Mustansiriyah Journal of Pure and Applied Sciences, Print ISSN: 2957-9910. Online ISSN: 2957-9929, Vol. 4, No. 1, 2026, Pages 174-182
DOI: 10.47831/mjpas.v4i1.M241

MJPAS

MUSTANSIRIYAH JOURNAL OF PURE AND APPLIED SCIENCES

Journal homepage:
https://mjpas.uomustansiriyah.edu.iq/index.php/mjpas

RESEARCH ARTICLE – MATHEMATICS

Apply XGBoost, Random Forest, and Neural Netwrk (MLP) Models for Predicting Employee Absenteeism Using Workforce Data Analytics

[*]Mohammed H. Alars [1] , Abbas M. Albakry [2]

1 *Department of Computer Science, Iraqi Commission for Computers and Informatics (ICCI), Informatics Institute for Postgraduate Studies, Iraq*
mohamed.amnya@gmail.com
2 Department of Artificial Intelligence, University of Information Technology and Communication (UOITC), Iraq
abbasm.albakry@uoitc.edu.iq

Article Info.

Article history:

Received
1 April 2025

Accepted
16 June 2025

Publishing
30 March 2026

Abstract

Absence issues among employees create numerous problems for workplace productivity and operational continuity. The paper analyzes machine learning approaches to predicting employee absenteeism based on controlled workforce information. Scientists applied a thorough preprocessing pipeline to a realistic dataset, which involved dealing with missing values as well as encoding categories along with feature numerical scaling. An exploratory analysis of the data and feature importance evaluation helped identify proper predictive attributes.

Random Forest, together with XGBoost and Multi-Layer Perceptron (MLP) Neural Network, served as the classification models to create predictive models, which were assessed using accuracy and precision and recall, and F1-score metrics. XGBoost proved its superiority as a predictive model by reaching 81.03% accuracy during testing while showing balanced results. This study highlights that ensemble-based machine learning approaches present opportunities to boost data-driven decision-making for human resource management operations. The coming

paper will concentrate on adding more content to datasets as well as improving model visibility while studying adaptive learning methods to enhance real-world systems and stability.

The official journal published by the College of Education at Mustansiriyah University

***Keywords**: Employee absenteeism, XGBoost, workforce analytics, predictive modeling, human resources, machine learning.*

1. Introduction

Organizational resilience together with effective workforce planning requires businesses to predict employee absenteeism levels. Labor resource departments can prevent staff shortages and decrease productivity loss by successfully predicting when employees will be absent from work. This research develops a machine learning-based solution which uses organized employee records to make absentee behavior predictions.

The research initiated with absorption of actual absenteeism data from the real world. The first operations required thorough data cleaning procedures which solved value gaps and fixed inconsistent records and adjusted all categorical entries. The following step involved exploratory data analysis to both examine absenteeism pattern distributions and detect irregularities within the data.

The assessment of important features that influence absentee prediction was achieved through Random Forest and XGBoost feature importance metrics together with correlation analysis which showed reason for absence, transportation cost, service time and hit target score as the key factors. The analysis results led us to chose critical features for modeling and development.

The analysis included three machine learning models: Random Forest and XGBoost and a Multi-layer Perceptron Neural Network (MLP) which were built and evaluated for analysis.

- Random Forest (RF) serves as an ensemble learning method which builds numerous decision trees for collective vote aggregation to predict absentees msed.vse.cz. The method was created by Breiman (2001) pmc.ncbi.nlm.nih.gov from CART decision trees through the combination of bagging (bootstrap aggregating) and split-time random feature selection to achieve decorrelation between treesmsed.vse.cz.. Averaging numerous decorrelated tree predictions provides more robust and accurate predictions compared to a

single tree while reducing the risk of overfitting ibm.com.. The RF approach successfully handles both classification through majority voting among trees and regression by averaging their output predictions while naturally working with high-dimensional structured data.

