using machine learning techniques," *Computers*, vol. 9, no. 4, p. 86, 2020. https://doi.org/10.3390/computers9040086

[2] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," *Appl. Sci.*, vol. 12, no. 13, p. 6424, 2022. https://doi.org/10.3390/app12136424

[3] N. Mansor, N. S. Sani, and M. Aliff, “Machine learning to predict employee attrition,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 435–445, 2021.

[4] N. B. Yahia, J. Hlel, and R. Colomo-Palacios, “From big data to deep data to support people analytics for employee attrition prediction,” *IEEE Access*, vol. 9, pp. 60 447–60 458, 2021. https://doi.org/10.1109/ACCESS.2021.3074559

[5] R. Babatunde, S. O. Abdulsalam, O. A. Abdulsalam, and M. O. Arowolo, “Classification of the customer churn prediction model for the telecommunications industry using analysis of variance,” *IAES Int. J. Artif. Intell.*, pp. 1323–1329, 2021. https://doi.org/10.11591/ijai.v12.i3.pp1323-1329

[6] F. Guerranti and G. M. Dimitri, “A comparison of machine learning approaches for predicting employee attrition,” *Appl. Sci.*, vol. 13, no. 1, p. 267, 2022. https://doi.org/10.3390/app13010267

[7] F. K. Alsheref, I. E. Fattoh, and W. M. Ead, “Automated prediction of employee attrition using ensemble model based on machine learning algorithms,” *Comput. Intell. Neurosci.*, pp. 1–9, 2022. https://doi.org/10.1155/2022/7728668

[8] J. Alvarado-Uribe, P. Mejía-Almada, A. L. Masetto Herrera, R. Molontay, I. Hilliger, V. Hegde, J. E. Montemayor Gallegos, R. A. Ramírez Díaz, and H. G. Ceballos, “Student dataset from tecnologico de monterrey in mexico to predict dropout in higher education,” *Data*, vol. 7, no. 9, p. 119, 2022. https://doi.org/10.3390/data7090119

[9] R. Ajoodha, “Identifying academically vulnerable learners in first-year science programmes at a South African higher-education institution,” *South Afr. Comput. J.*, vol. 34, no. 2, 2022. https://doi.org/10.18489/sacj.v34i2.832

[10] H. D. Tran, N. Le, and V. H. Nguyen, “Customer churn prediction in the banking sector using machine learning-based classification models,” *Interdiscip. J. Inf. Knowl. Manag.*, vol. 18, pp. 87–105, 2023. https://doi.org/10.28945/5086

[11] M. Pratt, M. Boudhane, and S. Cakula, “Employee attrition estimation using random forest algorithm,” *Balt. J. Mod. Comput.*, vol. 9, no. 1, pp. 49–66, 2021. https://doi.org/10.22364/bjmc.2021.9.1.04

[12] D. K. Dake and C. Buabeng-Andoh, “Using machine learning techniques to predict learner drop-out rate in higher educational institutions,” *Mob. Inf. Syst.*, vol. 2022, pp. 1–9, 2022. https://doi.org/10.1155/2022/2670562

[13] S. Shilbayeh and A. Abonamah, “Predicting student enrolments and attrition patterns in higher educational institutions using machine learning,” *Int. Arab J. Inf. Technol.*, vol. 18, no. 4, pp. 562–567, 2021. https://doi.org/10.34028/18/4/8

[14] E. A. Varela-Tapia, I. L. Acosta-Guzmán, C. I. Acosta-Varela, P. M. Marcillo-Sanchez, D. G. Patiño Pérez, and J. D. Tumbaco Bravo, “Intelligent predictive model of BMI in nutritionists’ patients using machine learning algorithms: Logistic regression and neural networks,” in *Proceedings of the 20th LACCEI International Multi-Conference for Engineering, Education and Technology, Boca, Raton*, 2022. https://doi.org/10.18687/LACCEI2022.1.1.791