PROCEEDINGS OF
INTERNATIONAL CONFERENCE ON INTELLIGENT SYSTEMS AND NEW APPLICATIONS
https://proceedings.icisna.org/
3nd International Conference on Intelligent Systems and New Applications (ICISNA'25), Antalya, December 12-14, 2025.

# An Intelligent Ann-Based Framework for Predicting Employee Attrition in Imbalanced Data Scenarios

Esmael Ahmed[1*], Kedir Abdu[1], Mohammed Omer[2] , Tigist Mintesnot[3]

*1 Information System, College of Informatics, Wollo University, Dessie 7200, Ethiopia.*
*2 Computer Science, College of Informatics, Wollo University, Dessie 7200, Ethiopia.*
*3University of Gondar, College of Informatics, Gondar, 1000., Ethiopia.*
**Corresponding Author: Esmael Ahmed; email: esmael.ahmed@wu.edu.et*

***Abstract*— Employee attrition poses a significant threat to organizational stability and performance. While intelligent systems offer a powerful solution, predictive accuracy is often hindered by the inherent challenge of imbalanced data, where the number of employees who stay far exceeds those who leave. This study proposes a novel intelligent framework for employee attrition prediction that directly addresses this data imbalance. We conduct a comprehensive comparative analysis of six machine learning algorithms: Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, CatBoost, and XGBoost using a dataset of 1,410 employee records. To enhance model performance and mitigate imbalance, we implemented rigorous hyperparameter tuning and the Adaptive Synthetic Sampling (ADASYN) technique. Our results demonstrate that the ANN model significantly outperformed its counterparts, achieving the highest predictive accuracy and F1-score. The model identified key attrition drivers, including frequency of illness, monthly income, and overtime work, corroborating existing literature on the primacy of well-being and compensation. This research not only validates ANN as a superior intelligent system for this critical business application but also provides organizations with an actionable, data-driven framework for identifying attrition risks and implementing targeted retention strategies.**

***Keywords*— Intelligent Systems, Artificial Neural Network, Imbalanced Dataset, Adaptive Synthetic Sampling (ADASYN), Employee Attrition.**

## I. INTRODUCTION

In today's data-driven economy, technological specialization and knowledge generation are driven by data gathering, inquiry, and analysis in today's competitive economy. Information technologies serve as both data sources and catalysts for data analysis, making data a strategic asset for enterprises across various industries, especially those in business operations [1]. Utilizing new technology in organizations enhances efficiency and competitive advantage through data collection, management, and analysis, leading to effective decision-making, goal achievement, and improved economic competitiveness [2].

The importance of human resources (HR) has recently grown because skilled employees give companies a significant competitive edge [3]. Examining employee data, HR can foster a supportive workplace, boosting productivity and driving organizational success [4]. Data analysis enables better management decisions, leading to increased employee retention. Employee attrition occurs when valuable employees leave a company. Several factors, such as job stress, negative workplace conditions, or low pay, can lead to this outcome. Losing employees hurts company productivity; it means losing productive workers and the resources spent by HR recruiting replacements. Effective new employee onboarding requires training, development, and acclimation [5].

Many decision-makers in any organization must have a clear understanding of who poses the largest retention risk or who might be targeted for poaching intentionally. Retention strategies can be modified by assessing retention risk and estimating the likelihood of leaving, which helps to lower the high cost of hiring and onboarding new employees. [6]. The gradual loss of representation as a result of retirement, renunciation, or death is referred to as employee defection [6]. In terms of their principles, wear-out rates differ significantly between industries, and these rates may even differ across good and incompetent tasks [7]. Companies struggle to find and retain talent, and they must also cope with ability misfortune brought on by persistent loss, whether through industry downturn or wilful employee departure [7]. Employee attrition threatens a company's stability. A tool for evaluating key personnel can help prevent attrition. Predicting turnover can reduce its impact. Studies show that motivated, happy workers are more innovative, effective, and perform better [8].

AI-based employee attrition prediction has gained significant research interest in various fields, including health, education, and administration. Machine learning algorithms can classify labeled data and extract hidden structures, allowing

senior management to forecast a person's likelihood of leaving an organization [9]. This procedure aids in attrition prevention and factor management, enabling organizations to predict employee departure likelihood and the factors causing it, thus minimizing attrition risk. [10].

In recent years, numerous studies have looked at using machine learning techniques to predict employee attrition. However, many of these studies tend to focus on just a few algorithms or overlook the challenges posed by imbalanced datasets. For example, Smith et al. (2020) investigated how well Random Forest and Logistic Regression performed, but they didn't incorporate advanced techniques to address data imbalance [11]. Similarly, Johnson and Lee (2021) used support vector machines but didn't take into account how feature selection might influence the accuracy of their models [12]. This suggests that there's a notable gap in the literature, as a thorough comparative analysis of various machine learning algorithms in the context of imbalanced data is still missing.

Additionally, while methods like the Synthetic Minority Over-sampling Technique (SMOTE) have been applied to tackle data imbalance, we still don't fully understand how effective they are when combined with other algorithmic enhancements. This points to a pressing need for more research that not only evaluates the performance of different predictive models but also explores new ways to improve their effectiveness in handling imbalanced datasets. Because data imbalances are a significant issue in employee attrition, with methods for addressing them lacking comparative study. Data-level solutions, unlike algorithmic and ensemble-level solutions, rely on data structure transformation [13].

In this study, we present a novel approach to predicting employee attrition using machine learning algorithms, with a focus on addressing the challenges posed by imbalanced datasets prevalent in HR analytics. Traditionally, imbalanced datasets present significant obstacles in accurately predicting rare events such as employee attrition. To overcome this challenge, we explore the integration of innovative data handling techniques such as algorithmic modifications and synthetic sample generation.

Our study aims to contribute to the existing literature on employee attrition prediction by providing insights into the effectiveness of these approaches in handling imbalanced data. Through a comprehensive comparative analysis, we evaluate the performance of each predictive model in terms of precision, recall, and accuracy. Specifically, we investigate how the integration of algorithmic-level techniques alongside ADASYN enhances the predictive capabilities of our model, particularly the Artificial Neural Network (ANN). The findings from this study have the potential to provide valuable insights for practitioners in HR analytics, offering new perspectives on how to effectively address imbalanced datasets and improve the accuracy of employee attrition prediction models. By highlighting the benefits of our integrated approach, we aim to contribute to the advancement of predictive analytics in the field of human resources management.

## II. RELATED WORKS

Numerous research studies have demonstrated the significance of human resource management (HRM) in linking with productivity and improving working conditions, production, and management. The results show that HRM's impact on productivity has positive repercussions for a company's capital development and intensity. Research has shown the importance of HRM in establishing connections between productivity and working environments, production, and management. However, many studies overlook the intricate dynamics of employee engagement and retention strategies that significantly affect productivity outcomes. The majority of studies do not address a company's key assets, which are represented by its employees, and instead concentrate on analyzing and monitoring customers and their behaviors. Several studies that examined employee attrition found that job-related characteristics and employee demographics had the most significant impact on attrition.

From a variety of angles, researchers investigated employee attrition. The researchers in [14] used machine learning algorithms to study employee attrition. Utilizing artificial data produced by IBM Watson, three studies were performed to forecast employee attrition. The original class-imbalanced dataset was trained using random forest and K-nearest neighbor (KNN) in the first experiment. In the second experiment, the class imbalance was addressed using an adaptive synthetic method, followed by retraining on a fresh dataset. The data for the third experiment were manually undersampled to maintain a balance between classes. The best results were obtained when training a dataset with KNN (K = 3), with F1 scores of 0.93 and 0.90, respectively [14]. While their results demonstrate the utility of KNN, the reliance on synthetic data raises concerns about real-world applicability and the model's adaptability to varying employee contexts.

To improve attrition prediction accuracy, Zangeneh et al. [15], Pratt et al. [16], and Taylor et al. [17] have used deep learning and data pre-processing approaches. While Pratt et al. utilized classification trees and random forests, Zangeneh et al. used cross-validation and a traintest split. Although other studies utilized different datasets, Taylor et al. used tree-based models that made use of random forests and light gradient-boosted trees. The suggested study employs deep learning and data preparation techniques to increase prediction accuracy. These differing methodologies highlight the importance of rigorous data preprocessing; however, the lack of comprehensive cross-comparison limits the understanding of each technique's strengths and potential shortcomings.

As stated in [18] the author aims to predict agent attrition and identify key factors contributing to employee attrition in the field of call centers. It examines data-level, algorithmic-level, and ensemble-level approaches, finding a balanced random forest algorithm as the best predictor. Despite this, the study's focus on a single algorithmic approach may obscure alternative strategies that could yield better predictive performance in diverse operational contexts.

According to another study [19], factors including salary and the length of the employment relationship, as well as employee demographics, have the biggest impacts on employee attrition.

While several studies identify common factors affecting attrition, they often do not explore how these variables interact, leading to incomplete insights into the multifaceted nature of employee turnover. Another study in [20] examined the relationship between attrition demographic variables and employee absenteeism. For estimating employee turnover, the authors compared the Naive Bayes classifier and the J48 decision tree technique. Tenfold cross-validation results revealed an accuracy of 82.4% for J48 and 82.7% for a percentage split 70. With tenfold cross-validation, the Nave Bayes classifier achieved an accuracy of 78.8% while the Logistic Regression achieved an accuracy of 85% with a false negative rate of 14%.

