Research Article
Journal of Intelligent Systems: Theory and Applications 7(2) (2024) 145-158
DOI: 10.38016/jista.1440879

Leveraging Machine Learning Methods for Predicting Employee Turnover Within the Framework of Human Resources Analytics

Zeynep Taner[1], Ouranıa Areta Hızıroğlu[2*], Kadir Hızıroğlu[3]

[1, 2, 3] Department of Management Information Systems, İzmir Bakırçay University, İzmir, Türkiye

zeynepyt78@gmail.com, ourania.areta@bakircay.edu.tr, kadir.hiziroglu@bakircay.edu.tr

Abstract

Employee turnover is a critical challenge for organizations, leading to significant costs and disruptions. This study aims to leverage Machine Learning (ML) techniques within the framework of Human Resources Analytics (HRA) to predict employee turnover effectively. The research evaluates and compares the performance of six widely used models: Decision Trees, Support Vector Machines (SVM), Logistic Regression, Random Forest, XGBoost, and Artificial Neural Networks. These models were implemented using the R programming language on an open-source dataset from IBM. The methodology involved data preprocessing, splitting into training, validation and testing sets, model training, and performance evaluation using metrics such as accuracy, sensitivity, specificity, precision, F1-score, and ROC-AUC. The results indicate that the Logistic Regression model outperformed the other models, achieving high accuracy and a good F1-score. The study concludes by emphasizing the importance of HRA and ML techniques in predicting and managing employee turnover, while discussing limitations such as class imbalance and the need for more rigorous performance evaluation. Future research directions include exploring alternative models, feature selection techniques, and addressing class imbalance.

Keywords: Human resources analytics, Employee turnover prediction, Machine learning models.

Makine Öğrenimi Yöntemlerini İnsan Kaynakları Analitiği Çerçevesinde İşten Ayrılma Tahminleri için Kullanma

Öz

Çalışan devir oranı, kuruluşlar için önemli bir zorluk oluşturmakta ve önemli maliyetlere ve aksaklıklara yol açmaktadır. Bu çalışma, insan kaynakları analitiği çerçevesinde makine öğrenimi tekniklerini etkin bir şekilde kullanarak çalışan devirini öngörmeyi amaçlamaktadır. Araştırma, altı yaygın olarak kullanılan modelin performansını değerlendirmekte ve karşılaştırmaktadır: Karar Ağaçları, Destek Vektör Makineleri, Lojistik Regresyon, Rastgele Orman, XGBoost ve Yapay Sinir Ağları. Bu modeller, IBM'den açık kaynaklı bir veri kümesi üzerinde R programlama dili kullanılarak uygulanmıştır. Çalışmanın metodlolojisi, veri ön işleme, eğitim, doğrulama ve test setlerine bölme, model eğitimi ve doğruluk, hassasiyet, özgünlük, hassasiyet, F1-skoru ve ROC-AUC gibi ölçümleri kullanarak performans değerlendirmeyi içermektedir Sonuçlar, Lojistik Regresyon modelinin diğer modellerden daha iyi bir performans sergilediğini, yüksek doğruluk ve iyi bir F1-skoru elde ettiğini göstermektedir. Çalışma kasapmında, çalışan devir oranını öngörmek ve yönetmek için insan kaynakları analitiği ve makine öğrenmesi tekniklerinin önemi vurgulanarak, sınıf dengesizliği gibi sınırlamaları ve daha güvenilir performans değerlendirmesi gereksinimine yönellik tartışmalara da yer vermektedir. Çalışmanın son kısmında, gelecek araştırma konuları çerçevesinde alternatif modellerin keşfedilmesi, özellik seçim teknikleri kullanılarak sonuçların değerlendirilmesi ve sınıf dengesizliğini gidermeye dönük hususlar ele alınmaktadır.

Anahtar Kelimeler: İnsan kaynakları analitiği, Çalışan devir hızı tahmini, Makine öğrenimi modelleri.

* Corresponding Author
E-Mail: ourania.areta@bakircay.edu.tr

Recieved : 21 Feb 2024
Revision : 6 May 2024
Accepted : 9 Jul 2024

1. Introduction

Human Resource Management (HRM) has undergone transformations to cope with ongoing technological advancements and dynamic business requirements. One such transformation is the adoption of HRA, which involves analyzing HR data on a larger scale to support evidence-based decision-making related to human performance, satisfaction, engagement, and ultimately, turnover. HRA has become increasingly important in understanding various processes that contribute to overall business success and competitive advantage (Van Vulpen, 2023).

The suitability of leveraging ML techniques for analyzing employee turnover within the HRA framework lies in their ability to identify complex patterns and relationships in large datasets, which may not be apparent through traditional statistical methods. MLmodels can learn from historical data and provide accurate predictions, enabling organizations to proactively identify employees at risk of turnover and take appropriate measures to retain valuable talent.

A critical aspect of HRA is the prediction of employee turnover, as high turnover rates can incur significant costs and impact productivity (Yavuz, 2016). Numerous studies have examined employee turnover and its reasons, highlighting the importance of retaining and rewarding the best employees (Aarons et al., 2009; Peryön, 2017, 2018; Randstad, 2022, 2023; Gallup, 2024). Effectively predicting employee turnover probabilities helps businesses improve workforce planning, reduce costs, and increase overall employee satisfaction (Moturi et al., 2023).

To address this challenge, the use of ML techniques within the framework of HRA has gained significant attention in recent years (Avrahami et al., 2022; Wijaya et al., 2021; Choi et al., 2021; Gao et al., 2019; Alsaadi et al., 2022). ML models can effectively predict employee turnover by learning from historical data and identifying patterns and relationships that may not be apparent through traditional statistical methods.

This study aims to evaluate and compare the performance of six widely used ML models - Random Forest, Logistic Regression, Artificial Neural Networks, Support Vector Machines, XGBoost and Decision Trees - in predicting employee turnover within the context of HRA. The choice of these models for predicting employee turnover in this study was based on their popularity, proven performance, and diversity of approaches (Breiman, 2001; Cortes & Vapnik, 1995; Friedman, 2001). These models represent a range of techniques, including tree-based methods, probabilistic models, and neural networks, capable of capturing complex relationships in the data (Demir & Çalık, 2021; Uzak, 2022). Some models, such as Decision Trees and Logistic Regression, offer interpretable results (Demir & Çalık, 2021), while others, like Random Forest and XGBoost, are known for their scalability and robustness to outliers and noise (Breiman, 2001; Friedman, 2001). The inclusion of simpler models allows for a comparison with more complex ones, assessing the trade-off between complexity and predictive performance (Liao, 2023). Moreover, these models have been successfully applied in previous studies on employee turnover prediction, providing evidence of their effectiveness in this context (Jain et al., 2020; Stachová et al., 2021).

Within the framework of its aim, the following research objectives were set:

- Evaluate and compare the performance of the trained models in predicting employee turnover using various metrics, including accuracy, sensitivity, specificity, precision, F1-score, and ROC-AUC (Receiver Operating Characteristic - Area Under the Curve).
- Identify the most effective model for predicting employee turnover and discuss the implications and limitations of the study.
- Provide recommendations for businesses and researchers to leverage ML techniques for effective employee turnover prediction and management.

