NTU Journal of Pure Sciences
Double Blind Peer Review Journal
ISSN: 2789-1089 EISSN: 2789-1097
Available Online: https://journals.ntu.edu.iq/index.php/NTU-JPS
Volume 5, No. 3, 2026

Original Article

Machine Learning-Based Prediction of Employee Turnover in Iraq's Private and Engineering Sectors

Soud Abdalazez Mohamed Amen [a,*]

[a] Institute of Technical Management – Nineveh, Northern Technical University, 41002 Mosul, Iraq
*Corresponding: Soud Abdalazez Mohamed Amen, Email: s23@ntu.edu.iq

Received: 29 April 2026
Revised: 5 June 2026
Accepted: 10 June 2026
Published: 16 October 2026
DOI: https://doi.org/10.56286/7bfexg17

© 2026 The Author(s). Published by NTU Journal of Pure Sciences. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license: https://creativecommons.org/licenses/by/4.0/

Abstract

It forecasts staff turnover in the private and engineering sectors of Iraq using machine learning methods. The analysis of labor market conditions in Iraq, which exists in conditions of prolonged economic instability, labor mobility, which reveals the problems of staff retention in engineering, manufacturing and technical service companies, justifies the need for proactive decision support tools. Applied research was conducted on the basis of the public dataset Human Resource Analytics. Unlike most studies using the dataset in its original form, a methodology for expanding and transforming factors is proposed. In particular, the subjective indicator of employee satisfaction, rarely observed in real organizational conditions, has been eliminated and a set of derived and integral indicators formed reflecting productivity dynamics, loyalty, workload intensity, seniority effectiveness, professional growth, and stress-related factors. Special attention is paid to the development of the SSI integral indicator (Satisfaction Score Index), which acts as a substitute for the employee satisfaction indicator and allows maintaining the predictive ability of the model. Feature scaling, SMOTE class balancing, cross-validation, comprehensive assessment using ROC-AUC, PR-AUC, and F1-score metrics, as well as isotonic regression for probability calibration to provide reliable and interpretable risk estimates were performed. The experimental results showed that ensemble algorithms (Random Forest, Gradient Boosting) demonstrate the best predictive effectiveness, and the formed new features are identified as the most influential associated with staff turnover. While the study is motivated by the peculiarities of the Iraqi labor market, the methodology proposed can be generalized to private and engineering organizations in other regions under the same conditions of relocation, burnout, and high turnover. The approach proposed can be taken as a basis for the implementation of intelligent systems for predicting the risk of dismissal in HR analytical platforms.

Keywords: Staff Turnover, Machine Learning, Predictive Analytics, Human Resource Analytics, Random Forest, Gradient Boosting

Introduction

Long-term socio-economic instability, resulting from armed conflicts, political changes, low investment, and uneven development of the labor market, has a substantial impact on the activities of the private sector in the course of post-crisis recovery in such countries as Iraq. In this context, one of the major issues that engineering, manufacturing, and service companies face is high staff turnover, particularly among technical specialists and qualified personnel. Talking about the management method in private and engineering companies in Iraq, I would like to clarify that the staff turnover process leads to direct costs associated with recruitment and training. It also disrupts the continuity of production, which in turn reduces the quality of projects and therefore delays the execution of contracts. The acclimatization process of new employees takes a long time, up to This issue creates considerable stress on the existing teams and management structures of specialized engineering companies or those that require practical experience, over the course of several months. Another problem is that of implicit knowledge and experience; it can only be gained through the daily operation of work. In most companies,

Personnel Management is based on a retrospective approach, so the situation becomes more complicated because the analysis is limited to official reports and intuitive decisions of managers; this lets the organization respond to existing problems but not avert them — and, for this reason, most companies underestimate the importance of retaining key specialized employees [1] and do not think about the risk of layoffs. Many businesses suffer from the loss of highly qualified staff because they simply lack the tools for timely intervention in staff retention. The financial and organizational impact of staff turnover is quite enormous, especially in economies that are on the recovery path. Replacing one employee may cost a significant portion of the annual income of the company. These costs include staffing, training, periods of low productivity, and possible work errors. Such losses with limited resources weaken the company's competitiveness and stability in the market. This is why firms need a proactive strategy in managing employee attrition. It is proposed that the use of data analytics and machine learning will help forecast who among the workers is likely to quit and what the key reasons for it are. This will further assist the management in making decisions based on the data available to them, thereby enhancing stability and efficiency at work in private and engineering companies under an unstable economic environment.

The scientific novelty of this work lies in the development of a methodological approach to forecasting staff turnover, focused on the real limitations of corporate HR systems. Unlike most existing studies, the paper suggests:

- avoiding the use of a subjective indicator of employee satisfaction, which in real conditions is either absent or has low reliability.;
- formation of an expanded feature space based on derived and integral indicators reflecting the behavioral, temporal and organizational characteristics of employees;
- development of the SSI (Synthetic Satisfaction Index) integral indicator, which acts as a proxy assessment of satisfaction and allows you to maintain the predictive ability of the model without using subjective data;
- application of a comprehensive model evaluation protocol, including stratified cross-validation, class imbalance accounting (SMOTE), as well as probability calibration.

Thus, the novelty of the work does not consist in the application of individual machine learning algorithms, but in the methodological transformation of the input feature space and the approach to interpreting staff turnover factors.

Materials and Methods

Analysis of the Labor Market and the Problem of Staff Turnover in Iraq's Private and Engineering Sectors

Several features of the Iraqi labor market directly impact employee movement and turnover, particularly in the private and engineering sectors. While government employment means stability and job security over the long term, the private sector often means temporary projects with unsteady, unsupported demand. Hence, the stability of personnel in the fields of engineering, construction, and technology is lower and quickly affected by such conditions.

One of the most apparent features of the private labor market is the instability of the career path. Engineers and technicians work in short- and medium-term projects. Sometimes they change their companies within the same field or even decide to leave the country. This mobility is not only determined by the size of the salary. It is also associated with the volume of work, the duration of the project, and the opportunities for professional growth.

Hence, staff turnover can be regarded as a temporary situation only in relation to the particular company but as ongoing in the nature of the market. Another critical determinant is working pressure. Jobs in engineering and technical fields are typically high effort, long hours, and high responsibility. Most companies do not monitor the condition of the employee systematically. Therefore, fatigue and reduced enthusiasm are only apparent when the work is left. In the end, the company becomes aware of the problem after it has occurred, without really knowing the reasons. Most private firms still handle human resources data operationally, not analytically. Information about employees covers only the basics. Such things as work-life balance, experience utilization, or a change in productivity are seldom measured. That is why cases of leaving work are usually analyzed after their occurrence.