Robustness and Accuracy represent the main advantages of (RF) through ensemble averaging multiple trees that minimize variance and overfitting issues to enhance generalization capabilities roboticsbiz.comaslopubs.onlinelibrary.wiley.com. The ensemble methods along with RF prove to be superior to single decision trees specifically when used for predicting employee outcomes in HR analytics scenarios file-bogtrrctmrhq3vlqqare7y. RF successfully processes data complexity through its ability to manage both numerical and categorical features without sensitivity to data outliers or noise aslopubs.onlinelibrary.wiley.comroboticsbiz.com. Minimal preprocessing work (such as feature scaling) is needed since RF effectively creates nonlinear class boundaries. RF delivers two types of feature importance metrics called Gini importance and permutation importance to explain which attributes power prediction results roboticsbiz.com. RF provides value to HR domains because it helps identify essential predictive variables (such as factors linked to employee absenteeism and attrition) across the field. RF models with hundreds of trees require more time when training and especially during prediction runtime which could present problems for real-time applications. Due to increased numbers of trees the complete model loses interpretability when compared to single decision trees and linear models. The process for extracting straightforward explanations from predictions remains challenging due to which it hinders transparent decision-making (such as human resources decisions) applications. Even though RF reduces overfitting it does not prevent all cases of overfitting particularly when predictors contain excessive noise or when tree numbers are insufficient when compared to boosting methods. Using numerous trees creates relationship obscuration that results in decreased visibility of individual predictor actions.

- XGBoost Classifier (eXtreme Gradient Boosting) represents a cutting-edge version of gradient boosted decision trees which Chen and Guestrin (2016) msed.vse.cz. developed into their current form. A sequential tree-building process occurs in XGBoost through boosting where each subsequent tree corrects errors present in previous trees by gradient descent on a loss function. The innovations within XGBoost enable it to accomplish accurate classification of structured data while being highly scalable. XGBoost produces an ensemble of shallow trees by combining weighted and summed predictions through which users can control learning rate and tree depth and boosting round parameters for final output generation. The main advantages of XGBoost include its High Predictive Performance which frequently leads to best accuracy results on tabular data kdd.org. XGBoost has proven its superiority as the top algorithm in multiple structured dataset competitions on numerous Kaggle challenges kdd.org. Employee attrition or churn prediction problems benefit from XGBoost since its boosting mechanism helps detect nonlinear feature relationships in order to achieve high predictive accuracy. XGBoost implements built-in regularization functions which apply L1 and L2 penalties to tree leaf weights to control model complexity along with the ability to prevent overfitting. This strength provides XGBoost with superior resistance to overfitting than previous boosting approaches when dealing with noisy data. XGBoost models demonstrate excellent generalization properties when properly parameterized and they perform equally well compared to complex ensemble models msed.vse.cz. The algorithm also delivers high efficiency combined with scalability. The parallel processing capabilities optimized for XGBoost allow it to train quickly on big data sets without compromising memory usage kdd.orgsciencedirect.com. The system works with multiple data structures for swift tree construction and understands missing values through automatic split direction learning sciencedirect.com. These features make it feasible to apply XGBoost to big HR databases or real-time analytics with many features. Flixablity of XGBoost offers multiple classification objective functions including logistic and hinge that enables adaptation to different loss metrics. This tool offers model interpretation features such as feature importance and SHAP value analysis to understand how individual features contribute to the outcome. This capability helps validate that the model keeps within human resource domain understanding. Weaknesses of XGBoost requires thorough adjustment of its numerous parameters including tree depth and learning rate and number of trees and regularization terms because of its advanced feature

capacity. An XGBoost model with either excessive boosting iterations or deep trees without proper regularization will generally exhibit overfitting behavior particularly when dealing with small datasets. A proper stopping point must be chosen during boosting to maintain optimal performance and this selection should occur through either cross-validation or early stopping methods. The human interpretation of XGBoost along with other ensemble tree models remains limited because it functions as a black box. Although the model extracts feature importance the distribution of decision processes across hundreds of trees creates difficulties in obtaining complete decision transparency. The incomplete clarity of model explanations in HR analytics may become problematic when stakeholders need to understand why specific employees get identified as high-risk for leaving the organization. XGBoost requires substantial computational resources because both large ensemble size and extended hyperparameter search increase its computational requirements. XGBoost works at a slower training pace for small datasets because its tree construction follows a sequential process instead of parallel execution however it implements parallel processing for building individual trees. The deployment process becomes heavier due to the presence of numerous trees within the ensemble. Large tabular datasets require considerable computing resources to achieve optimal performance when using XGBoost because it outpaces previous boosting algorithms but demands computing power for its optimal deployment.