By addressing these objectives, this study contributes to the existing body of knowledge in HRA and employee turnover prediction, while also providing practical insights for businesses to implement data-driven strategies for workforce management.

The study initially includes a literature review section, covering previous research around employee turnover prediction. This section identifies gaps in the existing literature and the contributions of this study. In the methodological part, elements such as the data set description, data preprocessing, model selection and training, model performance and evaluation are presented according to the research methodology. The following section presents the study's findings and the performance of the models, determining the best-performing model based on comparisons, offering also recommendations for usability of the tools for employee turnover prediction based on their suitability. Finally, the authors present the summary of the outcomes and future directions and recommendations in the conclusions section.

2. Literature Review

Human Resources Analytics is the process of collecting, analyzing, and making more effective decisions through insights derived from human resource data. HRA involves analyzing data from various sources within the enterprise using different methods to answer the right questions (Van Vulpen, 2023). Decision-making based on data enables organizations to gain a competitive advantage through more strategic and informed HRM (Shrivastava, Nagdev, and Rajesh, 2017).

In this study, while emphasizing the importance of data-driven decision-making in HRA, it also focuses on the analysis of employee turnover prediction within HRA applications. Employee turnover prediction analysis is a data analytics application that enables a business to predict employee departures in advance. This analysis has become a significant topic for businesses in recent years, providing important insights into workforce management and employee retention for employers and researchers (Wijaya et al., 2021; Ye et al., 2019; Liu & Liu, 2021; Schlechter et al., 2016; Putri & Rachmawati, 2022; Liao, 2023; Judrups et al., 2021; Chaudhary, 2022). Such studies help reduce workforce costs, increase employee satisfaction and productivity, and also aid in strategic human resources planning. In conducting employee turnover prediction analysis, the concept of "workforce turnover" comes into play, which refers to the number of employees leaving a business in a given period for various reasons, including voluntary and involuntary departures (Roche et al., 2015; Russell et al., 2017; Scanlan et al., 2013; Chisholm et al., 2011; Woltmann et al., 2008; Bogaert et al., 2019; Chapman et al., 2022; Roche et al., 2021; Poku et al., 2022; Onnis, 2017; Bardoel et al., 2020; Mayson & Bardoel, 2021; Healy & Oltedal, 2010; Russell et al., 2012; Belbin et al., 2012; Ashworth, 2006). High workforce turnover incurs significant costs, affecting training, recruitment, separation costs, and productivity (Yavuz, 2016). Therefore, having a model that can accurately predict the likelihood of employee departures is of great importance.

Empirical studies conducted within the scope of data analytics for predicting employee turnover have been presented in Table 1. The literature review table includes various research studies that have utilized different MLmodels and techniques for the purpose of predicting employee turnover. Each study is aimed at reducing the likelihood of employee turnover, targeting specific sectors and objectives. The studies vary in terms of features included, data sources, models and methods used, development tools, and evaluation metrics. Most research has two main objectives: "Increasing productivity" and "Reducing costs." For instance, a study using the K-nearest neighbours algorithm (Balcıoğlu & Artar, 2022) aims to increase efficiency, while a study on ML model selection for employee loss prediction in the telecommunications sector (Uzak, 2022) aims to reduce costs. Regarding data sources and size, some studies use open-source datasets, while others use in-house data, with data sizes ranging from small-scale studies to large datasets. The variables used include demographic (related to personal attributes of employees) and job-related variables (pertaining to employees' work experience and performance).

For development tools, programming languages such as Python or R have been used. Each study employed different ML models and methods, including Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Support Vector Machines (SVM), etc., in an attempt to predict employee turnover. According to the results of these studies, when examining the effectiveness of different MLmethods in predicting employee turnover, the studies "Prediction of Employee Turnover Probability with Machine Learning: K-Nearest Neighbors Algorithm (Balcıoğlu & Artar, 2022)" and "Employee Attrition Prediction (Yedida et al., 2018)" achieved high accuracy using the KNN algorithm. These results indicate KNN as an effective option for predicting employee turnover. Similarly, the studies "MLModel Selection for Employee Loss Prediction in the Telecommunications Sector (Uzak, 2022)" and "Predictive Analysis on the Example of Employee Turnover (Maisuradze, 2017)" have shown high accuracy with the Random Forest (RF) model, suggesting RF as a highly effective model for turnover prediction. The study "Prediction of Employee Turnover using ML(Shanthakumara et al., 2022)" used Artificial Neural Networks (ANN), showing ANN as a viable alternative for turnover prediction. The "Employee Attrition Prediction" study utilized Logistic Regression, indicating its effectiveness in turnover prediction. Metrics such as accuracy, precision, recall, and F1-score were commonly evaluated.

Additionally, some studies have utilized specific metrics like AUC (Area Under Curve). Data attributes in these studies include various factors such as demographic information (age, gender, education) and job-related information (position, salary, job satisfaction). These attributes have been used to predict the likelihood of employees leaving their jobs.

Despite the existing research, several limitations and gaps warrant further investigation:

- Limited comparative studies: While individual studies have explored the performance of specific ML models, there is a lack of comprehensive comparative analyses evaluating the effectiveness of different models on the same dataset.
- Inconsistent results: The existing literature presents inconsistent results regarding the most effective ML model for employee turnover prediction, suggesting that the choice of model may be context-dependent or influenced by factors such as data quality, preprocessing techniques, and feature selection.
- Lack of generalizability: Many studies have focused on specific industries or contexts, which may limit the generalizability of their findings to other organizational settings.
- Limited discussion of practical implications: While the studies demonstrate the potential of ML techniques for employee turnover prediction, there is often a lack of discussion regarding the practical implications and implementation challenges for businesses.
- Absence of rigorous model evaluation: Some studies have relied primarily on accuracy as the sole performance metric, overlooking the importance of other relevant metrics such as accuracy, sensitivity,

specificity, precision, F1-score, and ROC-AUC, which can provide a more comprehensive understanding of model performance.

This study aims to address these limitations by conducting a comprehensive comparative analysis of six widely used ML models (Random Forest, Logistic Regression, Artificial Neural Networks, Support Vector Machines, XGBoost and Decision Tree) on a publicly available dataset, evaluating their performance using multiple metrics, and discussing the practical implications and future research directions.