Companies do not have any tools to help them figure out who might quit before then. Therefore, the problem of staff turnover in the private and engineering sector in Iraq is not only related to the economic situation-it is the result of several factors working together, including the nature of employment, the pressure of work, and the level of productivity and working conditions within the company. Understanding this problem calls for a better analysis that connects these factors and reveals unclear patterns. The labor market analysis for Iraq illustrates the necessity of applying data analysis techniques even when the required data is scarce. The construction of predictive models enables firms to shift from being reactive to taking early actions.

This will help in stabilizing the workforce, reducing risks, and enhancing the utilization of cadres in engineering and technical firms. Many studies have been conducted in Iraq and neighboring regions regarding the topic of staff turnover. Most of them have been related to organizational and behavioral factors, using questionnaires. These studies provide a good general picture, but they have limitations in the analysis method. In a study [2], the analysis involved the intention of employees to leave work in the private sector in Iraq. The study primarily focused on work-life balance. It was questionnaire-based and analyzed using statistical tools like SPSS.

The analysis comprised descriptive statistics, correlation, and regression. It was found that there is a significant relationship between work-life balance and the decision to leave a job as it is only based on intentions and not actual behavior and only on a linear relationship. Hence, it does not model the actual layoffs of workers or the interaction of multiple risk factors over time. In a related study, the authors examine the effects of several factors on decreasing staff turnover at Iraqi telecommunications firms.

Their analysis uses conventional statistical techniques, mainly regression analysis and hypothesis testing, to make decisions regarding employee retention. Although this study helps one to learn how organizational measures can reduce staff turnover, it is more of a descriptive one than a predictive model that can identify employees at high risk of being laid off. This paper relates to the correlation of the level of job satisfaction and intention to quit in the teaching staff of private higher education in Iraq. Survey-based quantitative analysis is applied to establish relationships between variables related to staff satisfaction and the intention to quit, once again using correlation and regression methods. The results do confirm the relevance of the components of employee satisfaction. However, the academic context is quite different from private engineering or technical companies.

Also, basing self-assessment of satisfaction limits the possibility of transferring such research and its use in the practice of personnel management systems. Organizational problems related to staff turnover in medical institutions are discussed in [5]. Even though the field transcends engineering, the study is applicable in the conceptualization of staff turnover as a multi-factorial organizational phenomenon. However, the methodological approach is mainly qualitative and descriptive, focusing on organizational issues rather than predictive assessment or risk modeling. The reviewed publications indicate that most of the research on staff turnover in Iraq

was conducted using basic classical statistical methods. These included correlation analysis and linear regression, mainly between variables related to the intention of employees to quit their jobs, rather than actual turnover.

These techniques are very helpful in testing hypotheses and interpreting outcomes. However, they are not very effective in understanding the complicated relationships between HR variables, more so when they are non-linear. Neither can they make accurate and personalized predictions of the risks of leaving work. In this study, a different method was used. The research employed machine learning in the assessment of employee turnover. Emphasis was placed on the accuracy of the prediction to ensure that the model was right and that there was a chance of understanding the results clearly. Using ensemble algorithms and built HR management indicators from operational HR data, the proposed approach removes the major drawbacks of previous work and enables moving from descriptive analysis to proactive and scalable forecasting of staff turnover, which is particularly appropriate for private and engineering organizations.

Analysis of Existing Publications on Staff Turnover Assessment

In the scientific literature, the issue of forecasting staff turnover is considered within the framework of modern approaches to HR analytics and human resource management. In HR analytics, there are four types of analytics (descriptive, predictive, predictive, prescriptive), and starting with predictive analytics, machine learning becomes a means of achieving research goals. In this study of staff turnover forecasting, we will be interested in predictive analytics using machine learning models.

Researchers mostly focus on applying machine learning methods to build models that can predict the likelihood of employees being fired [6, 7]. Classical classification algorithms are preferred – logistic regression, decision trees, ensemble methods (Random Forest, Gradient Boosting) and various modifications of neural networks. Each algorithm has its advantages and disadvantages, and the best algorithm for a particular demand forecasting task will depend on the specific characteristics of the dataset and the problem at hand. In [6], an accuracy of 88% was achieved using the publicly available IBM HR dataset using the support vector method, and the key factors consistently included monthly income, hourly rate, job level, and age. In the same data set in the reference [7], the logistic regression model achieved the best accuracy of 87.7%, while the random forest reached 83.2%. In reference [8], the Gaussian Naive Bayes classifier achieved the best results on IBM HR data. These results demonstrate that simple models, if tuned well, can give performance close to or even better than complex models.

Another direction in research is focused on data processing before analysis and the problem of category imbalance. In most cases, the number of employees who leave is much less than those who stay, and this affects the accuracy of the models [9]. Also, there is an increasing interest in the use of explainable AI methods (XAI) [10, 12], since they help to understand the results of models and use them in making managerial decisions. Also, some studies have begun to focus on data specific to specific areas and certain categories of employees, rather than relying solely on General Data.

This opens the way to the development of more adapted models that can take into account the specifics of specific markets or professional communities. Thus, in [11], staff turnover is considered in the context of behavioral and psychological factors, which goes beyond the use of

universal open datasets and shows an example of analysis in a specific domain – targeting special groups of employees and socio-psychological factors.

In general, a review of current research shows that the problem of staff turnover is being actively studied using a wide range of machine learning methods. However, gaps remain related to the interpretation of results, the use of real corporate data, and consideration of dynamic factors influencing employee decisions. This creates the basis for further research in predicting staff turnover using machine learning methods.

Methodology for Employee Turnover Prediction

We are interested in the internal risks of voluntary dismissal by the employee, as it will lead to a great risk for the company. The company's ability to predict the future is an added advantage to its development. Since the risks of staff turnover are under control, the company will be more assured of the future, which means that it will invest in its employees without fear of the risk of their rapid dismissal. Therefore, there is a need to predict the activity of the company's personnel, which will minimize the associated risks. The study aimed to compare several machine learning models to be used in different types of organizations for staff retention and reducing turnover.

Dataset Description

The study used the open data set Human Resource Analytics from the Kaggle platform, which has information on 15,000 employees. It contains 10 features (as shown in Table 1) with the target variable as left: that is predicting whether an employee will quit or not. The sample includes employees from 10 departments. This data set is widely used in scientific research, for example, to analyze staff turnover, which allows comparing the results obtained with existing ones.

Table 1. Description of dataset features.