- The Multi-Layer Perceptron (MLP) Neural Network stands as a fundamental artificial neural network with three main layers including the input layer and at least one hidden layer together with an output layer. The neurons in an MLP first perform weighted linear input combination and then activate these values through a non-linear function. A Multi Layer Perceptron (MLP) gains complexity in non-linear functionality features when one or more hidden layers exist and this architecture demonstrates universal approximation ability through sufficient neurons en.wikipedia.org. MLPs operate as feed-forward ANNs because data moves from input to output and they commonly train through backpropagation of the gradient descent weights (Rumelhart et al., 1986). The iterative training procedure alters the weights to reach low classification mistakes. MLPs represented some of the first neural network-based models for supervised learning while remaining popular models for classification applications according to mdpi.com. The main capabilities of MLP include Model Complex Relationships and Handle Complex Feature-Target Relationships scaler.com. The network obtains the capability to detect advanced feature combinations as each added hidden layer enhances its processing ability particularly in structured datasets where variable interactions matter. An MLP network will use attendance data and workload information to generate accurate attrition risk predictions when these inputs are weak predictors independently. The multilayer perceptron works as a domain-neutral function approximation model. The application of Multilayer Perceptron (MLP) extends across various fields which include financial analytics as well as Human Resources analytics and image recognition and speech recognition mdpi.commdpi.com. The MLP provides a suitable model for tabular structured data when suitable encoding transforms the input data into numerical values. MLP operates naturally for multi-class classification while simultaneously generating multiple labels as outputs. The framework enables users to tackle various predictive assignments (employee promotion eligibility to loan default prediction and medical diagnosis) through a unified system. An MLP derives new data representations from input features through its hidden layers rather than depending on existing feature splits like decision trees. The system generates new predictive features internally from its input components. The implementation of Multi-Layer Perceptrons on tabular data shows they achieve better results than tree ensembles yet still discover important linear attribute combinations that are hard to split with tree methods. Training an MLP with sufficient data enables the network to generate improved accuracy or gain new knowledge (such as an internal combination of HR metrics). An MLP delivers continuous output probabilities through its mathematical computation process before the final threshold operation for classification purposes. The output predictions from MLP models become either better calibrated or demonstrate an ability to produce probabilities as prediction scores. The output of tree-based

models generates constant values across defined sections of the input space. An MLP generates output that serves as a confidence indicator for decision-making when experts need probabilistic predictions like employee retention likelihood. The main limitations of (XGBoost) include its requirement for abundant training data while it also suffers from overfitting behavior. An MLP with complex configuration tends to match training data noise and artificial patterns when dealing with scarce structured input data typical in HR analytics. The networks easily overfit when the model is both too extensive and trained excessively without implementing regularization methods. However the implementation of methods including dropout and weight decay typically becomes necessary to address this problem which creates additional experimental hurdles scaler.com. Hyperparameter Sensitivity There are many design choices (number of hidden layers, number of neurons per layer, activation functions, learning rate, etc.) that significantly affect an MLP's performance. Finding the right architecture and training parameters can be challenging and typically requires experience or extensive tuning scaler.com. In contrast, tree-based models often work reasonably well with default settings. For a researcher in HR analytics without deep neural network expertise, this sensitivity can be a hurdle. Computational Cost Training an MLP can be computationally intensive, especially if the network has many layers or neuronsscaler.com. Each training epoch involves many floating-point operations. While modern libraries are optimized (and can use GPUs), a well-tuned MLP may still take longer to train than an XGBoost or RF on the same data. If the data is high-dimensional or one-hot encoded (common in structured data with many categories), the input layer becomes large, further slowing computation. Lack of Interpretability: MLPs are often criticized as "black boxes." The learned weights and nonlinear transformations do not have an obvious interpretation for humans, unlike a decision tree's splits or a regression's coefficients. This opacity can be problematic in domains like HR or healthcare where understanding *why* the model made a prediction is important for trust and actionable insights scaler.com. Although techniques like SHAP values or saliency maps can be applied to MLPs, the explanations are not as straightforward as, say, listing top features in a random forest. Need for Preprocessing Neural networks expect numerical input; thus categorical variables must be encoded (e.g. one-hot), and features usually should be normalized. If an HR dataset has categorical fields (department, job role, etc.), an MLP workflow must transform these appropriately. Improper preprocessing can hurt an MLP's performance significantly. Tree-based models are more robust to raw categorical inputs (when encoded as integers) and to unscaled inputs, so an MLP places a heavier burden on data preparation.