Table 1. Table of Studies Conducted in the Field of Human Resources Analytics

Title and Year	Purpose/ Objective	Attributes	Data Source/ Size	Development Tool	Model-Method and Techniques	Metric

Predicting Employee Attrition Using Machine Learning: A K-Nearest Neighbors Algorithm Approach (Balcıoğlu & Artar, 2022)	Increase efficiency	Demographic data; Age, Marital Status, Education level; Job-related data; Working hours, Position, Job satisfaction, Salary, Work arrangement	Open source - 1205	MATLAB R2020b	KNN (k=4) - %93
					KNN(K=1)
					KNN(K=6)
					KNN(K=8)
						Accuracy
						Precision
						Recall
						F1-Score

MLModel Selection for Predicting Employee Turnover in the Telecommunications Industry (Uzak, 2022)	Reduce costs	Demographic data; ID, Age, Gender, Marital status, Location, Child number, Military Service, School Type; Job-related data; Title, Function, Reason for Leaving, Status/Objective, Active/Inactive	Company data – 16655	Python	RF - %92,2 Logistic Regression - KNN - DVM – CART - Gradient Boosting Machine- YSA - XGBoost	Accuracy Precision Sensitivity F1-Score EAKA
Prediction of Employee Turnover using Machine Learning (Shanthakumara et al., 2022)	Increase efficiency	Demographic data; Age, Sex, Education; Job-related data; Position, Department, Salary, Overtime, Average Monthly Hours, Tenure, Number of Projects, Satisfaction, Work Accident	N/A – 15400	R	RF - %93 Naive Bayes - Logistic Regression	Accuracy
Employee Attrition Prediction (Yedida et al., 2018)	Increase efficiency	Job-related data; Average Monthly Hours, Number of Projects, Promotion in the Last Five Years, Seniority	Open source – 14999	Python	KNN - %94,32 Naive Bayes - Logistic Regression - MLP Classifier	AUC Accuracy F1-Score
Predicting the Perceived Employee Tendency of Leaving an Organization Using SVM and Naive Bayes Techniques (Emmanuel-Okereke & Anigbogu, 2022)	Reduce costs	Demographic data; Gender, Experience, Seniority, Education; Job-related data: Date of Entry, Job Safety, Working Hours, Job Satisfaction, Status/Objective	Survey - 514	Python	Naive Bayes - %100 DVM – RF – Decision Tree	Precision Recall F1-Score
Employee Turnover Prediction Using MLBased Methods (Kışaoğlu, 2014)	Reduce costs - Increase efficiency	Demographic data; Age, Race Job-related data; Performance, Job Satisfaction Survey Results, Job Transition/Change Networks, Status/Objective - "Will Leave", "Will Not Leave"	Open source - 25000	WEKA	DVM - Karar Ağacı - Naïve Bayes	Accuracy Precision Recall F1-Score
Employee Turnover Probability Prediction (Barın, 2022)	Reduce costs	Demographic data; Age, Seniority, Gender, Marital Status, Number of Children, Education; Job-related data:	Company data – 3282	R	Hierarchical Model - 69.4% Naive Bayes - RF	ROC-AUC

		Performance Score, Appreciation Score, Salary, Salary Increase, Promotion, First Year Information, Foreign Language; Status/Objective - Employed/Left				
Optimization of employee turnover through predictive analysis (Stachová, Baroková & Stacho, 2021)	Reduce costs	Demographic data; Age, Sex, Education, Marital status, Seniority; Job-related data; Business Travel, Position, Department, Commute Distance, Work-Life Balance, Hourly Wage, Monthly Income, Overtime, Working Hours, Salary, Salary Increase, Promotion	Open source – 1470	Python	RF -%87 Logistic Regression – Decision Tree - K-Means	Accuracy
Employee Churn Prediction using Logistic Regression and Support Vector Machine (Maharjan, 2021)	Reduce costs	Demographic data; Age, Education, Gender, Seniority, Marital Status; Job-related data: Position, Monthly Income, Job Satisfaction, Overtime, Performance, Training Duration (last year), Work-Life Balance, Work Experience, etc., Status/Objective, Employed/Left	Open source – 23436	Python	DVM -%84 Logistic Regression	Precision Recall F1-Score ROC-AUC Accuracy
Explaining and predicting employees' attrition: a MLapproach (Jain, Jain & Pamula, 2020)	Reduce costs	Job-related data; Satisfaction Level, Performance, Number of Projects, Average Monthly Hours, Work Accident, Promotion - Last 5 Years, Salary, Domain, Target Variable, Department Names (Sales, HR, Technical, Support, etc.)	Open source - 14.000+	Python	RF -%99 YSA - Decision Tree - Naive Bayes – Logistic Regression – DVM	Precision F1-Score Recall
Predictive Alaysis on the Example of Employee Turnover (Maisuradze, 2017)	Reduce costs	Demographic data; Age, Gender, Education, Seniority, Marital Status; Job-related data: Overtime, Job Satisfaction, Monthly Income, Performance, Distance from Home, Promotion, Work-Life Balance, Salary Increase, Position, Department	Open source - 1471	Python	RF- 98.62% DVM – YSA	ROC-AUC Accuracy
Employee turnover prediction and retention policies design: a case study (Ribes, Touahri & Perthame, 2017)	Reduce costs	Demographic data; Age, Experience, Gender, Ethnic Background, Education; Job-related data: Performance, Role Salary, Working Conditions, Job Satisfaction, Burnout, Seniority, Status/Objective,	Open source – 1000	R	Linear Discriminant-%75 DVM – KNN – RF -Naïve Bayes	Accuracy ROC-AUC

		Employed/Left				
Leveraging MLMethods for Predicting Employee Turnover Within the Framework of Human Resources Analytics (Current/Our Study)	Reduce costs	Demographic data; Age, Education, Gender, Seniority, Marital Status; Job-related data: Position, Monthly Income, Job Satisfaction, Overtime, Performance, Training Duration (last year), Work-Life Balance, Work Experience, etc., Status/Objective, Employed/Left	Open source - 1470	R	RF - YSA - Decision Tree - XGBoost – Logistic Regression – DVM	Precision Recall F1-Score ROC-AUC Accuracy

The following section presents the methodology that was employed so that the authors could meet the objectives of this study.

3. Research Methodology

The purpose of the research is to determine the most suitable and effective model for predicting employee turnover and to evaluate the performance of this model. The following sections describe the several stages that the authors undertook to meet the aim and objectives of this paper.

3.1 Data Source and Preprocessing

The dataset in question is from Kaggle platform, created by IBM data scientists and titled "IBM HR Analytics Employee Attrition & Performance" (Pavansubhash, 2016). The dataset comprises a total of 1470 employee records, (1233 employees and 237 leavers) with 35 features, including 34 independent variables and 1 dependent variable (Attrition). The independent variables encompass demographic information, job-related data, and other relevant factors, while the dependent variable is a binary indicator of employee attrition.

Data preprocessing involved removing variables with low analytical value, such as "EmployeeNumber," "EmployeeCount," "Over18," and "StandardHours." The remaining variables were then normalized for scaling to enable analysis that is more meaningful. Table 1 represents the preprocessed dataset and includes the types of variables in the dataset and their descriptions.

The preprocessed dataset was split then into training (60%), validation and testing (20% each) sets for models' development and evaluation where partitioning was carried out for each model separately.