Sign	***Description***
satisfaction level	The employee's satisfaction level with the company's work. The value ranges from 0 to 1, where 0 is total dissatisfaction and 1 is total pleasure
last evaluation	Evaluation of the employee's work based on the last completed assessment. Values range from 0 to 1
number project	The number of projects that the employee worked on
average monthly hours	The average number of hours an employee spends on work per month
time spend company	The total time that an employee has already worked for the company in years
work accident	Whether the employee had an accident at work. A value of 1 indicates its presence, and 0 indicates its absence.
left	Whether the employee has quit. A value of 1 indicates dismissal, and 0 indicates remaining in the company.
promotion last 5 years	Whether the employee has been promoted in the last 5 years. A value of 1 indicates the presence of a promotion, and 0 indicates its absence.
department	The department where the employee works
salary	The employee's salary level. Possible values: "low", "medium", "high"

Exploratory Data Analysis

If we calculate the staff turnover rate for all time as the ratio of the number of people who dropped out due to turnover and the average number of employees during the same period, it turns out that it is 23.8%. This is very important because the normal staff turnover rate is considered to be up to 10%. Even realizing that this indicator may vary depending on the field of activity of the enterprise, there is a significant excess of the average value. The chart in Fig. 1 shows the number of vacated and remaining employees in each department.

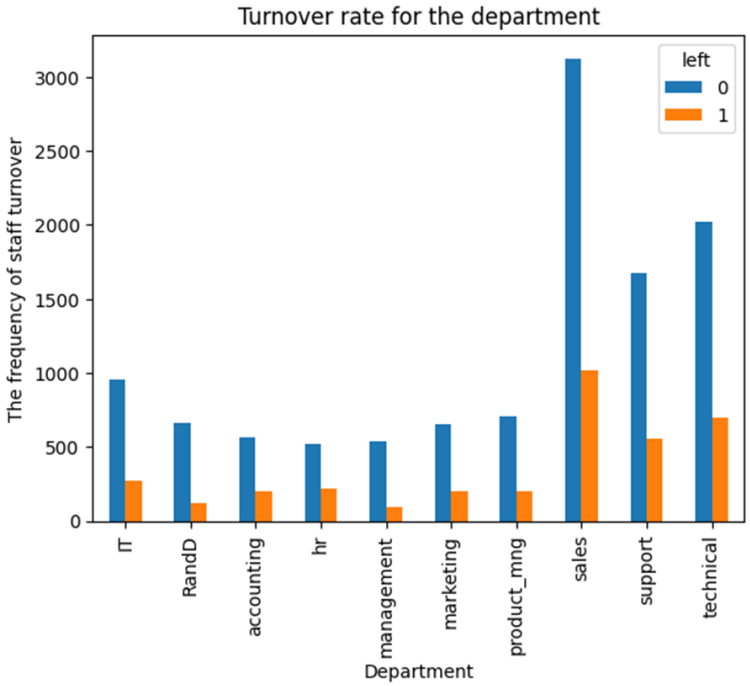
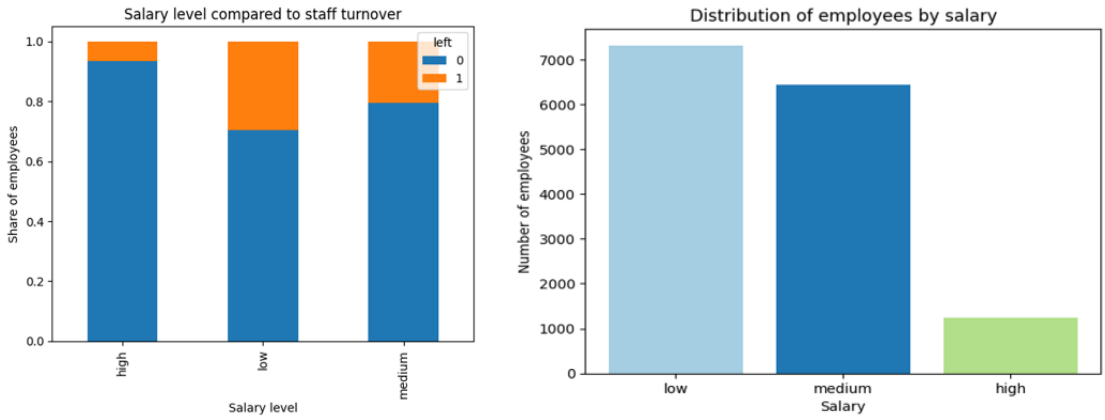


Figure 1. The number of dismissed and remaining employees in each department.

A noticeable difference for the two groups is observed in the "satisfaction level" feature. It is hard to say what this feature represents in numerical terms, so most probably the value of this feature was based on the results of other tests or assessments (which we do not know) aimed at quantifying employee productivity. It can be seen that those who were promoted mostly stayed, a difference of almost 5 times, which underscores just how important career growth is when it comes to holding onto good people. It can also be noted that the values of the attributes for each department are about the same and they are at a good level, which indicates successful management. This shows that department management in the context of employee maintenance is no small issue at all. It would be fair to note that the company is improving its sales due to the high involvement of many workers in customer support, while tech specialists take care of software support and development. Observe the correlation: workers with low pay and no promotion are much more likely to leave the company (Fig. 2, a). Let's do the math and find the relative proportions by dividing each value in a row of the table by the sum of the values in that row. This shift let us get a share of vacated and remaining employees for each level of salary (Fig. 2, b).



a) b)

Figure 2. Salary level compared to staff turnover and employee salary distribution.

There are a number of important correlations in this data set for the study of staff turnover (Fig. 3). The most negative relationship was found between employee satisfaction and the fact of dismissal. This supports the hypothesis that unsatisfied workers are much more likely to leave a company. There is also a positive moderate relationship between the results of the latest assessment and the number of projects, and between the number of projects and the average number of hours worked. This may mean that the most productive or overworked workers are more likely to get higher grades and work more hours. Meanwhile, experience at the company and recent promotions or accidents have almost no effect on the decision to leave the job. This proves the assumption that the primary factor in staff turnover is not official indicators of length of service or personnel-related events but rather, an internal feeling of satisfaction and a balance between workload and recognition. Thus, the analysis of the correlation matrix supports the working hypothesis regarding the crucial role of employee satisfaction in predicting their decision to leave the company and also underscores the importance of workload management and an adequate assessment of work results.