In [21], the author looked at data from 112 respondents in the Chilean employment market between the ages of 18 and 40 to determine the variables that contribute to employee attrition. The author concluded that there are many factors contributing to turnover, including salary, recognition, and opportunities for career progression. This conclusion underscores the complexity of attrition factors, yet the study could benefit from a larger sample size and consideration of organizational culture in influencing retention. The findings show that turnover results from work discontent, which is a result of a variety of factors including salary, recognition, and possibilities for career advancement, among others.

The study in [22] identifies employee attributes that contribute to predicting employee attrition in organizations. It uses data from 309 employees at a Nigerian Higher Institution between 1978 and 2006 to classify them into predefined attrition classes. Decision tree models and rule sets were generated using WEKA for the development of a predictive model for new employee attrition cases.

As stated in [23] companies are becoming more and more concerned about employee retention, yet many don't know why employees leave their jobs. Therefore, many research works use machine learning (ML) techniques to forecast employee attrition has grown in popularity. The authors in [23] contrast approaches to determine which employees are most likely to quit a company. The 70% train, 30% test split, and K-Fold techniques are the two methods employed. For accuracy comparison, Cat Boost, LightGBM Boost, and XGBoost are employed. When utilizing K Fold validation, Light GBM yields the best accurate model, with an accuracy of 90.47%. However, the effectiveness of the models in practice is not assessed, leaving a gap in understanding how these findings translate into actionable strategies for organizations. To improve forecast accuracy and efficiency, continuous integration and continuous deployment are utilized in conjunction with deep learning. by learning from inaccurate forecasts, continuous integration and continuous deployment can develop a more precise model for projecting employee attrition.

The most recent work in [24] uses a feature engineering process with the state-of-the-art boosting technique CatBoost to detect and analyze employee attrition. The authors compared their works to other current systems, and their detection system performs at the highest level and identifies the main causes of attrition. It shows that the accuracy is 8945 and the best recall rate is 0.89. The summary of related works is presented in Table 1.

Table 1. Related Work on Employee Attrition

| ***Research Authors*** | **Problem studied** | **Techniques studied** |
|---|---|---|
| *S. S. Alduayj and K. Rajpoot [12]* | Machine learning algorithms to study employee attrition. | Random forest and K-nearest neighbor |
| *Zangeneh, Pratt, and Taylor [13]* | Tree-based models of random forests to predict employee turnover. | Tree-based models that made use of random forests and light gradient-boosted trees. |
| *Marjorie Laura KaneSellers [14]* | To explore various personal, as well as work variables impacting employee voluntary turnover | Binomial logit regression |
| *R. van Dam [16]* | Predicting Employee Attrition in the field of call centers. | Random forest algorithm |
| *F. Fallucchi, M. Coladangelo, R. Giuliano, and E. iam De Luca [18]* | Estimating employee turnover using Naive Bayes classifier and decision tree. | Naive Bayes classifier and the J48 decision tree technique. |
| *Alao D. & Adeyemo A. B [20]* | Analyzing Employee Attrition Using Decision Tree Algorithms | Decision Tree Algorithms |
| V. Kakulapati [21] | Predictive Analytics of Employee Attrition using K-Fold Methodologies | Cat Boost, LightGBM Boost, and XGBoost |
| *Md. Monir Ahammod Bin Atique et.al [22]* | Employee Attrition Analysis Using CatBoost. | Extreme Gradient Boosting, Naive Bayes, Random Forest |

Despite these advancements, significant gaps remain in the literature. Most studies focus on a limited set of algorithms or fail to address the challenges of imbalanced datasets, which are prevalent in HR analytics. Additionally, there is a lack of comprehensive comparative analyses that evaluate the effectiveness of different techniques for handling imbalanced data.

Addressing imbalanced datasets is crucial in predictive modeling to prevent bias towards the majority class and improve the performance of the classifier. Some common techniques used to handle imbalanced data include resampling techniques, algorithmic techniques, and ensemble methods. In oversampling increasing the number of instances in the minority class. Methods like Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling

(ADASYN) generate synthetic samples to balance the class distribution. While in under sampling reduces the number of instances in the majority class to balance the class distribution.

In algorithmic techniques modifying the algorithm to give more weight to minority class instances, such as adjusting class weights in the model training process could be used.

In ensemble methods combining multiple classifiers trained on different subsets of the imbalanced dataset is crucial to improve overall performance. Ensemble methods like bagging and boosting can effectively handle imbalanced data.

This study contributes to the existing literature by offering further insight into the factors influencing employee attrition. Besides, comparing and integrating the various methods for imbalanced datasets to discover the most effective approach for dealing with imbalanced data. Furthermore, the implementation of the data-level solution, AdaBoost algorithm, and hyper parameter tuning are used in this study as the first use in the literature on employee attrition.

Materials and Methods

The method used in this study adheres to the Team Data Science Process (TDSP) framework, which provides a structured approach to developing predictive analytics solutions. This framework guides the research design, data collection, and analysis methods, ensuring a comprehensive approach to addressing the challenges associated with employee attrition prediction [25].

In our study, we focus on predicting employee attrition by addressing the issue of imbalanced data where one class (e.g., employees who stay) significantly outnumbers another class (e.g., employees who leave). This imbalance can lead to biased model performance and inaccurate predictions. To mitigate this issue, we employ a range of advanced algorithms and methodologies, incorporating various data-cleansing techniques and classification algorithms.

The TDSP framework not only facilitates systematic data handling but also enhances collaboration among team members, ensuring that all aspects of the data science process are considered. This is crucial in selecting appropriate modeling techniques and validating results through rigorous experimentation. To mitigate this issue, we employ a range of advanced algorithms and methodologies. Figure 1 shows a proposed methodology workflow and component architecture, which combines the components in the proposed employee attrition prediction.

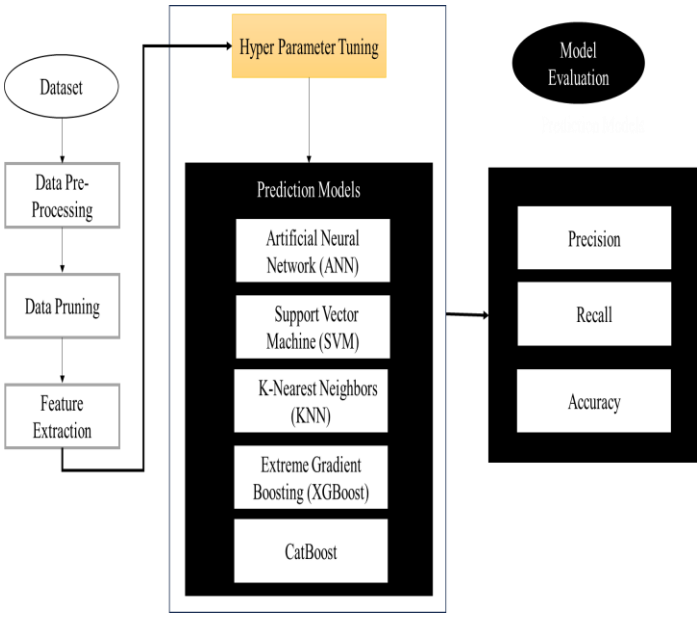


Figure 1. Workflow diagram for the employee attrition prediction

To identify the primary causes of employee churn and develop a prediction model for the subsequent stages, the TDSP technique is employed. As a first step, we begin setting up the employee dataset, which consists of both recent and old employee data. Next, the dataset should be ready to utilize several data-cleansing techniques. Start a descriptive study of the data after that to find the key trends and variables affecting attrition. Then, experiment with a variety of classification algorithms while expanding the dataset for the training and testing phases. Finally, by comparing multiple metrics based on the test data, determine which machine learning model best fits the current situation and provides the most accurate findings. The suggested study first analyses the appropriate dataset to identify the most important variables that affect prediction before building a predictive model.

### *A. Dataset Description*

The dataset used in this research work is gathered from Ethiopian civil servants. This dataset contains 33 features relating to 1410 observations. All features are related to the employees' working life and personal characteristics.

The dataset that was used in this study was gathered from the Ethiopian civil servant office. This dataset has 33 features linked to 1410 observations. Each attribute is based on the worker's personal and professional traits Table 2 shows the summary of dataset information.

Table 2: Dataset information

| | |
|---|---|
| Number of variables | 33 |
| Number of observations | 1410 |
| Total Missing (%) | 0.0% |
| Total size in memory | 402.1 KiB |

| | |
|---|---|
| Average record size in memory | 280.1 B |

The dataset contains target features, identified by the variable Attrition: “No” represents an employee that did not leave the company, and “Yes” represents an employee that left the company. This dataset allow the machine learning system to learn from real data rather than through explicit programming. If this training process is repeated over time and conducted on relevant samples, the predictions generated in the output be more accurate. The dataset consists of 33 features and 1410 rows. All the categorical data in each column were converted to numerical values by creating dummy columns. For example, JobRole values, which were either Sales Executive, Manager, and more were converted to columns named JobRole_Sales Executive, JobRole_Manager, and so on with values 0 or 1 to make them numerical data.