Three machine learning algorithms were chosen because they possess unique strengths which make them effective for structured data classification problems. Random Forest achieved selection because its ensemble structure utilizes multiple decision trees to decrease overfitting and enhance generalization. This approach shows outstanding performance when working with high-dimensional tabular data while delivering interpretable feature importance that uncovers the main factors affecting absenteeism. XGBoost proved most suitable because it delivered the best predictive results with optimal computational performance. The gradient boosting framework of this method lets it construct sequential trees to fix previous iteration errors thus enabling detection of complex feature relationships. Regularization features embedded within the model lower the chance of overfitting which makes it suitable for working with our actual HR dataset. This research included MLP Neural Network to evaluate the performance of a non-tree-based non-linear predictive model. MLP networks use multiple layered transforms to identify and learn complex relationships between input variables. The utilization of these models in this research allows researchers to determine if deep non-linear feature extraction methods surpass traditional tree-based methods.

Our goal was to find an optimal solution between model interpretability and performance as well as computation efficiency through the implementation of these three models. A comparative assessment helped determine XGBoost as the most appropriate method for dependable absenteeism prediction.

The analysis included stratified data splits for training each model followed by standard classification metric performance evaluation. XGBoost achieved both high predictive accuracy of 81.03% and strong model robustness.

2. Related Work

Several studies have explored predictive analytics and machine learning in HR applications. Zupančič and Panov (2024) implemented tree-based methods using historical timesheet profiles and demographic data to forecast sick and vacation leave. Their study demonstrated the effectiveness of predictive clustering trees and ensemble models for different forecasting intervals, establishing a practical foundation for integrating such models in HRIS environments.

Mohan et al. (2024) proposed machine learning approaches to optimize time and attendance systems through predictive analytics and anomaly detection. Their work highlighted the operational improvements in accuracy and efficiency when biometric systems and automated data processing were integrated. These studies provide strong support for the predictive capabilities of ML in personnel management.

Cho et al. (2023) reviewed HR analytics frameworks with a specific emphasis on public sector applications. They outlined a five-step implementation strategy define, collect, analyze, share, reflect and emphasized the role of machine learning in enabling evidence-based decision-making. Their work underscores the importance of integrating analytics within government HRM contexts and highlights potential ethical and technical challenges.

These recent contributions, alongside studies by Kocakulah et al. (2016), Gandomi and Haider (2015), and De Masi et al. (2021), reinforce the suitability of machine learning especially ensemble methods like XGBoost for enhancing workforce intelligence, proactive staffing, and operational efficiency.

3. Dataset and Preprocessing:

The dataset (Absenteeism_at_work) was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil. Employee information was included in Dataset which contained details about Reason for absence, Month and day of absence, Seasonal information, Commute-related expenses and Service tenure and demographic data. The target label consisted of two categories where employee absenteeism exceeding three hours received a value of 1 but all other instances retained a value of 0. Analyzing data using various analytics methods to benefit from the outputs in identifying important features in the database and identifying the relationships and correlations between them.