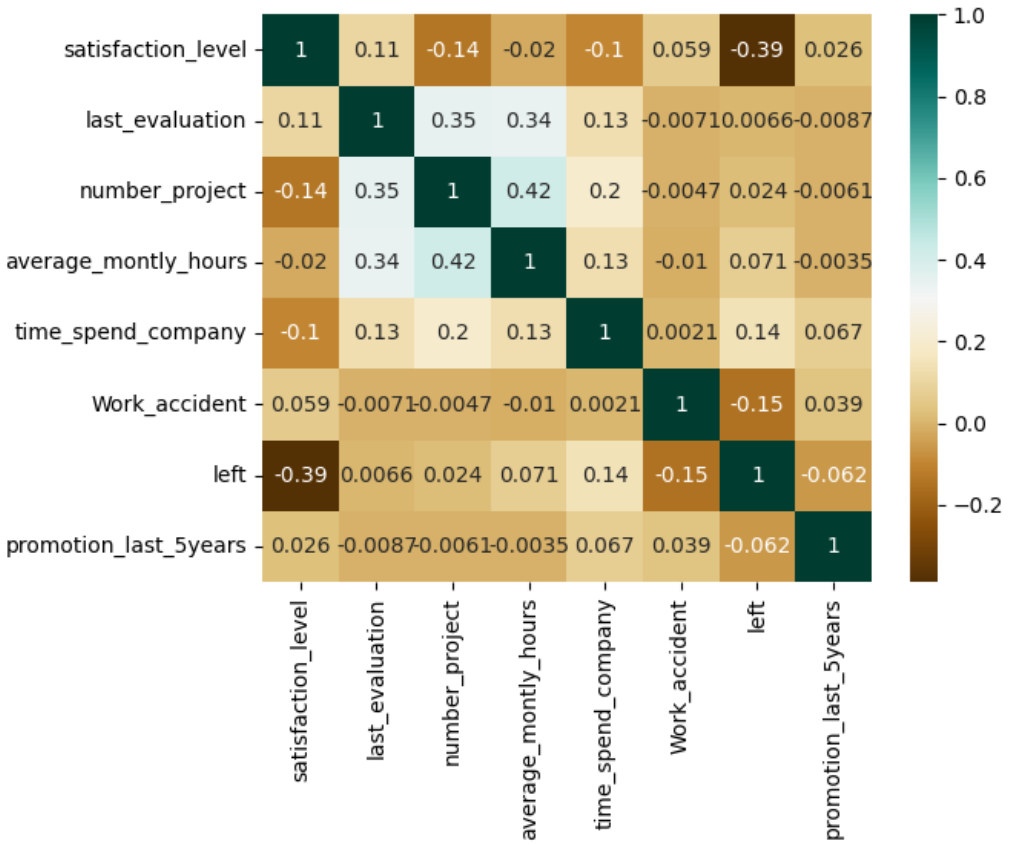


Figure 3. Heat map of correlations of the initial HR dataset for predicting staff turnover.

Feature Space Expansion

Even though this set has been widely used in studies, its predictive powers primarily rely on the variable satisfaction_level, which is subjective and often not available in real HR systems. In an attempt to enhance the practical value of the study, the feature space was enriched by creating additional variables that reflect the behavioral, temporal, and motivational-psychological characteristics of employees.

In particular, such derived indicators have been developed that can be provided by an HR analyst:

- performance_gap – the difference between an individual employee's performance assessment and the average value in his department, reflecting the degree of professional compliance or deviation from the norm;
- workload_index – the ratio of the average monthly workload to the number of projects, characterizing the intensity of work and the level of workload;
- tenure_efficiency – the ratio of individual performance assessment to the duration of work in the company, describing the dynamics of productivity over time;
- loyalty_score is an integral metric that takes into account work experience, the absence of accidents and the presence of promotions over the past five years, forming an indicator of employee loyalty;
- burnout_risk is a binary sign indicating the potential risk of professional burnout when combined with a high workload, a large number of projects, and a low efficiency score;
- stress_index is a relative indicator of work intensity, calculated as a function of the average workload, the number of projects and the time spent in the company;
- growth_opportunity is an indicator of the possibility of professional development, taking into account the fact of increasing and the ratio of seniority to productivity;
- work_balance is a measure of the balance between workload and stability, reflecting the deviation of work intensity from the average level within the company;

– experience_saturation is a characteristic of the level of "saturation of experience", which increases with the duration of work in the company, but decreases with low assessment rates.

To replace the subjective variable satisfaction_level, the construction of an integrated Satisfaction Score Index (SSI) is proposed, reflecting the generalized level of employee satisfaction based on behavioral and temporal characteristics.

The index was formed as a normalized combination of pre-calculated indicators characterizing the balance between productivity, workload, stability and professional development.

$$SSI = \alpha1 \cdot (1 - stress_index) + \alpha2 \cdot work_balance + \alpha3 \cdot growth_opportunity + \alpha4 \cdot loyalty_score - \alpha5 \cdot burnout_risk + \alpha6 \cdot experience_saturation$$

where the coefficients αi are determined empirically based on normalized correlations between the corresponding indicators and the initial variable satisfaction_level in the training sample. To simplify interpretation, all weights are scaled.

Thus, the SSI index is an aggregated indicator of an employee's psychological state, integrating the impact of workload, development opportunities, loyalty, and burnout risk.

Preliminary statistical analysis (Fig. 4) showed that the constructed SSI has an average positive correlation with satisfaction_level (r=0.56), which confirms its ability to reproduce the emotional and motivational aspects of staff engagement.

In addition, there is a strong inverse relationship between SSI and the layoff rate (left, r = -0.43), which even exceeds the correlation between satisfaction_level and left (-0.39).

hus, SSI can be considered as an effective substitute for subjective satisfaction assessment, suitable for practical forecasting of staff turnover.