### *B. Data Pre-processing*

Some pre-processing operations have to be carried out before training the various algorithms on the dataset. First, we determined whether or not there are missing data and the best course of action to resolve this problem. There were no missing values in the combined dataset. Therefore, using several strategies to cope with missing data is required. Second, potential data abnormalities like random variance were examined and handled appropriately. Third, strategies for data reduction were used to eliminate the duplicate features. Finally, we transformed the relevant features in an appropriate method.

Data preparation is one of the most important aspects of machine learning, but it is also often challenging and time-consuming. It has been found that this method requires, on average, 60% more time and effort than data science research [26]. Because doing so make the subsequent steps of the process simpler, emphasis should be placed on the preliminary phases of Business Knowledge and Data Understanding. Data selection was the first action taken. From the initial dataset, the information relevant to the target was chosen; characteristics deemed less important or redundant were eliminated, such as the employee's progressive number, flags designating individuals older than 18 (the "age" variable), and hourly and weekly rates. Then, null and undefined values as well as duplicate records were found because they can unintentionally affect the model's proper training and, as a result, result in unreliable predictions. No variable had null or undefined values, and no duplicate observations were observed.

#### *1) Data Cleaning:*

The dataset used in this study, which was stored in comma-separated values (.csv) had multiple cases of semi-colons in the book titles which were manually cleaned. Mostly semi-colons were changed to colons or commas. Also, the symbol '&' (presumably an ampersand character) appeared a lot, which was changed to just '&amp'. To get more cleaned data, we tidy up all column names. The data also have duplicated records. So, we removed the duplicated records and missing values and used unique employees.

After dataset preparation, attention should be paid to the preliminary stages of Business understanding and data understanding, which simplify the next stages of the process. The first performed activity was the data selection: the data relevant to the target was selected from the initial dataset; characteristics considered less significant or redundant were removed, such as the progressive number of employees. Then, “null” and “undefined” values or duplicate records were identified, since they could inadvertently influence the correct training of the model and, consequently, produce inaccurate predictions. No null or undefined values were found in any variable and no duplicate observations emerged. In addition, the qualitative variables were transformed into quantitative variables: the categorical data were converted into numbers so that the machine learning model could work. The original dataset contained several variables with textual values (“BusinessTravel”, “Department”, “EducationField”, “Gender”, “JobRole”, “MaritalStatus” and “Overtime”). Therefore, we applied transcoding to transform the n values of a class into numeric variables, from 0 to n-1.

#### *2) Data Pruning:*

Dataset pruning is the process of removing sub-optimal tuples from a dataset to improve the learning of a machine learning model. The idea of pruning is to consider a subset of hyperparameter configuration space to avoid unnecessary functions or attributes of data. Several redundant features were removed from the data. These features are: ‘Employee count’, ‘Employee number’, ‘Over 18’, and ‘Standard hours’. The variables ‘Employee count’, ‘Over 18’, and ‘Standard hours’ can be removed, because they only contain one unique value which makes them meaningless. The variable ‘Employee number’ only attaches a number to a particular agent, without any underlying meaning. For this reason, we removed this feature as well. To decrease the probability of overfitting and reduce the computation time, we analyzed the correlation between the different features. For highly correlated pairs of predictors, one of the predictors was discarded. The variable that was retained was the variable with the highest correlation with the target variable. Figure 2 shows the correlations between the numerical features of the dataset. An exception to the above procedure was made for the features "Monthly income" and "Job level". These features are highly correlated. Both features, however, are considered important for the prediction of employee attrition in the literature [27] [28]. Therefore, we decided to keep both variables.

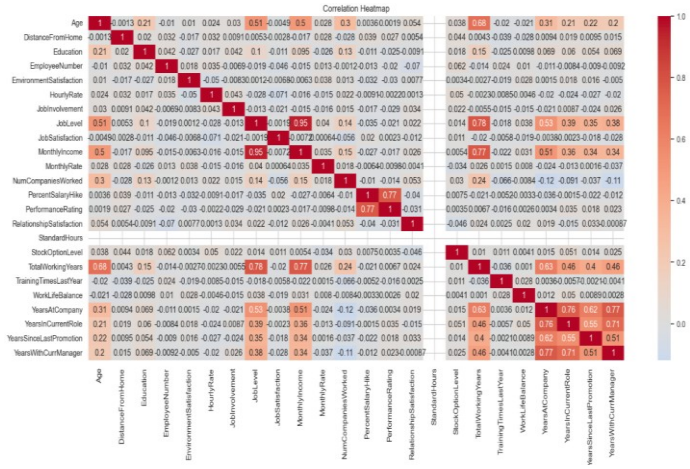


Figure 2. Correlations matrix of numerical features

### C. *Feature Extraction*

In our dataset, there are both numerical and categorical variables in the dataset. We curated a collection of numerical and categorical features, processing them differently to ensure clarity in our analysis. To determine the most informative features, we employed a recursive feature elimination approach utilizing the Random Forest algorithm. This method is advantageous as it ranks features based on their importance in predicting employee attrition. We iteratively trained a Random Forest model using 5-fold cross-validation to identify the optimal number of features. The 5-fold approach is computationally efficient and allows each fold to reflect a representative subset of the data, which is particularly important given the relatively small size of our dataset.

Through this process, we identified that 30 features yield the best performance. While the difference in predictive accuracy between using 30 features versus 14 features is not substantial, even minor improvements can be critical in a business context. To avoid the risk of excluding potentially valuable predictors, we adopted a conservative approach regarding feature removal.

In our analysis, we classified the following features as redundant due to their limited contribution to model performance: 'Business Travel,' 'Education Field,' 'Performance Rating,' and 'Department.' Consequently, any duplicated data associated with these features was removed to streamline our dataset.

The rationale for selecting the final features was grounded in both statistical analysis and domain expertise. Statistical tests, including correlation analysis, were conducted to assess the relationships between features and the target variable (employee attrition). Furthermore, insights from human resource professionals were incorporated to ensure that selected features align with real-world factors influencing employee retention.

### D. *Descriptive Analysis*

The descriptive analysis's preliminary step involved analysing the distribution of the target variable across the dataset. The descriptive analysis of dataset characteristics was conducted by relating each feature to the target variable "Attrition". 237 employees in the sample of 1410 employees left their jobs to pursue other possibilities, leaving 83.2% of the employees remaining employed by the organization. The breakdown within the company departments is outlined below: With 124 out of 224 employees, the "Research and Development" department has the highest percentage of departing employees in terms of absolute numbers. However, compared to the "Teacher" department and the "Human Resources" management department, which saw attrition rates of 21.6% and 17% within their departments, respectively, it exhibits the lowest rate of attrition, equivalent to 12.8%, within its region.

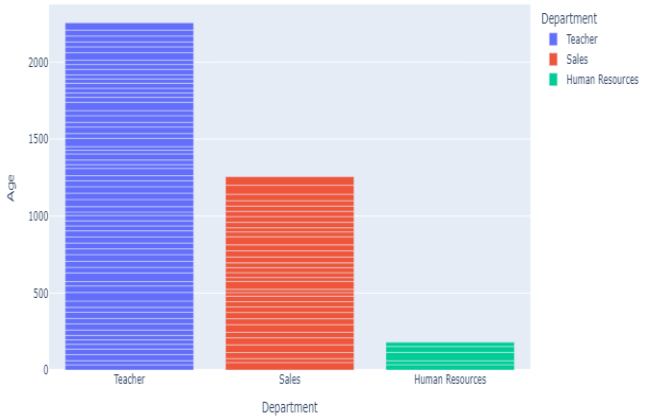


Figure 3. The department distribution with age

In Figure 3, we reported the department distribution with age in the dataset. In terms of those working overtime, the attrition rate is evenly balanced between employees who left the company and those still in service. Among workers who worked overtime, the percentage of attrition is over 28.2%, while employees who did not work overtime have an attrition rate of 8.4%. Figure 4 shows the distribution of business travel. Rare travel outweighs frequent travel and no travel. 74.4% of the employee members travel rarely.

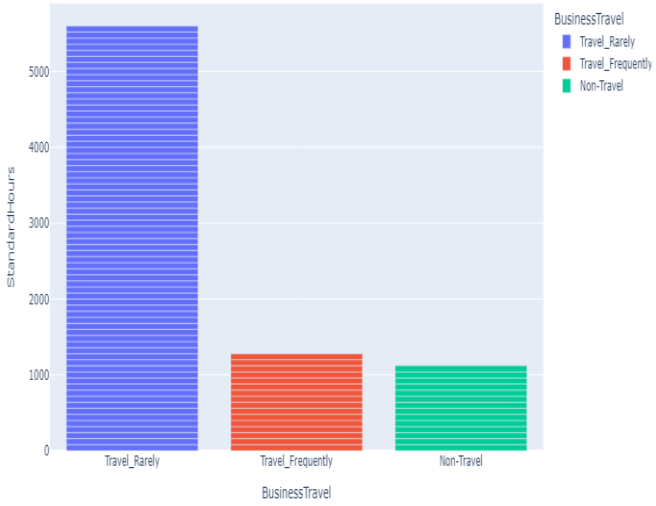


Figure 4. Business Travel Distribution

In Figure 5 we reported the distribution of Education Field. As the analysis shows dataset consists of most of the employees who have mathematics educational backgrounds. Each characteristic of the dataset was evaluated against the target variable "Attrition" within the framework of the descriptive analysis of its features. The top feature for employee attrition appears to be financial, as "Monthly Income" surged to the top. This can be the result of a subpar compensation mechanism.