Table 2. Preprocessed Data set

Order	Variable	Definition	Variable Type
	Demographic – Independent variable		
1	Age	Employee's Age	Numeric
2	Marital status	Marital Status (Single, Married, Divorced)	Categorical
3	Gender	Gender	Categorical
4	Education	Education Level (1: Below University, 2: University, 3: Bachelor's, 4: Master's, 5: Doctorate)	Numeric
5	Travel Status	Business Travel Frequency (No Travel, Rare Travel, Frequent Travel)	Categorical
	Job-related - Independent Variable		
1	Daily Wage	The amount of money a company is obligated to pay an employee for a day's work.	Numeric
2	Department	Department (Research and Development, Sales, Human Resources)	Categorical
3	Commute Distance	Distance between home and company	Numeric
4	Field of Study	Field of Education (Science, Medicine, Human Resources, Technical Degree, Marketing, Other)	Categorical
5	Environmental Satisfaction	Environmental Satisfaction Score (1: Low, 2: Medium, 3: High, 4: Very High)	Numeric
6	Engagement Level	Level of Job Involvement (1: Low, 2: Medium, 3: High, 4: Very High)	Numeric
7	Work-Family	Job Level (1 - 5)	Numeric
8	Role	Job Role (Sales Manager, Human Resources Manager, etc.)	Categorical
9	Job Satisfaction	Job Satisfaction (Low, Medium, High, Very High)	Numeric
10	Monthly Income	Employee's Monthly Income	Numeric
11	Salary Raise	Percentage of Salary Increase	Numeric
12	Number of Companies	Total number of companies the employee has worked for before	Numeric

	Worked At		
13	Job Satisfaction	Job Satisfaction (Low, Medium, High, Very High)	Numeric
14	Overtime	Employee's Overtime Status (Yes, No)	Categorical
15	Salary Raise %	Percentage of Salary Increase	Numeric
16	Performance Rating	Level of Performance Appraisal (Low, Good, Excellent, Outstanding)	Numeric
17	Communication Satisfaction	Level of Relationship Satisfaction (Low, Medium, High, Very High)	Numeric
18	Working Hours	Standard Working Hours	Numeric
19	Stock Option Level	Employee's Stock Option Level (0 - 3)	Numeric
20	Work Experience	Total Years of Working	Numeric
21	Training Duration (Last Year)	Training Duration Last Year	Numeric
22	Work-Life Balance	Work-Life Balance Level (1: Poor, 2: Good, 3: Better, 4: Best)	Numeric
23	Seniority	Years at the Company	Numeric
24	Tenure in Role	Years in Current Role	Numeric
25	Years with Current Manager	Years with Current Manager	Numeric
	Dependent Variable		
1	Attrition Status	Employee Attrition (Yes, No)	Categorical

3.2. Model Selection and Training

Six widely used ML models were selected for this study, namely as Random Forest, Logistic Regression, Artificial Neural Networks, Support Vector Machines, XGBoost and Decision Tree:

- Random Forest, an ensemble learning method, is known for its robustness, ability to handle large datasets with many features, and its effectiveness in both classification and regression tasks (Breiman, 2001).
- Logistic Regression, a classical statistical method, is often used when the dependent variable is categorical and provides interpretable results (Demir & Çalık, 2021).
- Artificial Neural Networks, inspired by the structure and function of biological neural networks, are capable of learning complex non-linear relationships between input features and the target variable (Demir & Çalık, 2021; Uzak, 2022).
- Support Vector Machines, a non-probabilistic binary linear classifier, are known for their ability to handle high-dimensional data and their effectiveness in both linear and non-linear classification tasks (Cortes & Vapnik, 1995).
- XGBoost, an ensemble learning method that combines multiple weak learners (decision trees) to create a strong learner, is known for its ability to handle complex interactions among features and its effectiveness in both classification and regression tasks (Friedman, 2001).
- Decision Trees, a simple yet powerful supervised learning algorithm, is known for their interpretability, ability to handle both categorical and numerical data, and effectiveness in capturing non-linear relationships between features and the target variable. They repeatedly divide the feature space into subsets based on the most informative features, creating a tree-like model that can be easily visualized and understood (Rokach & Maimon, 2005).

The researchers implemented the models using the R programming language. The installation, training, and performance evaluation of each model was carried out on the original dataset. The training process involved fitting each model to the training dataset, with 5-fold cross-validation to ensure the robustness and generalizability of the results. Cross-validation helps to assess the model's performance on different subsets of the data, reducing the risk of overfitting and providing a more reliable estimate of the model's performance on unseen data.

Training involved using the specified models with utilized to compare the models based on specific metrics (see part 3.3). The training set is the data used by the ML algorithm during its learning process. This dataset includes the input and output values for each example. The learning algorithm uses the data in the training set to learn the correct outputs for the inputs. For example, in text classification studies, the content of the input texts and the output categories are included in the training set. In contrast, the test set is used to validate and assess the performance of the trained model. The test dataset comprises data that are distinct and previously unseen in comparison to the training set. The model, trained during the learning process, makes predictions for the inputs in the test set. To evaluate the model's accuracy and performance, these predictions are compared with the actual outputs of the test data (Kutlugün et al., 2017).

More specifically, and with regards to each one of the ML models, the setup and evaluation took place as followed:

- Random Forest: A model containing 500 trees was established with the RandomForest package, and the classification performance of the model was examined in detail.

- Logistic Regression: Within the framework of the generalized linear model, a logistic regression model was created using the glm() function, and probability predictions were made.
- Artificial Neural Networks: A 10-neuron neural network model was established with the nnet package, and the classification predictions of the model were evaluated.
- Support Vector Machines (SVM): On the data divided into training and test sets, the SVM model was established by determining the optimal gamma and cost values through the e1071 package, and the classification performance was evaluated.
- XGBoost: The xgboost package was used, and various hyperparameters were adjusted with the train() function. These parameters include the maximum depth of trees (max_depth), learning rate (eta), and editing parameters (gamma). Additionally, optimal parameter combinations were determined using a comprehensive grid search method to further optimize the model.
- Decision Trees: A model to predict attrition was created using the Tree library, trained, and visualized by adding information to its branches. The accuracy of the model was evaluated on the test dataset with confusionMatrix.

3.3. Performance and Evaluation

The trained models were evaluated on the test dataset using various performance metrics, including accuracy, sensitivity, specificity, precision, F1-score, and ROC-AUC. These metrics provide a comprehensive assessment of the models' predictive capabilities, considering factors such as correct classifications, false positives, and false negatives. For the calculation of each of the aforementioned metrics, the following need to be defined:

- True Positives (TP): The number of instances that are actually positive and correctly predicted as positive by the model.
- True Negatives (TN): The number of instances that are actually negative and correctly predicted as negative by the model.
- False Positives (FP): The number of instances that are actually negative but incorrectly predicted as positive by the model.
- False Negatives (FN): The number of instances that are actually positive but incorrectly predicted as negative by the model.

Then, the metrics can be defined and calculated as followed:

- Accuracy: The proportion of correctly classified instances out of the total instances.
  Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Sensitivity (Recall or True Positive Rate): The proportion of true positive predictions among all actual positive instances.
  Sensitivity = TP / (TP + FN)
- Specificity: The proportion of true negative predictions among all actual negative instances.
  Specificity = TN / (TN + FP)
- Precision: The proportion of true positive predictions among all positive predictions.
  Precision = TP / (TP + FP)
- F1-score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.
  F1-score = 2 * (Precision * Recall) / (Precision + Recall)
  The F1-score ranges from 0 to 1, with 1 being the best value and 0 being the worst.
- ROC: An aggregate measure of the model's performance, considering both its ability to identify positive instances (employee retention) and negative instances (employee turnover). It is calculated as the sum of the True Positive Rate (TPR) and the True Negative Rate (TNR) divided by 2. TPR measures the proportion of actual positive instances that are correctly identified, while TNR measures the proportion of actual negative instances that are correctly identified. The AUC (Area Under Curve) value measures the probability of the model correctly classifying a randomly selected positive example into a randomly selected negative example. The closer the AUC value is to 1, the better the model performs.