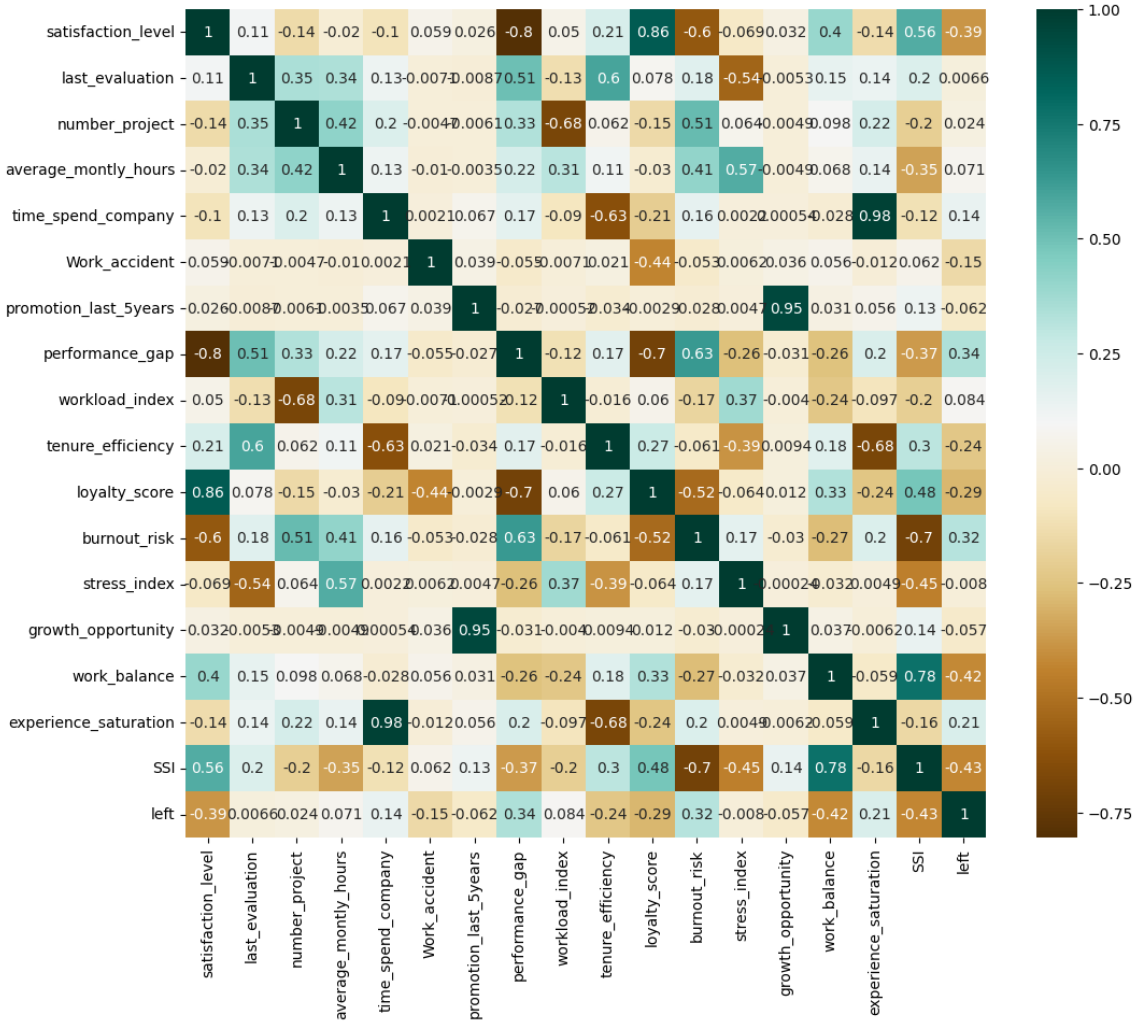


Figure 4. Heat map of correlations of the modified HR dataset for forecasting staff turnover.

Research Methods and Steps

The study of staff turnover is based on the integrated approach that unites the stages of data collection and its processing, the selection of relevant indicators, and the construction of forecasting models. At the initial stage, we worked with the initial data on personnel and organizational characteristics; then, we analyzed them after cleaning for further preparation. > Before modeling, all numerical variables were normalized using Min-Max Scaling in the range [0;1], which ensured the same scale of features. Categorical variables (department, salary) are transformed by the one-hot encoding method. The satisfaction_level attribute is completely excluded from further calculations, and SSI is used as the main characteristic of the psychological state (satisfaction) of employees. The next step was to identify key predictors of employee churn and to obtain a balanced dataset for model training. In order to address the issue of class imbalance, the SMOTE algorithm was used, which carried out oversampling only on the training sample. This method ensured that examples from both classes would be presented equally when training the model, while the test sample remained unchanged; this way, there would be no information leakage. Further research has been directed toward developing predictive models using machine learning techniques and assessing the accuracy of the models. The obtained results allowed for the effectiveness evaluation of different methods and determination of the most appropriate ones for turnover prediction with the existing information base.

Protocol for building and validating models

A standardized experimental protocol was used to ensure reproducibility and correctness of the results.

The simulation included the following steps:

- Data preprocessing: normalization of numerical features (Min-Max scaling); one-hot encoding of categorical variables; exclusion of subjective satisfaction index.
- Formation of an extended set of features (feature engineering).
- Sample separation: stratified train/test partitioning; using 5-fold cross-validation.
- Class balancing: using the SMOTE method exclusively on the training sample to prevent data leakage.
- Model training: Random Forest, Gradient Boosting, Decision Tree, Logistic Regression, MLP.
- Quality assessment: PR-AUC (basic metric), ROC-AUC, F1-score, Recall.
- Probability calibration: using isotonic regression to improve the reliability of probability estimates.
- Optimization of the classification threshold: selection of the threshold maximizing the F1-score.

This protocol avoids typical errors (data leakage, bias due to class imbalance) and ensures correct interpretation of the results.

Results

The accuracy of staff turnover predictions was evaluated for five machine learning algorithms: Random Forest, Decision Tree, Gradient Boosting, Logistic Regression, and Multi-Layer Perceptron. The main classification metrics that were used in this assessment are Precision, Recall, F-score, and Accuracy. For the correct comparison of models, we used a unified evaluation protocol stratified 5-fold cross-validation with SMOTE over sampling within each fold (without data leakage).

Model quality was assessed using PR-AUC (Average Precision) as the main metric, and ROC-AUC, F1, and Recall for the "quit" class. Classification probabilities were calibrated by isotonic regression as implemented in the CalibratedClassifierCV ensemble, after which the decision threshold was set to maximize F1 on out-of-fold predictions. Such strategy assures reproducibility and comparability of results without any bias due to class imbalance or information leakage. Results of 5-fold stratified cross-validation with SMOTE balancing are presented in Table 2 which demonstrates high model stability.

Table 2. Comparison of the effectiveness of staff turnover forecasting models.

Model	*PR-AUC*	*ROC-AUC*	*F1(class=1)*	**Recall (class=1)**
Logistic Regression	0.786 ± 0.015	0.929 ± 0.005	0.742 ± 0.004	0.899 ± 0.008
Decision Tree	0.903 ± 0.010	0.973 ± 0.003	0.946 ± 0.006	0.971 ± 0.004
Random Forest	**0.988 ± 0.003**	**0.993 ± 0.003**	**0.975 ± 0.002**	**0.969 ± 0.003**

Model	PR-AUC	ROC-AUC	F1(class=1)	Recall (class=1)
Gradient Boosting	0.974 ± 0.003	0.985 ± 0.002	0.921 ± 0.007	0.932 ± 0.008
MLP Classifier	0.974 ± 0.005	0.986 ± 0.002	0.920 ± 0.006	0.945 ± 0.008

Random Forest showed the best consistency between all metrics, reaching values of PR-AUC = 0.988 and F1 = 0.975, which indicates an excellent ability to distinguish employees with a high probability of dismissal (Fig. 5).