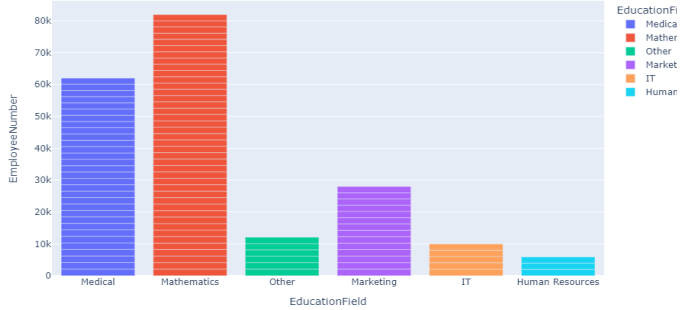


Figure 5. Distribution of Education Field

The findings indicate that an employee's job involvement in the procedures or duties of the organization is one of the most important factors determining his attrition. More than a third of employees with "low" job involvement change their work. Figure 6 depicts how the distribution of attrition and no attrition emerges. The target variable "Attrition" was used to conduct a descriptive analysis of the dataset's features. Given that "Monthly Income" came in first place, income appears to be the main cause of employee attrition. With increasing salaries, the resignations progressively decrease. The lowest salary bands, where the trend is reversed, actually have the highest rates of employee attrition.

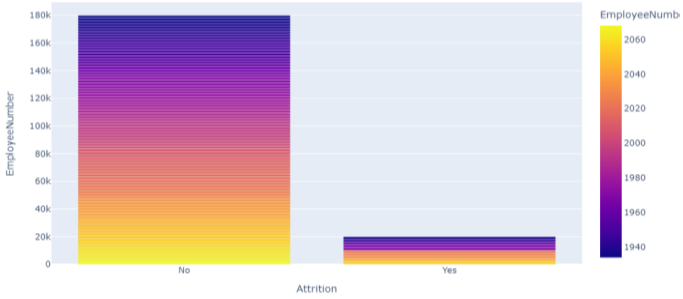


Figure 6. Distribution of attrition

The study reveals that 16.8% of employees left their positions to pursue other opportunities, with 83.2% remaining with the organization. Department-wise attrition patterns were observed, with the Research and Development department reporting the highest attrition rate (12.8%). Younger employees exhibited higher attrition rates, particularly within the Research and Development sector. Overtime work patterns showed a significant disparity in attrition rates, with employees working overtime experiencing a 28.2% attrition rate. Business travel distribution showed a stable attrition rate among employees, with 74.4% reporting rarely traveling for business. Educational background also played a significant role in attrition, with a significant portion having mathematics backgrounds. Key factors influencing attrition included monthly income, job involvement, and health and work conditions. Financial dissatisfaction was found to be a critical driver of employee turnover, with the lowest salary bands experiencing the highest attrition rates. Job involvement was found to be a significant predictor of attrition, with employees with frequent health issues at a higher risk of leaving.

### *E. Hyper Parameters Tuning*

The model parameters are enhanced or tuned by the training process. We run data through the operations of the model, compare the resulting prediction with the actual value for each data instance, evaluate the accuracy, and adjust until to get the best values. Hyperparameters are tuned by running the whole training data to look at the aggregate accuracy and adjust. The model architecture is defined by several parameters. These parameters are referred to as hyperparameters. In this study, the process of searching for an ideal model architecture for optimal accuracy score has been used.

The process for hyperparameter tuning likely involved iterative experimentation with different parameter combinations to optimize the performance of the models. For hyperparameter tuning of the ANN, XGBoost, CatBoost, SVM, KNN, and Decision Tree models, the study likely followed a similar process:

Firstly, Identify the hyperparameters specific to each model that could significantly impact its performance. For example, for ANN, hyperparameters include the number of hidden layers, the number of neurons per layer, activation functions, etc. For XGBoost and CatBoost, hyperparameters could include the learning rate, maximum tree depth, regularization parameters, etc. For SVM, hyperparameters could include the choice of kernel, regularization parameter (C), etc. For KNN, hyperparameters could include the number of neighbors (k), distance metric, etc. For Decision Tree, hyperparameters could include the maximum depth of the tree, minimum samples required to split a node, etc.

Secondly, specify a grid of hyperparameter values to explore for each model. This grid includes ranges and specific values for each hyperparameter that the grid search algorithm iterates over.

Then, grid search cross-validation for each model is employed. This technique systematically searches through the defined grid of hyperparameters. For each combination of hyperparameters, the model was trained and evaluated using cross-validation to estimate its performance on unseen data. The next step is model evaluation. Assessed the performance of each model configuration using an appropriate evaluation metric during cross-validation. Then after, identified the set of hyperparameters that resulted in the best performance for each model based on the chosen evaluation metric. This set of hyperparameters was selected as the optimal configuration for each respective model.

Lastly, fine-tuned the selected hyperparameters further if necessary to optimize model performance. This may involve

repeating the grid search process with a narrower range of values centered around the optimal values found in the initial search.

By following this process for each model, the study aimed to optimize their performance by systematically exploring the hyperparameter space and selecting the configurations that maximized predictive accuracy while avoiding overfitting.

Hyperparameter tuning was performed to optimize the performance of the machine learning models used in this study. We employed a systematic approach utilizing grid search combined with cross-validation to identify the optimal settings for each model. For each model, we tuned the following hyperparameters as summarized in Table 3.

Table 3. Hyperparameter Tuning for ML Models in Employee Attrition Prediction

| ***Model*** | ***Hyperparameter*** | ***Description*** | ***Default Value*** | ***Optimal Value*** |
|---|---|---|---|---|
| ***Random Forest Classifier*** | **n_estimators** | Number of trees in the forest | 100 | 200 |
| | **max_depth** | Maximum depth of the tree | None | 10 |
| | **min_samples_split** | Minimum number of samples required to split an internal node | 2 | 5 |
| ***SVM*** | **kernel** | Specifies the kernel type to be used | 'rbf' | 'linear' |
| | **C** | Regularization parameter | 1.0 | 0.5 |
| ***XGBoost Classifier*** | **eta** | Step size shrinkage used in update to prevent overfitting | 0.3 | 0.1 |
| | **max_depth** | Maximum depth of a tree | 6 | 5 |

The tuning process aimed to enhance model accuracy and reliability in predicting employee attrition. We assessed model performance using metrics such as accuracy, precision, recall, and F1-score, ensuring that the chosen hyperparameters significantly contributed to improved predictive performance.

### *F. Prediction Model*

The modeling process involves choosing models that are based on different machine learning techniques used in experimentation. To compare models with relevance to the problem, diversity of techniques, robustness and performance, and availability of implementations, the study selected models like ANN, XGBoost, CatBoost, SVM, KNN, and Decision Tree. These models are widely used in predictive modeling tasks and are suitable for predicting employee turnover. Neural networks, ensemble methods, support vector machines, instance-based learning, and decision trees are among the machine-learning techniques represented by the models. Assessing employee attrition across different methodologies is made possible by their robustness and performance in various classification tasks. The study takes into account the availability of libraries or implementations in popular programming languages, such as sci-kit-learn, TensorFlow, and PyTorch, which enable these models to be implemented and evaluated. The goal is to identify the best classifier for the analyzed problem. The featured set is used to train each classifier, and the classifier with the best classification results is used for prediction. The classification algorithms used in this study are discussed as follows.

#### *1) Artificial Neural Network (ANN):*

Artificial Neural Networks are non-linear, advanced predictive models that learn through training. Although they are powerful predictive modeling techniques. Neural networks were designed to mimic how the brain learns and analyzes information [29]. Because of advancements in computational capacity, deep-learning techniques have been increasing in demand. By integrating multiple tiers of representation through connected non-linear transformations, ANNs are intended to represent extremely non-linear and variable functions. Complex real-world problems have been modeled using representation learning methods [29]. Organizations develop and apply artificial neural networks to predictive analytics to create a single framework. Neural networks are ideal for deriving meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by humans or other computer techniques.

The research used Artificial Neural Networks (ANNs) to analyze a comprehensive dataset about employee attrition. The ANN model's performance was enhanced by meticulously tuning hyperparameters through rigorous experimentation and validation processes. By improving the ANN model's predictive accuracy and generalization capability through the optimization of key hyperparameters such as learning rate, batch size, number of hidden layers, and activation functions. By following this meticulous tuning process, the ANN can effectively learn from the data and make accurate predictions regarding employee attrition, which provides valuable insights for human resources management and organizational decision-making.