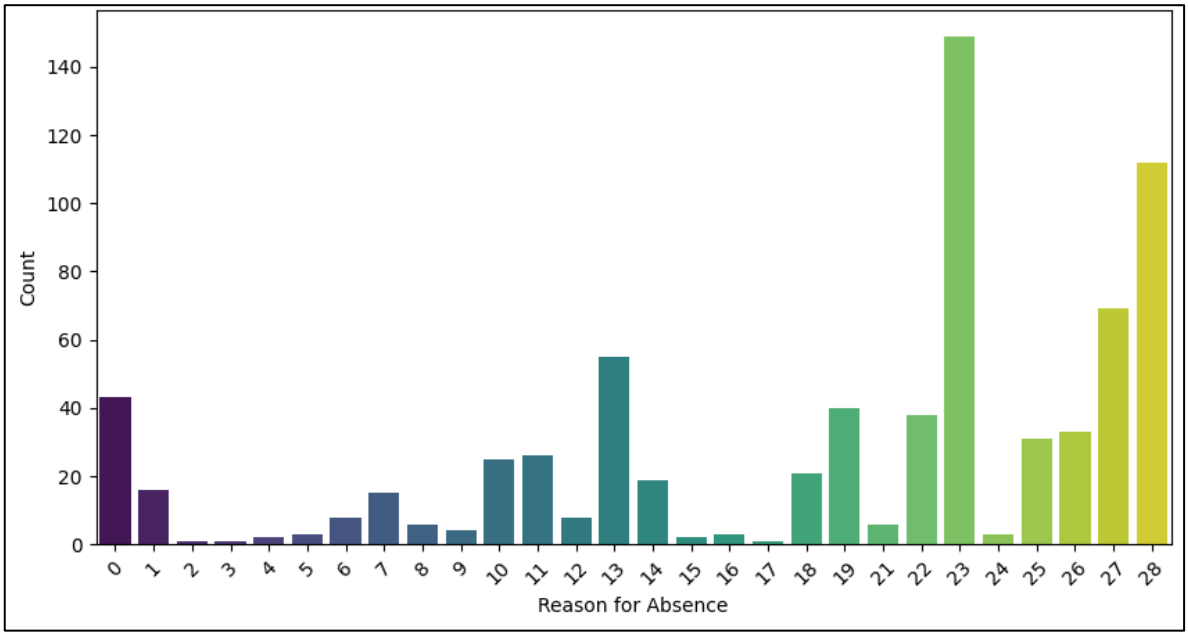

Fig. 1. (Absenteeism by Reason for Absence).

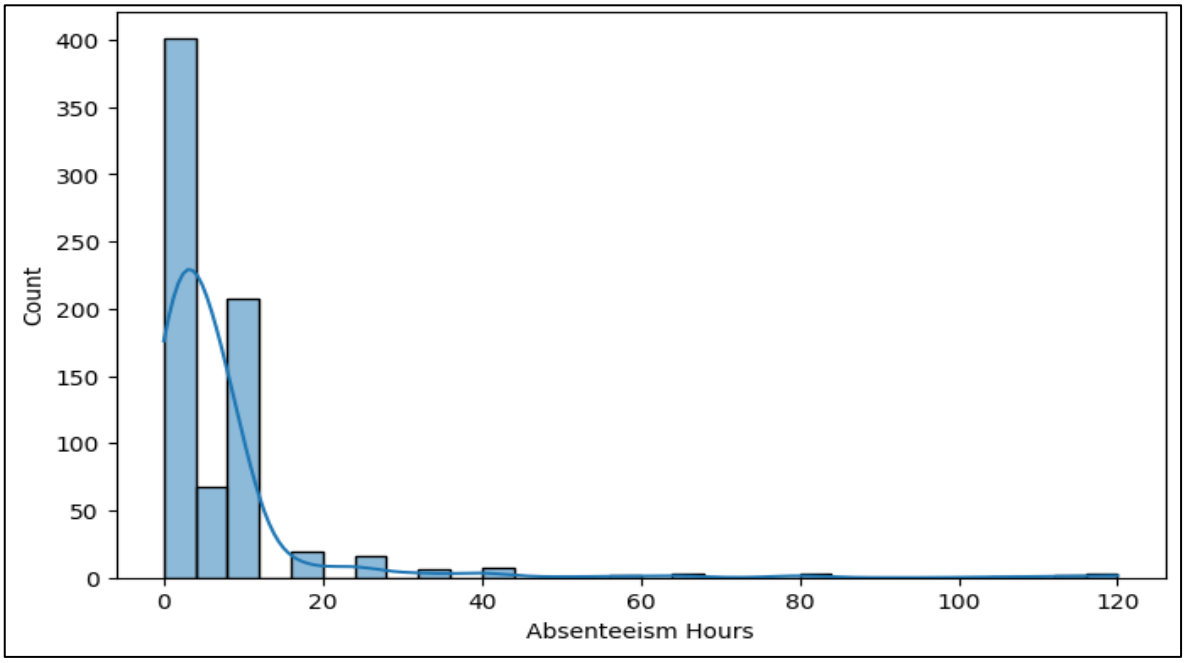

Fig. 2. (Distribution of Absenteeism time in Hours).