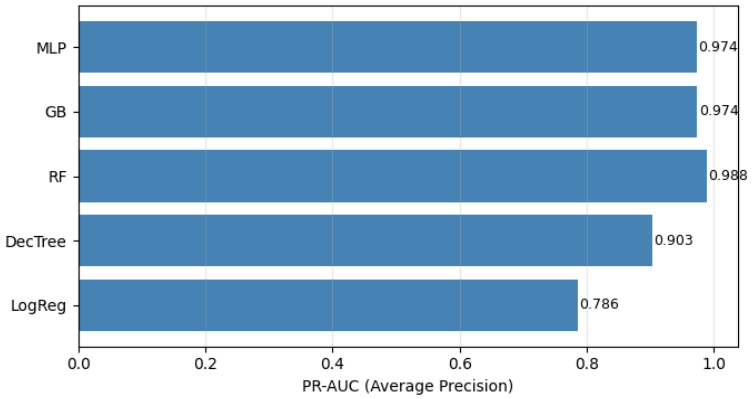


Figure 5. Comparison of model quality by PR-AUC metric (5-fold stratified CV).

The Gradient Boosting and MLP models also showed high quality, confirming that ensemble and neural network approaches are most effective in conditions of nonlinear the relationships between the features. Logistic Regression demonstrated the lowest accuracy among all models, which may be due to the limited ability of linear methods to describe complex interdependencies between the behavioral characteristics of employees.

The use of an improved dataset with new integrated features (performance_gap, workload_index, tenure_efficiency, loyalty_score, stress_index, growth_opportunity, work_balance, experience_saturation, SSI) significantly increased the predictive ability of the models.

Comparative analysis has shown that ensemble methods (Random Forest, Gradient Boosting) are superior to classical statistical models.

High values of PR-AUC (0.974-0.988) confirm the effectiveness of the models in identifying cases of probable dismissal of employees, even with a significant class imbalance.

The results also show the efficiency of the built-in Employee Satisfaction Index feature, which replaces the basic variable "level of satisfaction" without compromising the quality of forecasting.

The random forest model maintains an accuracy above 0.95 for repetition rates up to 0.9, which indicates a high ability to identify employees who are likely to leave without a large number of false positive results (Fig. 6).

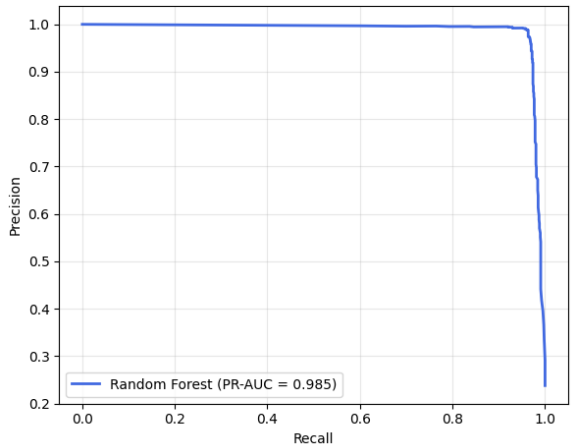


Figure 6. Precision-recall curve of the random forest model.

To find out if the forecast is close to reality, the random forest model was modified using CalibratedClassifierCV. The calibration curve in Fig. 7 shows that the results were accurate at high values. This is important when making decisions that directly concern each employee.

But in the average values there is a clear difference between expectation and reality. This means that the model needs improvement in this area. It can be improved using methods such as Platt Scaling or isotonic regression.

The model achieved balanced performance with very few negative errors. These cases are for employees who the model predicted would stay, but actually left as shown in Fig. 8.

In order for the forecasts to be closer to reality, the random forest model was modified using is isotonic regression. After this adjustment, the results remained strong or improved. The value of ROC-AUC reached 0.992 and PR-AUC reached 0.986, this indicates a high ability in discrimination even after calibration. Also, the accuracy was 0.988 and the recall was 0.963, and this helps to reliably identify employees at risk of leaving work. Therefore, this model can be practically used in the analysis of human resources.

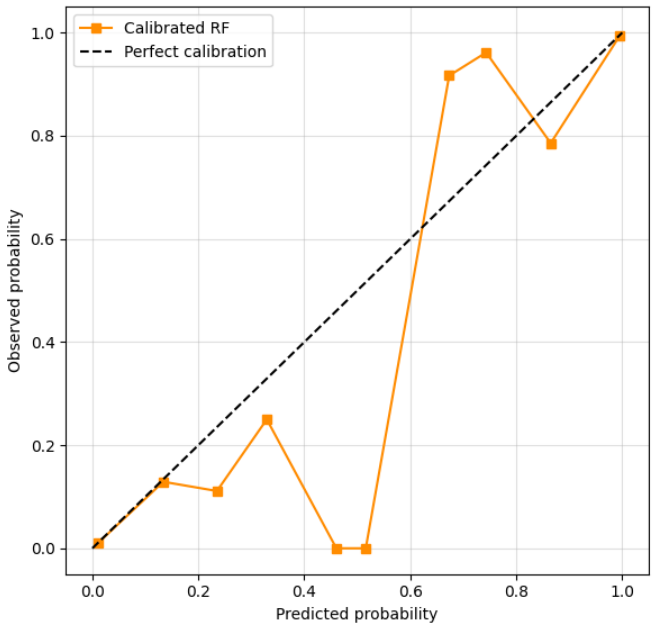


Figure 7. Calibration curve of the Random Forest Model.

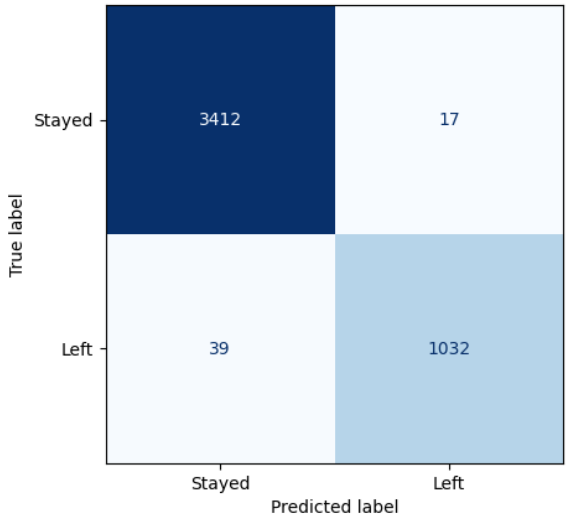


Figure 8. Confusion matrix for the Random Forest Model after applying SMOTE.