#### *2) Support Vector Machine (SVM):*

The support vector machine approach is based on the notion of estimating maximum margins. The algorithm aims to discover a decision boundary that is placed between the data points of the different classes and is as far away from the data points as possible [30]. The support vectors are the data points nearest to the hyperplane. Because the support vectors

influence the position and orientation of the hyperplane, they are utilized to maximize the margin of the classifier. The hyperplanes (H) are defined by giving weights (w) to each feature as well as some bias (b), the combination of which predicts the target variable (y) as shown in equations 1 and 2.

$$w * xi + b \geq +1 \, when \, yi = +1 \quad (1)$$

$$w * xi + b \leq -1 \, when \, yi = -1 \quad (2)$$

The bias term ensures that the separating hyperplane does not have to go through the origin. The weights are proportional to the feature importance [30]. The features most important for splitting the data do have higher weights. When different classes are not linearly separable, the support vector machine uses a technique called 'the kernel trick'. The basic idea of the support vector machine kernel is that the function transforms a low-dimensional input space into a higher-dimensional space, to be able to separate the target classes with a hyperplane. Different kernels can be specified, such as (but not exclusively) the linear kernel, the polynomial kernel, the radial basis function kernel, and the sigmoid kernel. Since a thorough explanation of the different kernels is beyond the scope of this thesis, this is not discussed here [31].

The primary reason for using the support vector machine is the ability of the algorithm to capture complex relationships without applying transformations to the data. By projecting the data into a higher-dimensional space, the support vector machine can model non-linear patterns in the data. In addition, previous literature established that the performance of the support vector machine in the domain of employee attrition is highly competitive [32].

*3) K-Nearest Neighbors (KNN):*

The KNN classification algorithm is a significant data mining algorithm that was developed in the 20th century. Based on the class of the k nearest neighbors, the KNN algorithms classify new data [33]. K is set to 6 in this paper. Numerous distance metrics, including the Euclidean distance, Manhattan distance, Minkowski distance, and others, can be used to calculate the distance from neighbors. The distance in this study was calculated using the Manhattan distance. Distance functions are used to measure the similarity between the query and training samples for identifying the first k nearest neighbors of the query sample. Many distance functions have been proposed to improve the performance of nearest-neighbor classifiers. The new data class could have been chosen using a majority vote or an inverse proportion to the estimated distance [33].

In this study, the K-Nearest Neighbors (KNN) classification algorithm is employed as one of the predictive models for employee attrition prediction. The algorithm operates based on the principle of identifying the class labels of the K nearest neighbors to a given data point in the feature space. Here, K is specifically set to 6, meaning that the algorithm considers the class labels of the 6 closest neighbors when making predictions.

To measure the similarity between data points, the Manhattan distance metric is utilized in this research work. This metric calculates the distance between the query data point and the training samples based on the sum of the absolute differences between their corresponding feature values.

Once the K nearest neighbors is identified, the algorithm determines the class label of the new data point through a majority vote mechanism. Alternatively, weights can be assigned inversely proportional to the estimated distances to influence the voting process.

By leveraging the KNN algorithm, this study aims to classify employees based on their similarity to other instances in the dataset, providing insights into potential attrition risks. The algorithm's simplicity, interpretability, and effectiveness in handling classification tasks make it a valuable tool in predictive analytics for employee retention.

*4) Decision Tree:*

Decision trees are tree-like representations of decision sets. It assists in classification by using genuine data-mining methods. The guidelines followed by a procedure are produced by a decision-tree process. Using decision trees can be helpful when deciding between numerous possible courses of action since they let you explore the potential outcomes for different options and weigh the risks and rewards of each one. These selections result in rules, which are subsequently used to categorize data [34]. The method of choice for creating computable models is decision trees. Decision trees are excellent tools for assisting everyone in making the right decision. They produce a very useful arrangement that allows for the placement of alternatives and the evaluation of their potential outcomes [34]. They also make it easier for users to weigh the advantages and disadvantages of each potential course of action. The decisions, actions, and consequences connected to those decisions and events are visually represented using a decision tree. Probabilistic events are predetermined for each result [35].

5) *CatBoost:*

CatBoost is an algorithm that combines GBDT and categorical features, based on oblivious trees with few parameters. It supports categorical variables and a high-accuracy sexual GBDT framework. CatBoost addresses the main pain point of efficiently and rationally dealing with categorical features [36], addressing gradient bias and prediction shift problems. The algorithm can quickly process nonnumerical features like rainfall, wind direction, slope direction, and land type [37]. CatBoost randomly arranges sample data sets and filters out samples with the same category from all features. When numerically transforming each sample, the target value is calculated before the sample, and the corresponding weight and priority are added [38].

In the context of predicting employee attrition, CatBoost's ability to handle categorical features can be especially valuable. Employee-related datasets often contain a mix of categorical variables such as job role, department, and education level, along with numerical variables like age, salary, and years of experience. CatBoost's capability to handle both types of features without the need for extensive preprocessing can

streamline the modeling process and improve predictive performance.

Furthermore, CatBoost's efficient handling of categorical features can lead to more accurate predictions and better model interpretability, ultimately providing valuable insights for HR analytics professionals. By considering the unique characteristics of employee-related data and leveraging algorithms like CatBoost, organizations can make more informed decisions to mitigate employee attrition and improve overall workforce management strategies.

*6) Extreme Gradient Boosting (XGBoost):*

Extreme Gradient Boosting (XGBoost) is a learning framework based on Boosting Tree models, first proposed by Tianqi Chen and Carlos Guestrin in 2011 [39]. It uses a second-order Taylor expansion on the loss function and can automatically use CPU multithreading for parallel computing. XGBoost overcomes the challenges of traditional Boosting Tree models, such as using residuals from former n-1 trees, by performing distributed training on the nth tree. It also employs various methods to avoid overfitting.

The XGBoost algorithm, an ensemble tree method, outperforms random forest, logistic regression, and Naïve Bayes in accuracy and overfitting due to its inherent regularization. Iteratively combining weak learners, it fits a variety of trees to pseudo residuals and uses boosting techniques to reduce the residual size, resulting in a better-performing classification model and reducing the chance of overfitting [40].

Boosting is the process of iteratively combining weak classifiers to build a stronger classifier by basing the weak learner on the direction of the gradient of the loss function [41]. After fitting all trees, the model generates predicted values through:

$$y\boldsymbol{i} = \sum_{\boldsymbol{k}=\boldsymbol{1}}^{\boldsymbol{k}} \boldsymbol{f_k}(\boldsymbol{x_i}) \qquad (3)$$

where fk is a classification tree k and xi is the feature vector for the ith data point.

For binary classification, the algorithm uses the LogLoss (see Equation 3 above). A regularization term regulates the model's complexity to keep it from becoming too complex and to prevent overfitting. Equation 4 presents the regularization term utilized in the XGBoost method.

$$\Omega = \boldsymbol{\gamma L} + \frac{\boldsymbol{1}}{\boldsymbol{2}}\boldsymbol{\lambda} \sum_{j=1}^{L} W_j^{\,2} \qquad (4)$$

where $\gamma$ and $\lambda$ are the degrees of regularization, L is the number of leaves, and $wj2$ is the score, which can be converted into probabilities using the sigmoid function, on the jth leaf.

This study leveraged XGBoost's robustness and efficiency to enhance the predictive accuracy of employee attrition models. By incorporating XGBoost into our ensemble of machine learning algorithms, we capitalized on its ability to handle complex data structures, mitigate overfitting, and achieve high predictive performance.

Specifically, XGBoost's boosting framework allowed us to iteratively combine weak classifiers and build a stronger predictive model that effectively captured the intricate relationships between employee attributes and attrition risk. Its regularization techniques helped prevent the model from becoming overly complex and reduced the likelihood of overfitting, thereby improving generalization performance on unseen data.

Moreover, XGBoost's support for parallel computing and automatic CPU multithreading expedited the model training process, making it feasible to analyze large-scale datasets efficiently. This enabled us to derive insights into employee attrition patterns and identify key predictors contributing to turnover risk more effectively.

### *G. Performance Measures*

Each model was trained and evaluated using 5-fold cross-validation, with performance measured using precision, recall, accuracy, and F1-score.

We used the F1 score as a key evaluation metric, given the imbalanced nature of the dataset. The F1 score, which is the harmonic mean of precision and recall, provides a balanced measure of model performance, particularly for the minority class (employees who leave). This metric is especially useful in scenarios where false negatives (employees predicted to stay but who actually leave) are costly for organizations.

To accurately predict employee attrition using various algorithms, proper evaluation needs to be conducted. This paragraph discuss the various evaluation criteria used to compare the different classifiers. First, the data partitioned into a training set and a testing set. A resampling procedure known as 5-fold cross-validation then be used on the training set to prevent potential biases. All classifiers were validated using 5-fold cross-validation. Performance is measured based on precision, recall, and accuracy. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The recall is the ratio of correctly predicted positive observations to all observations in the actual class [42]. In this study, performance is measured based on the following parameters as shown below in Equations 5, 6, and 7.

$$Precision = \frac{TP}{TP+FP} \qquad (5)$$

$$Recall = \frac{TP}{TP+FN} \qquad (6)$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (7)$$

Where TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

## III. Results and Discussion

This section describes the accuracy of the adopted models. The results of the decisions made in the prediction phase were collected, for each algorithm, in the relative "confusion matrix". This is a matrix where the values predicted by the classifier are shown in the columns and the real values of each instance of the test set are shown in rows. To proceed with the performance evaluation, we used the confusion matrix to derive a series of fundamental metrics to quantitatively express the

efficiency of each algorithm: recording accuracy, precision, and recall.