Within this framework, confusion matrices for each model have been included and explained in detail, providing insights into the models' performance in terms of true positives, true negatives, false positives, and false negatives. The confusion matrix is used to understand the model's performance more deeply and to examine the classification results in more detail. It is very valuable for determining which classes the model predicts better or worse, and which classes are associated with false positives or false negatives.

With the methodology clearly defined, next section presents the results obtained from the MLmodel testing and evaluation. The results and discussion section will analyze the performance of the selected models and interpret the findings.

4. Results and Discussion

4.1 Confusion Matrices

The confusion matrices provide a detailed breakdown of the models' performance in terms of TP, TN, FP, and FN. They help in understanding how well each model classified the instances into the correct categories. In the case of employee churn problem of this study, the positive class represents the employees who have not left the company whereas the negative class is the ones who left the company. Therefore, in the confusion matrices that will be provided below, indications with “yes” represent the negative classes (employee turnover) and with “no” refer to the positive classes (employee retention).

The confusion matrices for each of the ML models are as follows:

a. Random Forest:

- TN: The model correctly predicted 9 instances as “Yes” (employees who left).
- TP: The model correctly predicted 244 instances as “No” (employees who did not leave).
- FP: The model incorrectly predicted 2 instances as “Yes” when it was actually “No”.
- FN: The model incorrectly predicted 39 instances as “No” when they were actually “Yes”.

The Random Forest model has a high number of True Positives (244), correctly identifying employees who have not left the company. However, it has a relatively low number of True Negatives (9), indicating that it correctly identifies only a small proportion of employees who have left. The model has a very low number of False Positives (2), which means it rarely misclassifies employees who have not left as having left. On the other hand, the model has a higher number of False Negatives (39), incorrectly classifying employees who have left as still being with the company. This suggests that the model may have difficulty capturing all the instances of employee turnover.

Table 3. Confusion Matrix for Random Forest

	Predicted “No”	Predicted “Yes”
Actual “No”	244	2
Actual “Yes”	39	9

b. Logistic Regression:

- TN: The model correctly predicted 27 instances as “Yes” (employees who left).
- TP: The model correctly predicted 237 instances as “No” (employees who did not leave).
- FP: The model incorrectly predicted 21 instances as “Yes” when they were actually “No”.
- FN: The model incorrectly predicted 9 instances as “No” when they were actually “Yes”.

The Logistic Regression model has a good balance between True Positives (237) and True Negatives (27), indicating decent overall accuracy. It has a relatively low number of False Positives (21) and False Negatives (9). This model seems to have a balanced performance in identifying both positive and negative instances.

Table 4. Confusion Matrix for Logistic Regression

	Predicted “No”	Predicted “Yes”
Actual “No”	237	21
Actual “Yes”	9	27

c. Artificial Neural Networks (ANN):

- TN: The model correctly predicted 26 instances as “Yes” (employees who left).
- TP: The model correctly predicted 235 instances as “No” (employees who did not leave).
- FP: The model incorrectly predicted 22 instances as “Yes” when they were actually “No”.
- FN: The model incorrectly predicted 11 instances as “No” when they were actually “Yes”.

The Artificial Neural Networks model shows a high number of True Positives (235), accurately identifying employees who have not left. It has a relatively low number of False Positives (22), minimizing the misclassification of employees who have left as still being with the company. The model has a moderate number of True Negatives (26) and False Negatives (11), demonstrating a reasonable ability to identify employees who have left.

Table 5. Confusion Matrix for ANN

	Predicted “No”	Predicted “Yes”
Actual “No”	235	22
Actual “Yes”	11	26

d. Support Vector Machines (SVM):

- TN: The model correctly predicted 4 instances as “Yes” (employees who left).
- TP: The model correctly predicted 246 instances as “No” (employees who did not leave).
- FP: The model incorrectly predicted 44 instances as “Yes” when they were actually “No”.
- FN: The model incorrectly predicted 0 instance as “No” when they were actually “Yes”.

The SVM model has a high number of True Positives (246), accurately identifying employees who have not left. However, it also has a high number of False Positives (44), suggesting that it often misclassifies employees who have left as still being with the company. The model has a low number of True Negatives (4) and False Negatives (8), indicating poor performance in correctly identifying employees who have left.

Table 6. Confusion Matrix for SVM

	Predicted “No”	Predicted “Yes”
Actual “No”	246	44
Actual “Yes”	0	4

e. XGBoost:

- TN: The model correctly predicted 12 instances as “Yes” (employees who left).
- TP: The model correctly predicted 244 instances as “No” (employees who did not leave).
- FP: The model incorrectly predicted 36 instance as “Yes” when it was actually “No”.
- FN: The model incorrectly predicted 2 instances as “No” when they were actually “Yes”.

The XGBOOST model has a high number of True Positives (244), correctly identifying employees who have not left. However, it also has a relatively high number of False Positives (36), indicating a tendency to misclassify employees who have left as still being with the company. The model has a low number of True Negatives (12) and False Negatives (2), suggesting difficulty in accurately identifying employees who have left.

Table 7. Confusion Matrix for XGBoost

	Predicted “No”	Predicted “Yes”
Actual “No”	244	36
Actual “Yes”	2	12

f. Decision Tree:

- TN: The model correctly predicted 16 instances as “Yes” (employees who left).
- TP: The model correctly predicted 237 instances as “No” (employees who did not leave).
- FP: The model incorrectly predicted 32 instance as “Yes” when it was actually “No”.
- FN: The model incorrectly predicted 9 instances as “No” when they were actually “Yes”.

The Decision Tree model has a relatively balanced performance. It has a good number of True Positives (237), correctly identifying employees who have not left. The number of False Positives (32) is moderate, showing some misclassification of employees who have left as still being with the company. The True Negatives (10) and False Negatives (9) are relatively balanced, indicating a fair ability to identify employees who have left.

Table 8. Confusion Matrix for Decision Tree

	Predicted “No”	Predicted “Yes”
Actual “No”	237	32
Actual “Yes”	9	16

From the confusion matrices, we can see that the Artificial Neural Networks and Logistic Regression models exhibit a more balanced performance in correctly identifying both employees who have not left and those who have left. The Random Forest model performs well in identifying employees who have not left but may struggle to capture all instances of employee turnover. The XGBoost and Decision Tree models show a tendency to misclassify employees who have left as still being with the company, while the SVM model exhibits a strong bias towards predicting employees as staying with the company.

These confusion matrices provide insights into the models' performance and can help in identifying areas for improvement, such as addressing class imbalance or tuning the models to better identify employee turnovers and retetntions.

4.2 ML Model Testing Results

The results from the ML model testing within the framework of the accuracy, sensitivity, specificity, precision, F1 Score and ROC-AUC metrics are presented in the Table 10.

Based on the results, the Logistic Regression model outperformed the other models in terms of accuracy (89.80%), sensitivity (96.34%), and F1 Score (0.5614). It also achieved a high ROC-AUC value of 0.902, indicating its strong overall performance in distinguishing between the positive and negative classes.