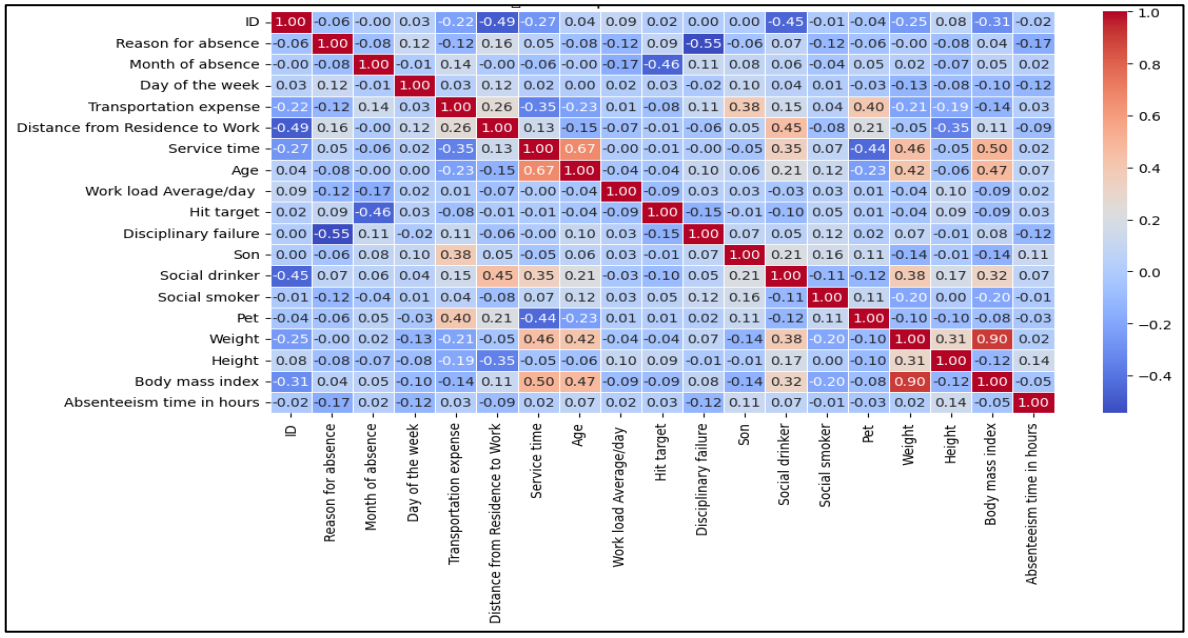

Fig. 3. (Heatmap of feature Correlation).

4. Methodology

The paper methodology includes a systematic process for machine learning technique implementation which creates predictive models to identify employee absenteeism patterns. The system follows these sequential phases for its workflow:

- The process starts with obtaining a complete **employee information dataset**. The dataset provides historical worker data on personal attributes and workplace attendance records with other contributing variables for absenteeism behavior.
- **The preprocessing** work includes cleaning and preparing data to make it suitable for machine learning algorithms by executing several essential steps.
  - **Handling Missing Values**: technique of imputation completes the processing of missing data points.
  - **Encoding Categorical Variables**: The conversion of categorical data attributes into numerical numbers happens through encoding techniques that include one-hot encoding and label encoding.

- **Numerical Features require Feature scaling** (standardization or normalization) to normalize their ranges so models converge better.

- Data exploration takes place during Employee **Data Analytics** to **discover valuable insights and pattern** recognition within the data. The analysis focuses on essential variables which demonstrate strong correlation to absenteeism because these data points will be essential for the model development process.
- The process includes **Feature Selection** that uses statistical or model-based methods to remove unimportant features then maintains attributes which boost prediction accuracy.
- The processed dataset serves as input for three machine learning models including **Random Forest, XGBoost and Neural Network** because these methods demonstrate high classification performance when handling complex data structures.
- A model evaluation process determines the performance of tested models through **Accuracy, Precision, Recall and F1-Score** metrics. The evaluation metrics create an all-encompassing view of the prediction ability for employee absenteeism across each model.
- **The evaluation** results from all models are compared to identify their specific strengths alongside weaknesses.
- **Selection of the Best Model** The model demonstrating the highest performance across the evaluation metrics is selected as the most suitable for absenteeism prediction.