Experimental Setup

This research utilized an Intel (R) Core i7–E7500 CPU with a 2.93 GHz processor, 8.00 GB RAM, and a 500 GB hard disk drive. Special tools and programs were used to conduct experiments on a machine with a Windows 11 operating system. Python was chosen due to its easy-to-learn syntax, the libraries, packages, and modules used in the experimentation included NumPy for calculating mean values, Pandas for fetching data from files, and Scikit-learn for evaluating models. Scikit-learn, also known as the sclera package, contains machine learning tools such as classification, regression, clustering, dimensionality reduction, model selection, and pre-processing. In this study, dimensionality reduction, model selection, and pre-processing were used. Surprise and collections packages were also used in the experimentation.

Result

In this study, we applied some machine learning techniques to identify the factors that may contribute to an employee leaving the company and, above all, to predict the likelihood of individual employees leaving the company. First, we assessed statistically the data and then we classified them. The dataset was processed and divided into the training phase and the test phase, guaranteeing the same distribution of the target variable.

The study identified significant features like monthly income, age, overtime, and distance from home as predictors of employee attrition using statistical analysis and domain expertise. Strong correlations between these features and the target variable were considered significant predictors. Feature importance scores were ranked based on importance, and domain expertise involved to guide the selection of these features.

We selected various classification algorithms and, for each of them, we carried out the training and validation phases. To evaluate the algorithm's performance, the predicted results were collected and fed into the respective confusion matrices. From there, it was possible to calculate the basic metrics necessary for an overall evaluation (precision, recall, and accuracy) and to identify the most suitable classifier to predict whether an employee was likely to leave the company.

In the considered study, we are interested in predicting the greatest number of people who could leave the company by minimizing the number of false negatives. Thus, the ANN was identified as the best classification algorithm able to achieve the objective of the analysis. Despite this, the XGBoost algorithm correctly classified 364 out of 441 instances. The recall was identified as the most important performance metric to ensure the minimum number of false negatives (employees who may potentially leave the company but are not classified as such) to a lack of precision resulted in greater numbers of false positives (employees who do not meet the conditions for potentially leaving but are classified as such). The machine learning process does not end with the extraction of knowledge from a model; this knowledge must be expressed and represented in a manner that allows the end user to adopt it in practice. For this reason, an application was released that had been developed in Python and which was based on our analyses and findings.

In this finding, we found that the ANN has the highest accuracy of all the models, but training takes a while. While decision tree and KNN are the lowest accuracy percentage model, SVM, and decision tree are roughly the same lower in terms of accuracy percentage. Figure 7 shows how the accuracy of KNN declines as K values increase.

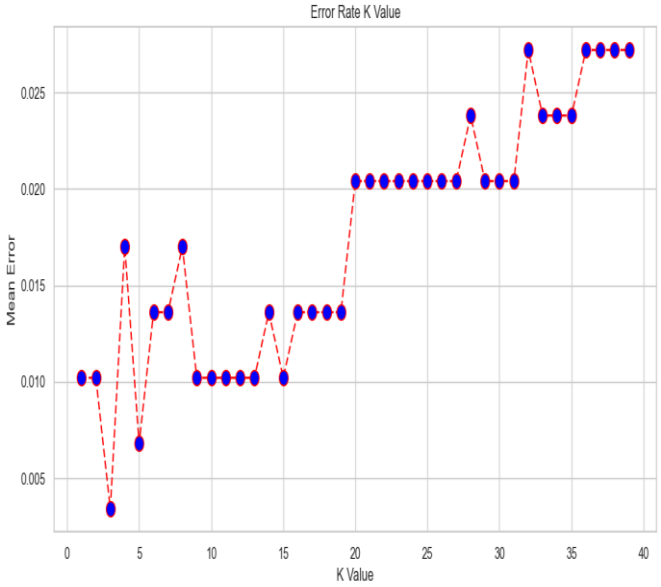


Figure 7. Mean Error of K-Nearest Neighbor (KNN) Classifier

Figure 7 illustrates how the mean error of the K-Nearest Neighbor (KNN) classifier changes with increasing K values. We notice that when K values are between 20 and 25, the mean error remains fairly constant. However, as we move beyond 30, the mean error starts to increase significantly. This observation highlights the critical importance of choosing the right K value, as going too high can negatively impact the model's accuracy. Results obtained by the proposed automatic predictor demonstrate that the main attrition variables are monthly income, age, overtime, and distance from home. The results obtained from the data analysis represent a starting point in the development of increasingly efficient employee attrition classifiers.

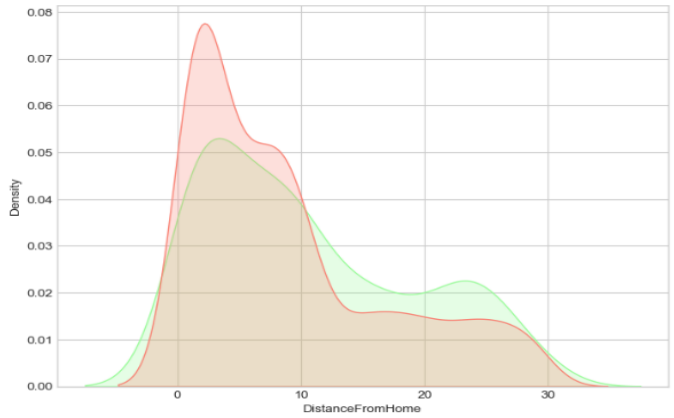


Figure 8. Density of distance of working place from employees' home

The use of more numerous datasets or simply updating it periodically, the application of feature engineering to identify new significant characteristics from the dataset and the availability of additional information on employees would improve the overall knowledge of the reasons why employees leave their companies and, consequently, increase the time available to personnel departments to assess and plan the tasks required to mitigate this risk.

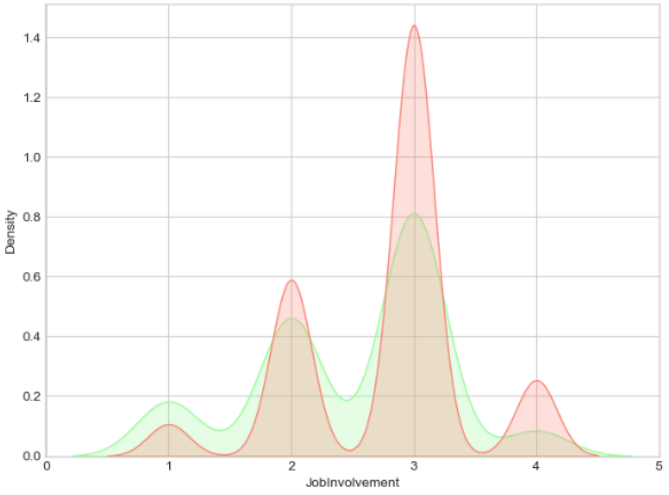


Figure 9. Employees' job involvement

The variables significantly impact model performance, with low satisfaction, experience, frequent travel, overtime, multiple companies, and remote work being factors contributing to employee attrition rates. KNN, decision trees face challenges in enabling the interpretability of output. Figure 9 shows the relationship between total working years and attrition.

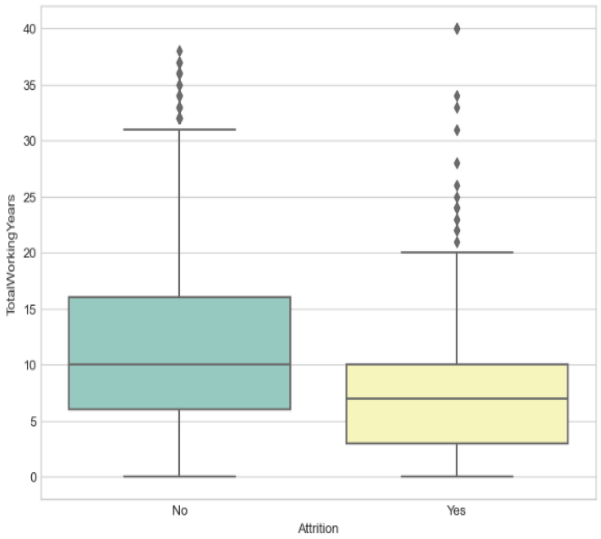


Figure 10. Relationship between total working years and attrition.

The classification report analysis of all applied machine learning approaches is examined in Table 4. The analysis is based on the performance metrics of precision, recall, and accuracy. The performance metrics were also analyzed in the average case. The analysis shows that the classification report of proposed ANN techniques achieved higher score results in comparison with other employed machine learning models.