The Artificial Neural Networks model also demonstrated good performance, with an accuracy of 88.78%, sensitivity of 95.53%, and the highest F1 Score among all models (0.6364). However, its ROC-AUC value (0.784) was lower compared to the Logistic Regression and Random Forest models.

The Random Forest model achieved an accuracy of 86.05% and a high ROC-AUC value of 0.900. It exhibited balanced performance in terms of sensitivity (86.22%) and specificity (81.82%). However, its F1

Table 9. Results for all ML models

Model	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1 Score**	**ROC-AUC**
Random Forest	0.8605	0.8622	0.8182	0.9919	0.3067	0.900
Logistic Regression	0.898	0.9634	0.5625	0.9186	0.5614	0.902
Artificial Neural Networks	0.8878	0.9553	0.5417	0.9144	0.6364	0.784
SVM	0.8503	1	0.08333	0.8483	0.1538	0.524
XGBoost	0.8707	0.9919	0.25	0.8714	0.3871	0.851
Decision Tree	0.8605	0.9634	0.3333	0.8810	0.439	0.684

Score (0.3067) was relatively lower compared to the Logistic Regression and Artificial Neural Networks models.

The XGBoost model showed an accuracy of 87.07% and a high sensitivity of 99.19%, indicating its effectiveness in correctly identifying positive instances. However, its specificity (25%) and F1 Score (0.3871) were lower compared to the other models.

The Decision Tree model achieved an accuracy of 86.05%, similar to the Random Forest model. It demonstrated high sensitivity (96.34%) but relatively lower specificity (33.33%) and F1 Score (0.439).

The SVM model exhibited the lowest accuracy (85.03%) among all models. While it achieved perfect sensitivity (100%), its specificity (8.33%) and F1 Score (0.1538) were the lowest, indicating a high rate of false positives.

The ROC-AUC values provide an aggregate measure of each model's performance, considering both its ability to identify positive instances (employee retention) and negative instances (employee turnover). The Logistic Regression and Random Forest models achieved the highest ROC-AUC values (0.902 and 0.900, respectively), indicating their superior overall performance compared to the other models.

It is important to note that the presence of class imbalance in the dataset can influence the models' performance, particularly in terms of sensitivity and F1-score for the minority class (employee turnover). Addressing class imbalance through techniques such as oversampling, undersampling, or using class weights can help improve the models' ability to correctly identify instances of employee turnover.

Also, one limitation of this study is the reliance on a single dataset. While the "IBM HR Analytics Employee Attrition & Performance" dataset provides a diverse set of employee records, the results' generalizability to other organizations or industries may be limited. Future research could validate the findings using datasets from different contexts or conduct multi-organizational studies to assess the models' performance across various settings.

From a practical standpoint, the findings of this study have several implications for businesses aiming to leverage ML techniques for employee turnover prediction as the study presents in the following section.

4.3 ML Tools Suitable for Employee Turnover Prediction

Based on the performance metrics evaluated in this study, the following machine learning tools are considered suitable for employee turnover prediction:

The Logistic Regression model demonstrated the highest accuracy, sensitivity, and F1 Score, along with a high ROC-AUC value. It is a simple and interpretable model that can provide insights into the factors contributing to employee turnover. Logistic Regression is particularly suitable when the relationship between the predictors and the target variable is linear.

The Artificial Neural Networks model achieved the second-highest accuracy and the highest F1 Score. It is capable of capturing complex non-linear relationships between the predictors and the target variable. Artificial Neural Networks can be effective when dealing with large datasets and when the underlying relationships are not well understood.

The Random Forest model exhibited balanced performance in terms of sensitivity and specificity, along with a high ROC-AUC value. It is an ensemble learning method that combines multiple decision trees, making it robust to outliers and noise. Random Forest can handle both categorical and numerical predictors and can provide feature importance rankings.

The XGBoost model demonstrated high sensitivity and a relatively high ROC-AUC value. It is an optimized implementation of gradient boosting that can handle complex interactions among predictors. XGBoost is known for its excellent predictive performance and its ability to handle missing values.

The SVM and Decision Tree models had lower overall performance compared to the above models, but they may still be considered in certain scenarios. SVMs can be effective when dealing with high-dimensional data, while Decision Trees offer interpretability and can handle both categorical and numerical predictors.

When selecting the most suitable ML tool for predicting employee turnover, it is essential to consider factors such as the size and complexity of the dataset, the interpretability requirements, the presence of non-linear relationships, and the computational resources available. It is also recommended to experiment with multiple tools and compare their performance using appropriate evaluation metrics to determine the best approach for the specific dataset and problem at hand.

The results and discussion section has provided valuable insights into the performance of various ML models for predicting employee turnover. In the following conclusion, the authors will summarize the key findings, discuss the implications of our study, and outline potential avenues for future research in this domain.

6. Conclusion

This study aimed to leverage ML techniques within the framework of HRA to predict employee turnover effectively. By evaluating and comparing the performance of Random Forest, Logistic Regression, Artificial Neural Networks, Support Vector Machines, XGBoost and Decision Tree models on the "IBM HR Analytics Employee Attrition & Performance" dataset, the study contributes to the existing body of knowledge in HRA and employee turnover prediction.

The findings suggest that the Logistic Regression model can be an effective tool in human resources analytics for turnover prediction. However, the choice of model should be based on the specific use case, considering the strengths and weaknesses of each model. Organizations should evaluate their requirements and prioritize the relevant performance metrics when selecting a model for implementation.

The findings of this study have practical implications for businesses seeking to implement data-driven strategies for workforce management. By leveraging ML techniques, organizations can proactively identify employees at risk of turnover and take appropriate measures to retain valuable talent. However, Organizations implementing ML models for employee turnover prediction should also consider the ethical implications and potential biases associated with these approaches. Ensuring fairness, transparency, and privacy in the use of employee data is crucial to maintain trust and comply with legal and ethical standards.

Future research directions include exploring alternative ML models, investigating the impact of feature selection techniques, and addressing class imbalance to further improve the predictive performance of the models. Additionally, validating the findings using datasets from different contexts or conducting multi-organizational studies can enhance the generalizability of the results.

In conclusion, this study demonstrates the potential of ML techniques within the HRA framework for predicting employee turnover. By continuously refining and improving these models, businesses can make data-driven decisions to optimize their workforce planning, reduce turnover costs, and enhance overall employee satisfaction and retention.