The methodology employed in this study follows a systematic and structured approach to develop a predictive model for employee absenteeism using machine learning techniques. The workflow consists of the following sequential phases:

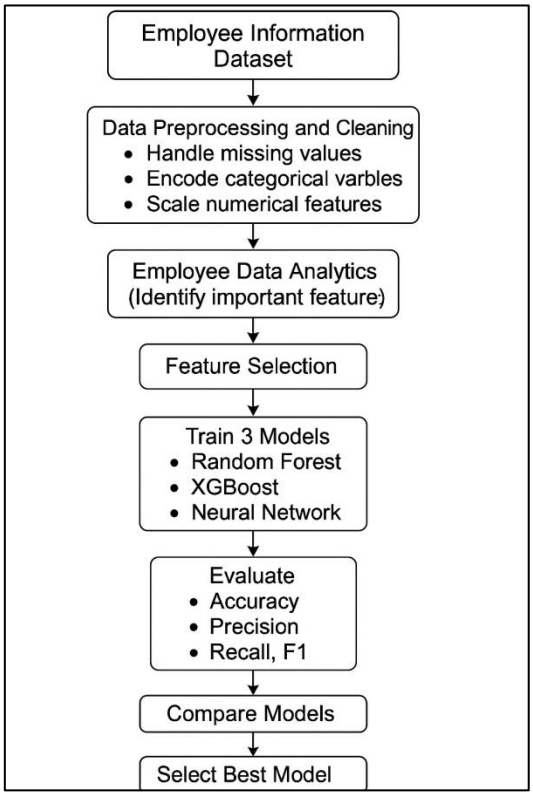

Fig.4. (General outline of the study's progress)

5. Results and Comparison

Table 1 Performance Metrics Across Models

Metric	Random Forest	XGBoost	Neural Network
Accuracy	80.17%	**81.03%**	68.10%
Precision (Present)	**86%**	83%	72%
Precision (Absent)	74%	**78%**	63%
Recall (Present)	77%	**83%**	73%
Recall (Absent)	**84%**	78%	62%
F1-Score (Weighted)	80%	**81%**	68%

Analysis and Comparison

The results reveal notable differences in model performance:

- XGBoost model demonstrated superior performance over most other models throughout the entire evaluation process. The model demonstrated the highest success rate of 81.03%, delivering balanced results across critical metrics. It maintained strong precision and recall for both absentee and present classes, which makes it reliable for real-world HR applications where both false positives and false negatives carry organizational cost.
- Random Forest model ranked second with an accuracy rate of 80.17%. The accuracy rate for detecting present employees (86%) was the highest, but XGBoost demonstrated slightly better detection of absent employees, thus missing fewer actual absences.
- MLP neural network model performed poorly compared to its peers. Its accuracy rate (68.10%) was significantly lower, while its recall for detecting absent employees (62%) was the weakest point, demonstrating the difficulty of generalization when working with the features and size of the dataset used in this research.

The ensemble-based XGBoost classifier performed best for pattern recognition with employee data, making it the model of choice for deployment. XGBoost achieved superior metrics across most measurement points, thus becoming the final model of choice.

6. Conclusion:

This Paper performed an extensive assessment of different machine learning techniques for employee absenteeism prediction that revealed XGBoost model surpassed other methods both in predictive power and stability. Advanced data analytics demonstrate their potential for using evidence-based decisions in human resource management through these research results. Future research should focus on expanding the dataset while improving interpretability of models and studying adaptive learning methods to adapt to changes in employee behavior patterns.

References:

[1] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System.
[2] Breiman, L. (2001). Random Forests. Machine Learning.
[3] Gandomi, A., & Haider, M. (2015). Big Data Concepts and Analytics. Int. J. Info. Mgmt.
[4] Kocakulah, M. C., et al. (2016). Absenteeism problems and costs. Int. J. Econ., Comm. & Mgmt.
[5] De Masi, F., et al. (2021). Predicting absenteeism with ensemble learning. J. Appl. Comp. & Info.
[6] Zupančič, P., & Panov, P. (2024). Predicting Employee Absence from Historical Absence Profiles. Appl. Sci., 14(16), 7037.
[7] Mohan, P., et al. (2024). Optimizing Time and Attendance Tracking Using ML. IJRMEET, 12(7).
[8] Cho, W., Choi, S., & Choi, H. (2023). HR Analytics for Public Personnel Mgmt. Admin. Sci., 13(2), 41.
[9] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning representations by back-propagating errors*. Nature, 323(6088), 533–536. DOI: 10.1038/323533a0.

[10] Hornik, K., Stinchcombe, M., & White, H. (1989). *Multilayer feedforward networks are universal approximators*. Neural Networks, 2(5), 359–366. DOI: 10.1016/0893-6080(89)90020-8.

Appendix:

- Python code snippets (Google Colab compatible)
- API deployment walkthrough
- Feature importance charts