These results are consistent with other studies [10], which have indicated the power of aggregation models in predicting employee turnover. Tree models not only give high accuracy, but also help to understand the influencing actors. This can be found out by feature importance assessment methods or XAI approaches (SHAP, LIME) [12]. This is of practical importance for HR analytics, because it

allows not only to anticipate the likelihood of an employee's dismissal, but also to understand what factors determine his decision to stay or leave the company.

Discussion

The obtained modeling results deliver several key conclusions in the practice of HR analytics. First, and foremost, it is the high efficiency of ensemble models. The Random Forest and Gradient Boosting algorithms demonstrated the highest accuracy rates (Accuracy ≈ 0.99 and 0.97, respectively), which indicates their ability to make accurate forecasts on potential layoffs. This allows HR departments to receive reliable predictive signals before the situation becomes critical. Furthermore, unlike "black boxes", ensemble trees do retain the ability to identify the most important signs of influence (Fig. 9). We constructed a Random Forest model to interpret the behavior of the model and assess the contribution of each feature to decision-making. The feature weights were obtained using the attribute feature_importances, which estimates the average increase in purity (Gini impurity) when broken down on each feature.

Fig. 9 shows a diagram illustrating the importance of the features in descending order. The following variables are most important to the model.

- work_balance (importance = 0.123): Work-life balance indicator is the most influential factor in predicting layoffs. This highlights the important role that working conditions play in shaping employees' intention to leave.
- time_spend_company (0.101): Tenure with a company also has a significant impact on the likelihood of being fired, suggesting a cumulative effect of experience and burnout.
- experience_saturation, tenure_efficiency, performance_gap, SSI (Overall Satisfaction Index): The weight of each of these variables ranges from 0.075 to 0.089, indicating a complex interaction between employee efficiency, expectations, and subjective ratings.
- There is a significant effect (≈ 0.07) between loyalty score and average monthly hours, confirming the role of emotional connection and workload.

Of the categorical characteristics such as department and salary category, salary_low and Department_technical matter most — but their overall effect is much smaller than the quantitative characteristics. These results are in line with what"s expected — and what other studies have found in this field. They speak to the value of rolling out work-life balance support programs, spotting burnout symptoms in your team, and creating a space for growth and recognition.

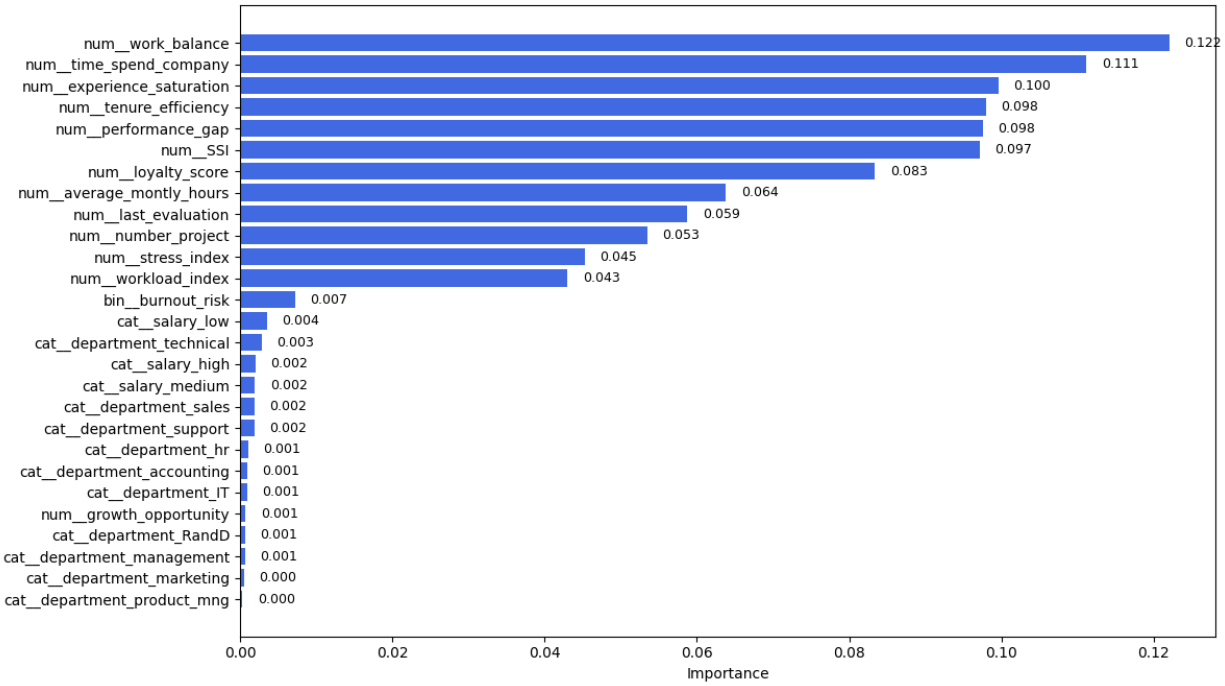


Figure 8. Importance of features for Random Forest.

This permits firms not just to forecast the risk of firing, but also to know what inspires workers to switch jobs. The results are consistent with studies in which tree-based ensemble algorithms are often optimal for tasks of staff turnover, especially on tabular HR data [10]. In [11], they obtained the following results on the same dataset used in this work: the XG Boost Classifier model – 98.8% and Random Forest - 98.7%, which corresponds to the results obtained in our study. Although satisfaction_level has been widely used in previous studies on the prediction of staff turnover, its application is rather limited. In practice, more specifically at private and engineering companies, such indicators are often not available or measured irregularly and are not very reliable since they are usually based on self-assessment or there is no systematical record.

Therefore, to exclude such subjective variables from the model means to give up one of the most frequently mentioned predictors of staff turnover, and this can probably reduce the forecasting accuracy. Still, as the results of the study show, the creation of alternative features that indirectly reflect aspects of employee satisfaction not only compensated for the loss of satisfaction_level but also enabled the achievement of high-quality indicators of the model (ROC-AUC = 0.99, PR-AUC = 0.99). These features are indices of work-life balance, seniority efficiency, congestion, stress, loyalty, and SSI (Satisfaction Score Index) as integral indicators specially introduced by us. Many of these new features are among the most important according to Random Forest, which confirms the high level of informativeness of the new feature vector. This poses a question mark on the portability of the model to other companies or labor markets.

The dataset used can be considered as generalized because it reflects standard HR management attributes, and not those of industry or country conditions. For future research in the private and engineering sectors of Iraq, it would be more relevant to add other variables, such as: Employee relocation and external mobility.

- Type of project and duration of the contract.
- Segmentation of technical staff by role.
- Workload variability and scheduling constraints.
- Compensation dynamics and non-material benefits.
- Indicators of a gradual decrease in engagement or decrease in participation in operational activities.