References

Aarons, G., Sawitzky, A., 2006. Organizational climate partially mediates the effect of culture on work attitudes and staff turnover in mental health services. Administration and Policy in Mental Health and Mental Health Services Research, 33(3), 289-301. https://doi.org/10.1007/s10488-006-0039-1

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., Childe, S. J., 2016. How to improve firm performance using big data analytics capability and business strategy alignment? International Journal of Production Economics, 182, 113–131. https://doi.org/10.1016/j.ijpe.2016.08.018

Alan, A., 2020. Makine Öğrenmesi Sınıflandırma Yöntemlerinde Performans Metrikleri ile Test Tekniklerinin Farklı Veri Setleri Üzerinde Değerlendirilmesi (Yüksek Lisans Tezi). Fırat Üniversitesi, Fen Bilimleri Enstitüsü, s.19

Alsaadi, E., Khlebus, S., Alabaichi, A., 2022. Identification of Human Resoıurce Analytics using MLalgorithms. Telkomnika (Telecommunication Computing Electronics and Control), 20(5), 1004. https://doi.org/10.12928/telkomnika.v20i5.21818

Ashworth, M., 2006. Preserving knowledge legacies: workforce aging, turnover and human resource issues in the us electric power industry. The International Journal of Human Resource Management, 17(9), 1659-1688. https://doi.org/10.1080/09585190600878600

Avrahami, D., Pessach, D., Singer, G., Ben-Gal, H. C., 2022. A human resources analytics and machine-learning examination of turnover: implications for theory and practice. International Journal of Manpower, 43(6), 1405-1424. https://doi.org/10.1108/ijm-12-2020-0548

Bahadır, M. B., Bayrak, A. T., Yücetürk, G., Ergun, P., 2021. A Comparative Study for Employee Churn, Prediction, Researchgate, 1-4.

Balcıoğlu, Y. S., Artar, M., 2022. Çalışanların İşten Ayrılma Olasılığının Makine Öğrenmesi İle Tahmini: K-En Yakın Komşu Algoritması İle. Güncel İşletme, Yönetim ve Muhasebe Çalışmaları, 29-35. https://www.researchgate.net/publication/359362785

Bardoel, A., Russell, G., Advocat, J., Mayson, S., Kay, M., 2020. Turnover among australian general practitioners: a longitudinal gender analysis. Human Resources for Health, 18(1). https://doi.org/10.1186/s12960-020-00525-4

Barın, H. D., 2022. Employee Turnover Probability Prediction, A thesis submitted to the Graduate School of Engineering and Science of Bilkent University for the degree of Master of Science in Industrial Engineering, 1-75.

Belbin, C., Erwee, R., Wiesner, R., 2012. Employee perceptions of workforce retention strategies in a health system. Journal of Management & Organization, 18(5), 742-760. https://doi.org/10.5172/jmo.2012.18.5.742

Bogaert, K., Leider, J., Castrucci, B., Sellers, K., Whang, C., 2019. Considering leaving, but deciding to stay: a longitudinal analysis of intent to leave in public health. Journal of Public Health Management and Practice, 25(2), S78-S86. https://doi.org/10.1097/phh.0000000000000928

Breiman, L., 2001. Rastgele Ormans. Machine learning, 45(1), 5-32.

Catani F, Lagomarsino D, Segoni S, Tofani V., 2013. Landslide susceptibility estimation by Rastgele Ormans technique: sensitivity and scaling issues. Nat Hazards Earth Syst Sci, 13:2815–2831, doi:10.5194/nhess-13-2815-2013.

Chapman, G., Nasirov, S., Özbilgin, M., 2022. Workforce diversity, diversity charters and collective turnover: long-term commitment pays. British Journal of Management, 34(3), 1340-1359. https://doi.org/10.1111/1467-8551.12644

Chaudhary, M., 2022. Rationale of employee turnover: an analysis of banking sector in nepal. International Research Journal of MMC, 3(2), 18-25. https://doi.org/10.3126/irjmmc.v3i2.46291

Chisholm, M., Russell, D., Humphreys, J., 2011. Measuring rural allied health workforce turnover and retention: what are the patterns, determinants and costs?. Australian Journal of Rural Health, 19(2), 81-88. https://doi.org/10.1111/j.1440-1584.2011.01188.x

Choi, J., Ko, I., Kim, J., Jeon, Y., Han, S., 2021. MLframework for multi-level classification of company revenue. Ieee Access, 9, 96739-96750. https://doi.org/10.1109/access.2021.3088874

Demir, K., Çalık, E., 2021. İnsan Kaynakları Analitiği: Modelleme ve Örnek Uygulamalarla. 2. Baskı, Nobel Bilimsel Yayıncılık.

Emmanuel-Okereke, I. L., Anigbogu, S. O., 2022. Predicting the Perceived Employee Tendency of Leaving an

Organization Using SVM and Naive Bayes Techniques. Open Access, 1-15.

Erkal, H., Keçecioğlu, T., Yılmaz, M. K., 2014. Gelecek 10 Yıl İçerisinde İnsan Kaynaklarının Yüzleşeceği Zorluklar. EUL Journal of Social Sciences, V(II), LAÜ Sosyal Bilimler Dergisi, Aralık, 32-63.

Gallup., 2023. State of the Global Workplace Report - Gallup. Gallup.com. Retrieved February 21, 2024, from https://www.gallup.com/workplace/349484/state-of-the-global-workplace.aspx#ite-506924

Gao, X., Wen, J., Zhang, C., 2019. An improved random forest algorithm for predicting employee turnover. Mathematical Problems in Engineering, 2019, 1-12. https://doi.org/10.1155/2019/4140707

Hatch-Maillette, M., Harwick, R., Baer, J., Masters, T., Cloud, K., Peavy, M., Wells, E., 2019. Counselor turnover in substance use disorder treatment research: observations from one multisite trial. Substance Abuse, 40(2), 214-220. https://doi.org/10.1080/08897077.2019.1572051

Healy, K., Oltedal, S., 2010. An institutional comparison of child protection systems in australia and norway focused on workforce retention. Journal of Social Policy, 39(2), 255-274. https://doi.org/10.1017/s004727940999047x

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset?resource=download

Jain, P. K., Jain, M., Pamula, R., 2020. Explaining and Predicting Employees' Attrition: A MLApproach. Research Article, 1-11.

Judrups, J., Cinks, R., Birzniece, I., Andersone, I., 2021. MLbased solution for predicting voluntary employee turnover in organization.. https://doi.org/10.22616/erdev.2021.20.tf296

Karcı, Z., 2017. Lojistik Regresyon Modeli ile Elde Edilen Tahminlerin ROC Eğrisi Yardımıyla Değerlendirilmesi: Türkiye'de Hanehalkı Yoksulluğu Üzerine Bir Araştırma (Yüksek Lisans Tezi). T.C. Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü, Ekonometri Anabilim Dalı, Isparta, 46.

Kışaoğlu, Z. Ö., 2014. Employee Turnover Prediction Using MLBased Methods, A thesis submitted to the Graduate School of Natural and Applied Sciences of Middle East Technical University.

Kropp, B., McRae, E. R., 2022. 11 Trends that Will Shape Work in 2022 and Beyond, 11 Trends that Will Shape Work in 2022 and Beyond (hbr.org)

Kutlugün, M. A., Çakır, M. Y., Kiani, F., 2017. Yapay Sinir Ağları ve K-En Yakın Komşu Algoritmalarının Birlikte Çalışma Tekniği (Ensemble) ile Metin Türü Tanıma, 2 https://www.researchgate.net/publication/323990877.

Liao, C., 2023. Employee turnover prediction using MLmodels.. https://doi.org/10.1117/12.2672733

Liu, H. and Liu, Y., 2021. Visualization research and analysis of turnover intention. E3s Web of Conferences, 253, 02018. https://doi.org/10.1051/e3sconf/202125302018

Maharjan, R., 2021. Employee Churn Prediction using Logistic Regression and Support Vector Machine, San Jose State University, Master's Projects. DOI: https://doi.org/10.31979/etd.3t5h-excq.