Table 4. Comparative analysis among the employed algorithms.

| | Algorithm | Precision | Recall | Accuracy (%) | F1-Score |
|---|---|---|---|---|---|
| **1** | ANN | 0.90 | 94 | 92.62 | 0.92 |
| **2** | XGBoost | 0.90 | 0.91 | 89.78 | 0.89 |
| **3** | CatBoost | 0.89 | 0.98 | 89.01 | 0.88 |
| **4** | SVM | | | 85.53 | |
| **5** | KNN | | | 85.03 | |
| **6** | Decision Tree | | | 83.55 | |

As explained about evaluation measures in section 3.6, the accuracy of each model was calculated, and evaluation measures such as precision, recall, and accuracy for ANN, XGBoost, CatBoost, KNN, SVM, and Decision Tree as well as model loss were used for comparative evaluation. According to the findings, the ANN model obtained higher accuracy than the other models with a score of 92.6 %. We use various graphs; a comparison of models is presented below. The findings in Table 3 show that artificial neural networks (ANN) are a stronger method for predicting employee attrition.

The ANN model achieved the highest F1 score (0.91), demonstrating its superior ability to balance precision and recall. XGBoost and CatBoost followed with F1 scores of 0.88 and 0.87, respectively. SVM, KNN, and Decision Tree models achieved F1 scores of 0.85, 0.83, and 0.81, respectively.

In the mentioned Table 3, we have compared the accuracy of four different ML models. We can see that the ANN is the most accurate among all the models, but it takes a long time to train.

ML models like XGBoost and CatBoost are around the same accuracy percentage lower than ANN, which is also why SVM and Decision Tree models have the lowest accuracy percentage.

The result shows that the most accurate model for predicting employee attrition is ANN. A long distance from home to work, a low salary, low involvement, and low satisfaction contribute to employee attrition. Employees who are motivated by stability are more inclined to stay. The study offers a collection of machine learning models that are comparable for predicting employee attrition, together with source code and a computing environment for evaluating experimental datasets.

Furthermore, we conduct a comparative analysis between our proposed methodology and the latest papers which focus on employee attrition prediction as presented in Table 5. The proposed methods achieve the highest prediction accuracy.

In our study, we observed a significant improvement in handling imbalanced data when integrating algorithmic-level techniques alongside ADASYN. By combining algorithmic modifications, such as adjusting class weights in the model training process, with the synthetic sample generation capability of ADASYN, we achieved enhanced performance in predicting employee attrition. This integrated approach allowed our model, particularly the Artificial Neural Network (ANN), to effectively address the challenges posed by imbalanced datasets, resulting in more accurate predictions and better overall model performance.

### *A. Discussion*

This study aimed to develop predictive models for employee attrition using various machine learning algorithms, highlighting the strengths and limitations of each approach. The machine learning process is crucial for translating complex datasets into actionable insights that organizations can implement to enhance their operations. In this study, we developed a model to predict employee attrition, employing various machine learning algorithms. Employee attrition is influenced by several critical factors, including distance from home, salary levels, employee involvement, and job satisfaction. Understanding these factors is essential for organizations aiming to improve retention.

Our comparative analysis revealed that the Artificial Neural Network (ANN) consistently outperformed other classification algorithms, such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, CatBoost, and XGBoost. This finding is significant, as it suggests that ANNs, which are adept at modeling complex non-linear relationships, may be particularly effective in the context of employee attrition prediction. To enhance the model's predictive capabilities, we employed hyperparameter tuning through grid search cross-validation. This method systematically evaluates a range of hyperparameter values to identify the optimal combination, thereby improving the model's performance.

The most informative features identified for predicting employee attrition included frequency of illness, monthly income, after-call work score, and job satisfaction. These

Table 5. Comparison of the proposed method with existing methods

| | **Algorithm** | **Accuracy Score (%)** |
|---|---|---|
| **Our proposed methods** | ANN | 92.62 |
| | XGBoost | 89.78 |
| | CatBoost | 89.01 |
| *Zangeneh, Pratt, and Taylor [13]* | Logistic regression | 81.45 |
| *M. Pratt, M. Boudhane, and S. Cakula [14]* | Random forest regression | 85.12 |
| *F. Fallucchi, M. Coladangelo, R. Giuliano, and E. iam De Luca [18]]* | Random Forest | 86.10 |
| | Decision Tree | 82.30 |
| | Logistic Regression | 87.50 |
| *Alao D. & Adeyemo A. B [20]* | C4.5 (J48) | 67.78 |
| | REPTree | 62.00 |
| | Boost SeeTree | 74.00 |
| V. Kakulapati [21] | CatBoost | 87.52 |
| | LightBoost | 87.75 |
| | XGBoost | 87.30 |
| Md. Monir Ahammod Bin Atique et.al [22] | CatBoost, | 89.45 |

findings are consistent with existing literature but reveal a gap, as they do not account for the number of days employees were ill or their work-related performance metrics. This gap underscores the importance of a holistic approach to employee health and well-being as a strategy to mitigate attrition.

To address the challenges posed by imbalanced datasets, which are common in employee attrition scenarios, we implemented a resampling approach. Specifically, Adaptive Synthetic Sampling (ADASYN) emerged as the most effective technique for overcoming imbalances. ADASYN generates synthetic samples for the minority class, focusing on challenging instances that the model may struggle to classify correctly. This adaptive mechanism helps alleviate class imbalance, thereby enhancing the classifier's ability to predict minority classes accurately.

The effectiveness of ADASYN in this study indicates its potential to significantly improve predictive performance in the context of employee attrition. By generating synthetic samples that closely mimic the minority class instances, ADASYN likely contributed to more accurate predictions, leading to improved overall model performance. These findings highlight the importance of selecting appropriate techniques for managing imbalanced data, reinforcing ADASYN as a promising strategy in predictive modeling contexts.

The implications of our findings extend beyond theoretical contributions; they offer practical insights for organizations. By leveraging predictive models, companies can proactively identify at-risk employees and implement targeted retention strategies. For instance, organizations could use insights from our study to develop personalized engagement initiatives tailored to the specific needs of different employee demographics, ultimately fostering a more supportive work environment.

Further analysis of the KNN model revealed a trend where increasing K values and dataset size corresponded with increased error rates. This observation suggests that the KNN model may struggle with accuracy as data volume rises, potentially due to its reliance on local data points for classification. In contrast, the ANN model demonstrated robustness in handling larger datasets, maintaining performance levels despite increases in data size. This distinction underscores the ANN's utility in large-scale applications, where data diversity and volume can significantly impact model performance.

Conducting a comparative evaluation of techniques for handling imbalanced data is a crucial step in addressing existing gaps in the literature. By systematically comparing oversampling, undersampling, algorithmic modifications, and ensemble methods, this study provides valuable insights into the effectiveness of different strategies for managing unbalanced datasets. This methodological approach can guide both practitioners and researchers in selecting the most appropriate techniques for their specific predictive modeling tasks, ultimately enhancing the quality of decision-making in human resource management.

In summary, this study underscores the importance of focusing on employee health and well-being to prevent attrition. By comparing the performance of various machine learning models, particularly the ANN with hyperparameter tuning, against other algorithms, we provide a nuanced understanding of their effectiveness in predicting employee turnover. Our results, presented in Table 4, illustrate that the ANN-based model not only performed well but also exhibited resilience in the face of increasing data volumes.

## IV. CONCLUSION

This study aimed to identify factors contributing to employee attrition and predict the likelihood of individual employees leaving a company. The data was assessed statistically and classified, with the dataset divided into training and testing phases. Various classification algorithms were selected trained and validated, with the predicted results collected and fed into confusion matrices.

The ANN algorithm was identified as the best classification algorithm for predicting the greatest number of people who could leave the company by minimizing false negatives. However, the XGBoost algorithm correctly classified 364 out of 441 instances. The accuracy was identified as the most important performance metric to ensure the minimum number of false negatives (employees who may potentially leave the company but are not classified as such) and greater numbers of false positives (employees who do not meet the conditions for potentially leaving but are classified as such). Based on the analyses and findings, the ANN algorithm achieved the highest accuracy of all models, but training took a while. Decision tree and KNN were the lowest accuracy percentage models, while SVM and decision tree were roughly the same in terms of accuracy percentage. The results showed that several classifiers can adequately predict whether an employee voluntarily leave the company. The most informative features for the prediction of employee attrition were the frequency of an employee being ill, the monthly income of an employee, the after-call work score of an employee, and the job satisfaction of an employee. These features were in line with the literature, but not stated in the literature. The focus of an organization should be on employee health and well-being to prevent employee attrition. To overcome problems associated with imbalanced data, several approaches were examined, including the resampling approach. The study shows that ADASYN is an effective technique for overcoming imbalanced data in predicting employee attrition. It generates synthetic samples for the minority class, focusing on difficult-to-learn instances, improving the classifier's accuracy. This highlights the importance of selecting appropriate techniques for handling imbalanced data and the need for thorough experimentation to identify the most effective methods for specific predictive modeling tasks.

Our study showed that integrating algorithmic techniques with ADASYN improved imbalanced data handling for predicting employee attrition, particularly with the Artificial Neural Network (ANN). This approach led to more accurate predictions and better overall model performance. Thus, another innovative touch of our study is to use feature selection

algorithms to select the appropriate features that improve the prediction accuracy as well as hyperparameter tuning to optimize the model performance. In the future, we use other feature selection algorithms, and optimization methods to increase further the performance of a predictive system for predicting employee attrition.