Further development of indicators is the use of time models (for example, time-to-event approaches) that would make it possible to predict not only the fact of resignation but also the estimated period when it may occur. Integration of such approaches with operational monitoring (HR dashboards) will give a possibility to form active management of the risk of staff turnover. The results obtained have not only practical but also theoretical significance. They confirm that staff turnover can be considered as the result of non-linear interaction between multiple factors, including workload, dynamics of productivity, and the accumulated experience of the employee.

The high importance of integral indicators (for example, work_balance, tenure_efficiency, SSI) shows that the traditional linear models, which many studies use, cannot fully capture the complexity of personnel processes. Also, the fact that the synthetic SSI index replaces the subjective satisfaction index successfully proves that it is possible to approximate hidden psychological and motivational factors through observed behavioral characteristics, which is quite interesting for further research in the field of interpreted machine learning.

Conclusion

The paper investigates the problem of forecasting staff turnover based on the open data set Human Resource Analytics.

The main scientific contribution of the work is as follows:

- An approach to building staff turnover models without using subjective indicators is proposed, which increases the applicability of models in real conditions.
- A system of derived and integral features has been developed that reflect the multidimensional nature of employee behavior.
- A new integrated SSI indicator has been introduced, which allows modeling employee satisfaction through observable data.
- The effectiveness of ensemble machine learning methods in an expanded feature space is demonstrated.
- A reproducible protocol for evaluating models has been proposed, taking into account class imbalances and the need to calibrate probabilities.

- Thus, the work contributes to the development of methods for analyzing tabular HR data and expands the possibilities of using machine learning to predict complex socio-economic processes.

Special attention was drawn not to the enhancement of model precision alone but also to the methodological improvement of feature structure. In this case, the variable satisfaction_level was removed from the feature set since it is unlikely to be measured in real business processes, and a new set of derived and aggregated indicators on various aspects of employee labor behavior was created. These include performance_gap, workload_index, tenure_efficiency, loyalty_score, stress_index, growth_opportunity, work_balance, experience_saturation, and SSI (Satisfaction Score Index) — an integral measure that compensates in part for what's missing. It was found that the new expanded feature space helps to keep or even improve the quality of forecasting models. Experiments have confirmed that ensemble algorithms (Random Forest, Gradient Boosting) demonstrate the best results for metrics ROC-AUC (0.992), PR-AUC (0.986), F1-score (0.975) after calibration of isotonic regression. An analysis of the importance of the features has shown that work_balance, time_spend_company, experience_saturation, tenure_efficiency, and SSI are among the most influential, which suggests a need for indicators based on complex employee characteristics. The result can be considered as a confirmation of the possibility of achieving high predictive accuracy without the use of direct satisfaction metrics if the model based on well-constructed derived metrics. From the practical point of view of human capital analysis, it allows the creation of adaptive models under specific company or industry conditions characterized by economic uncertainty or workforce mobility. It is recommended to consider further research on broadening the scope of functions in terms of specific parameters (employee reassignment, project duration, contract type, role segmentation, compensation dynamics, etc.) and applying time-to-occurrence models not only to predict the facts of the problem but also its termination time. This lays the foundation for an intelligent early warning system on workforce fluctuations and management decision support in HR analytics. Modeling results should be used in prediction and targeted employee retention strategies, such as job relocation, career development planning, salary adjustments, and improving working conditions. These would contribute to increased stability of the workers and hence of the organization as well.

Conflicts of Interest

The author declares no conflict of interest.

Acknowledgments

The author would like to thank the academic staff at Northern Technical University for their support and guidance during this research.

References

[1] Madanchian M. From recruitment to retention: AI tools for human resource decision-making. Appl Sci. 2024;14(24):11750. doi:10.3390/app142411750

[2] Abubaker MQF, Rashid AHA, Ishak WZA, Rahim NSA, Rashid YKA. Work–life balance and employees' turnover intention: a study of private banks in Sulaymaniyah, Iraq. Int J Acad Res Bus Soc Sci. 2025;15(12). doi:10.6007/IJARBSS/v15-i12/27194

[3] Abd I, Joumaa B, Ali N. The impact of job enrichment in reducing the turnover rate in Iraqi telecommunications companies. Afr J Econ Sustain Dev. 2025;8(3):37–54. doi:10.52589/AJESD-UBX2JIHJ

[4] Serin H, Qasim Z, Mansoor M. Impact of job satisfaction on turnover intention among academic personnel of private higher education sector in Iraq: a case of Tishk International University. Int J Res Bus Soc Sci. 2022;11(2):402–413. doi:10.20525/ijrbs.v11i2.1641

[5] Samadi SA, Biçak CA, Osman N, Abdalla B. Organizational challenges in healthcare services providers for individuals with autism spectrum disorder (ASD) considering personnel turnover rate. Brain Sci. 2023;13(4):544. doi:10.3390/brainsci13040544

[6] Raza A, Munir K, Almutairi M, Younas F, Fareed MMS. Predicting employee attrition using machine learning approaches. Appl Sci. 2022;12(13):6424. doi:10.3390/app12136424

[7] Kumar P, Gaikwad SB, Ramya ST, Tiwari T, Tiwari M, Kumar B. Predicting employee turnover: a systematic machine learning approach for resource conservation and workforce stability. Eng Proc. 2023;59(1):117. doi:10.3390/engproc2023059117

[8] Fallucchi F, Coladangelo M, Giuliano R, De Luca EW. Predicting employee attrition using machine learning techniques.

Computers. 2020;9(4):86. doi:10.3390/computers9040086

[9] Kaggle. Human resource dataset [Internet]. Available from: https://www.kaggle.com/datasets/kuniowu/human-resource [cited 2026 Jan 28].

[10] Krishna S, Sidharth S. Analyzing employee attrition using machine learning: the new AI approach. In: Proc IEEE 7th Int Conf Convergence Technol (I2CT); 2022; Mumbai, India. p. 1–14. doi:10.1109/I2CT54291.2022.9825342

[11] Iparraguirre-Villanueva O, Chauca-Huete L, Prieto-Chavez R, Paulino-Moreno C. Employee attrition prediction using machine learning models. In: Proc 22nd LACCEI Int Multi-Conf Eng Educ Technol; 2024 Jul 17–19; San José, Costa Rica. doi:10.18687/LACCEI2024.1.1.498

[12] Baydili IT, Tasci B. Predicting employee attrition: XAI-powered models for managerial decision-making. Systems. 2025;13(7):583. doi:10.3390/systems13070583Smith J, Brown L. Title of the article. Journal Name. 2020; 12(3):123-130.