Maisuradze, M., 2017. Predictive analysis on the example of employee turnover (Master's thesis). Tallinn University of Technology, Faculty of Information Technology, Department of Computer Systems, 3-76.

Mayson, S., Bardoel, A., 2021. Sustaining a career in general practice: embodied work, inequality regimes, and turnover intentions of women working in general practice. Gender Work and Organization, 28(3), 1133-1151. https://doi.org/10.1111/gwao.12659

McCarthy, A., Moonesinghe, R., Dean, H., 2020. Association of employee engagement factors and turnover intention among the 2015 u.s. federal government workforce. Sage Open, 10(2), 215824402093184. https://doi.org/10.1177/2158244020931847

Moturi, D. G., Wekesa, S., Juma, D., 2023. Influence of self efficacy on employee acceptance levels and use of human resource analytics in microfinance institutions in kenya. International Journal of Business Management, Entrepreneurship and Innovation, 5(1), 31-50. https://doi.org/10.35942/jbmed.v5i1.304

Onnis, L., 2017. Human Resourse Management policy choices, management practices and health workforce sustainability: remote australian perspectives. Asia Pacific Journal of Human Resources, 57(1), 3-23. https://doi.org/10.1111/1744-7941.12159

Pavansubhash, 2016. IBM HR Analytics Employee Attrition & Performance

Peryön., 2018. Çalişan Devir Orani Araştirmasi Sonuç Raporu. In https://www.peryon.org.tr/upload/files/PERYO%CC%88N_C%CC%A7al%C4%B1s%CC%A7an_Devir_Oran%C4%B1_Sonuc%CC%A7_Raporu_2017-2018.pdf.

Poku, C., Alem, J., Poku, R., Osei, S., Amoah, E., Ofei, A., 2022. Quality of work-life and turnover intentions among the ghanaian nursing workforce: a multicentre study. Plos One, 17(9), e0272597. https://doi.org/10.1371/journal.pone.0272597

Putri, M. and Rachmawati, R., 2022. Psychological contract, employee engagement, and perceived organizational support influence on employee turnover intention in pharmaceutical industry.. https://doi.org/10.4108/eai.27-7-2021.2316894

Randstad., 2022. Randstand Trends 2022 Report. In https://www.randstad.gr/. Retrieved February 21, 2024, from https://www.randstad.gr/s3fs-media/gr/public/2022-07/hr-trends-2022-salary-report-eng.pdf

Randstad., 2023. Randstand Trends 2023 Report. In https://www.randstad.com.tr/. Retrieved February 21, 2024, from https://www.randstad.com.tr/s3fs-media/tr/public/2023-04/TR_Turkey%20HR%20Trends%202023_0.pdf

Ribes, E., Touahri, K., Perthame, B., 2017. Employee turnover prediction and retention policies design: a case study, 1-10.

Roche, A., McEntee, A., Kostadinov, V., Hodge, S., Chapman, J., 2021. Older workers in the alcohol and other drug sector: predictors of workforce retention. Australasian Journal on Ageing, 40(4), 381-389. https://doi.org/10.1111/ajag.12917

Rokach, L., Maimon, O., 2005. Decision Trees. In O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook (pp. 165-192). Springer US. https://doi.org/10.1007/0-387-25465-X_9

Russell, D., Zhao, Y., Guthridge, S., Ramjan, M., Jones, M., Humphreys, J. Wakerman, J., 2017. Patterns of resident health workforce turnover and retention in remote communities of the northern territory of australia, 2013–2015. Human Resources for Health, 15(1). https://doi.org/10.1186/s12960-017-0229-9

Scanlan, J., Meredith, P., Poulsen, A., 2013. Enhancing retention of occupational therapists working in mental health: relationships between wellbeing at work and turnover intention. Australian Occupational Therapy Journal, 60(6), 395-403. https://doi.org/10.1111/1440-1630.12074

Schlechter, A., Syce, C., Bussin, M., 2016. Predicting voluntary turnover in employees using demographic characteristics: a south african case study. Acta Commercii, 16(1). https://doi.org/10.4102/ac.v16i1.274

Shanthakumara, A. H., Divya, J., Harshitha, H. T., Pallavi, L. V., Spoorthy, B. C. S., 2022. Prediction of Employee Turnover using Machine Learning. Grenze Scientific Society, 1-13.

Shrivastava, S., Nagdev, K., Rajesh, A., 2017. Redefining HR using people analytics: the case of Google. Human Resourse Management International Digest, 1-4.

Stachová, K., Baroková, A., Stacho, Z., 2021. Optimization of Employee Turnover through Predictive Analysis, Institut of Management, University of Ss. Cyril and Methodius in Trnava, Slovakia. Faculty of Management, Comenius University, Bratislava, Slovakia.

State of the Global Workplace Report - Gallup., 2024. Gallup.com. https://www.gallup.com/workplace/349484/state-of-the-global-workplace.aspx#ite-506924

Uzak, B., 2022. Telekomünikasyon Sektöründe Çalışan Kaybı Tahmini İçin Makine Öğrenmesi Modeli Seçimi (Yüksek Lisans Tezi). T.C. Bursa Uludağ Üniversitesi Fen Bilimleri Enstitüsü, 4-5.

Van Vulpen, E., 2023. What is HR Analytics? All You Need to Know to Get Started. AIHR. https://www.aihr.com/blog/what-is-hr-analytics/

Wijaya, D., Ds, J., Barus, S., Pasaribu, B., Sirbu, L., Dharma, A., 2021. Uplift modeling vs conventional predictive model: a reliable MLmodel to solve employee turnover. International Journal of Artificial Intelligence Research, 5(1). https://doi.org/10.29099/ijair.v4i2.169

Woltmann, E., Whitley, R., McHugo, G., Brunette, M., Torrey, W., Daras, L., Drake, R., 2008. The role of staff turnover in the implementation of evidence-based practices in mental health care. Psychiatric Services, 59(7), 732-737. https://doi.org/10.1176/ps.2008.59.7.732

Yavuz, H. V., 2016, Sanayi ve Hhizmet sektöründe işgücü devir oranlarinin yüksek olmasinin nedenleri ve çözüm önerileri: Denizli örneği (Yüksek Lisans Tezi). Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü. Yüksek Lisans Tezi, Çalışma Ekonomisi ve Endüstri İlişkileri Anabilim Dalı, DENİZLİ, 5-14.

Ye, J., Pu, B., Guan, Z., 2019. Entrepreneurial leadership and turnover intention in startups: mediating roles of employees' job embeddedness, job satisfaction and affective commitment. Sustainability, 11(4), 1101. https://doi.org/10.3390/su11041101

Yedida, R., Reddy, R., Vahi, R., J, R., Abhilash, Kulkarni, D., 2018. Employee Attrition Prediction, https://www.academia.edu/73094870/Employee_Attrition_Prediction, 1-3.

Zhu, Q., Shang, J., Cai, X., Jiang, L., Liu, F., Qiang, B., 2019. CoxRF: Employee Turnover Prediction based on Survival Analysis. In Proceedings of the 2019 IEEE, 1123-1130.