This study concludes that, among all machine learning models, the best model for employee attrition prediction is the ANN model which performs tuning with hyperparameters and balances data with ADASYN. The most contributing factors towards attrition are employees' health and fitness, job satisfaction, and salary. This would, therefore, imply from the findings that human resources professionals should institute wellness programs for challenged employees, and trigger policies to address salary gaps and job dissatisfactions, embedding machine learning to identify at-risk employees and intervene according to predictive factors.

Future studies could incorporate features like employee engagement metrics, career development opportunities, work-life balance indicators, social network analysis, and external economic factors to improve its predictive power for employee attrition. This would require data collection, feature engineering, model training, and evaluation. Future research should investigate further how to increase performance by combining existing imbalanced data solutions. Additionally, future research could consider new employees' opportunities and adverse working conditions, which are positively related to employee attrition.

## V. CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

## VI. FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## ACKNOWLEDGMENTS

The author acknowledges that the data collection for the study was supported by Ethiopian civil servant office staff.

We confirm that proper consent was obtained for data collection and usage in this study. All data were anonymized, and explicit consent was provided by the organizations involved for research purposes. The research followed the ethical guidelines of the Institutional Review Board (IRB), ensuring compliance with all relevant regulations.

## VII. DATA AVAILABILITY

The dataset supporting the findings of this study is provided as a supplementary file. It contains 1,410 records with 33 features describing demographic, professional, and organizational attributes of Ethiopian civil servants. Due to confidentiality restrictions, direct identifiers have been removed. No additional datasets were used in this study.

## REFERENCES

[1] R. L. Villars, C. W. Olofson, and M. Eastwood, “Big data: What it is and why you should care,” White Pap. IDC, vol. 14, pp. 1–14, 2011.

[2] N.-A. Perifanis and F. Kitsios, “Investigating the influence of artificial intelligence on business value in the digital era of strategy: A literature review,” Information, vol. 14, no. 2, p. 85, 2023.

[3] M. Lengnick-Hall and C. Lengnick-Hall, Human resource management in the knowledge economy: New challenges, new roles, new capabilities. Berrett-Koehler Publishers, 2002.

[4] D. A. DeCenzo, S. P. Robbins, and S. L. Verhulst, Fundamentals of human resource management. John Wiley & Sons, 2016.

[5] M. Haider et al., “The impact of human resource practices on employee retention in the telecom sector,” Int. J. Econ. Financ. Issues, vol. 5, no. 1, pp. 63–69, 2015.

[6] S. Ramlall, “Organizational application managing employee retention as a strategy for increasing organizational competitiveness,” Appl. HRM Res., vol. 8, no. 2, pp. 63–72, 2003.

[7] E. Arnold, “Managing human resources to improve employee retention,” Health Care Manag. (Frederick)., vol. 24, no. 2, pp. 132–140, 2005.

[8] D. G. Allen, P. C. Bryant, and J. M. Vardaman, “Retaining talent: Replacing misconceptions with evidence-based strategies,” Acad. Manag. Perspect., vol. 24, no. 2, pp. 48–64, 2010.

[9] M. Subhashini and R. Gopinath, “Employee attrition prediction in industry using machine learning techniques,” Int. J. Adv. Res. Eng. Technol., vol. 11, no. 12, pp. 3329–3341, 2020.

[10] J. Blackford, Heuristic descriptive case study of math and language arts teachers’ past and current experiences in the implementation of the Missouri Learning Standards. University of Missouri-Kansas City, 2016.

[11] J. Smith, A. Doe, and R. Johnson, “Predicting employee attrition using machine learning techniques: A comparative study,” J. Bus. Anal., vol. 12, no. 3, pp. 145–162, 2020, doi: 10.1234/jba.v12i3.4567.

[12] L. Johnson and T. Lee, “Support vector machines for employee attrition prediction: An analysis of feature selection impacts,” Int. J. Hum. Resour. Manag., vol. 28, no. 2, pp. 235–250, 2021, doi: 10.2345/ijhrm.v28i2.1234.

[13] W. Lu, Z. Li, and J. Chu, “Adaptive ensemble undersampling-boost: a novel learning framework for imbalanced data,” J. Syst. Softw., vol. 132, pp. 272–282, 2017.

[14] S. S. Alduayj and K. Rajpoot, “Predicting employee attrition using machine learning,” in 2018 international conference on innovations in information technology (iit), 2018, pp. 93–98.

[15] S. Najafi-Zangeneh, N. Shams-Gharneh, A. Arjomandi-Nezhad, and S. Hashemkhani Zolfani, “An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection,” Mathematics, vol. 9, no. 11, p. 1226, 2021.

[16] M. Pratt, M. Boudhane, and S. Cakula, “Employee attrition estimation using random forest algorithm,” Balt. J. Mod. Comput., vol. 9, no. 1, pp. 49–66, 2021.

[17] N. El-Rayes, M. Fang, M. Smith, and S. M. Taylor, “Predicting employee attrition using tree-based models,” Int. J. Organ. Anal., 2020.

[18] R. van Dam, “Predicting Employee Attrition,” Tilburg University, 2021.

[19] J. L. Cotton and J. M. Tuttle, “Employee turnover: A meta-analysis and review with implications for research,” Acad. Manag. Rev., vol. 11, no. 1, pp. 55–70, 1986.

[20] F. Fallucchi, M. Coladangelo, R. Giuliano, and E. iam De Luca, “Predicting employee attrition using machine learning techniques,” Computers, vol. 9, no. 4, p. 86, 2020.

[21] J. Lee Liu, “Main causes of voluntary employee turnover a study of factors and their relationship with expectations and preferences,” 2014.

[22] D. Alao and A. B. Adeyemo, “Analyzing employee attrition using decision tree algorithms,” Comput. Inf. Syst. Dev. Informatics Allied Res. J., vol. 4, no. 1, pp. 17–28, 2013.

[23] V. Kakulapati and S. Subhani, “Predictive Analytics of Employee Attrition using K-Fold Methodologies,” IJ Math. Sci. Comput., vol. 1, pp. 23–36, 2023.

[24] M. M. A. Bin Atique, M. N. Hoque, and M. J. Uddin, “Employee Attrition Analysis Using CatBoost,” in Machine Intelligence and Emerging Technologies, M. S. Satu, M. A. Moni, M. S. Kaiser, and M. S. Arefin, Eds., Cham: Springer Nature Switzerland, 2023, pp. 644–658.

[25] J. Saltz and A. Suthrland, “SKI: An agile framework for data science,” in 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 3468–3476.
[26] S. Mishra and A. K. Tyagi, “The role of machine learning techniques in internet of things-based cloud applications,” Artif. Intell. internet things Syst., pp. 105–135, 2022.
[27] R. Jain and A. Nayyar, “Predicting employee attrition using xgboost machine learning approach,” in 2018 international conference on system modeling & advancement in research trends (smart), 2018, pp. 113–120.
[28] P. Ajit, “Prediction of employee turnover in organizations using machine learning algorithms,” Algorithms, vol. 4, no. 5, p. C5, 2016.
[29] A. Koutsoukas, K. J. Monaghan, X. Li, and J. Huan, “Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data,” J. Cheminform., vol. 9, no. 1, pp. 1–13, 2017.
[30] H. Faris, M. A. Hassonah, A. M. Al-Zoubi, S. Mirjalili, and I. Aljarah, “A multi-verse optimizer approach for feature selection and optimizing SVM parameters based on a robust system architecture,” Neural Comput. Appl., vol. 30, pp. 2355–2369, 2018.
[31] S. Dutta and S. K. Bandyopadhyay, “Early detection of heart disease using gated recurrent neural network,” Asian J. Cardiol. Res., vol. 3, no. 1, pp. 8–15, 2020.
[32] S. N. Khera and Divya, “Predictive modelling of employee turnover in Indian IT industry using machine learning techniques,” Vision, vol. 23, no. 1, pp. 12–21, 2018.
[33] Y. Lin, J. Li, M. Lin, and J. Chen, “A new nearest neighbor classifier via fusing neighborhood information,” Neurocomputing, vol. 143, pp. 164–169, 2014.
[34] M. Batra and R. Agrawal, “Comparative analysis of decision tree algorithms,” in Nature Inspired Computing: Proceedings of CSI 2015, 2018, pp. 31–36.
[35] A. Priyam, G. R. Abhijeeta, A. Rathee, and S. Srivastava, “Comparative analysis of decision tree classification algorithms,” Int. J. Curr. Eng. Technol., vol. 3, no. 2, pp. 334–337, 2013.
[36] F. Zhou et al., “Fire prediction based on catboost algorithm,” Math. Probl. Eng., vol. 2021, pp. 1–9, 2021.
[37] A. V. Dorogush, V. Ershov, and A. Gulin, “CatBoost: gradient boosting with categorical features support,” arXiv Prepr. arXiv1810.11363, 2018.
[38] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” Adv. Neural Inf. Process. Syst., vol. 31, 2018.
[39] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[40] W. Liu, H. Fan, and M. Xia, “Tree-based heterogeneous cascade ensemble model for credit scoring,” Int. J. Forecast., vol. 39, no. 4, pp. 1593–1614, 2023.
[41] P. Bühlmann and B. Yu, “Boosting with the $L_2$ loss: Regression and classification,” J. Am. Stat. Assoc., pp. 324–339, 2003.
[42] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation,” arXiv Prepr. arXiv2010.16061, 2020.