University of Tehran Press

Interdisciplinary Journal of Management Studies (IJMS) Online ISSN: 2981-0795

Home Page: https://ijms.ut.ac.ir

Predicting University Entrance Examination Ranks by Developing a Stacking-Based Ensemble Machine Learning Algorithm

Mohammad Reza Mehregan[1*] | Arman Rezasoltani[2] | Amir Mohammad Khani[3]

1. Corresponding Author, Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. Email: mehregan@ut.ac.ir
2. Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. Email: armanrezasoltani@ut.ac.ir
3. Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. E-mail: amir.mo.khani@ut.ac.ir

ARTICLE INFO

Article type:
Research Article

Article History:
Received 24 November 2024
Revised 06 July 2025
Accepted 20 October 2025
Published Online 31 May 2026

Keywords:
educational planning,
ensemble learning,
national university entrance exam,
Optuna,
stacking.

ABSTRACT

A key issue for planning and consulting is the accurate prediction of students' rankings in important national university entrance exams, such as Iran's nationwide university entrance examination, commonly known as the Konkur. Although machine learning has been increasingly used in educational data mining, most existing models have shown limited accuracy, are inadequately formulated, and lack sufficient optimization for practical application. This study introduces a novel stacking-based ensemble learning model that incorporates XGBoost, LightGBM, and CatBoost as base learners, with a linear regression model as a meta-learner to improve national rank prediction. The proposed model's main hyperparameters were adjusted using the Optuna optimization framework to enhance the performance of each model. The model was trained and validated on a large dataset of over 73,000 student records from Ghalamchi Institute and evaluated using five-fold cross-validation with NRMSE and R^2 as performance measures. The results showed that the proposed model significantly outperformed baseline models, such as Random Forest, Gradient Boosting, and MLP Regressor, achieving NRMSE of 0.0659 and R^2 of 0.7735, which could be attributed to the effective integration of advanced learners with systematic hyperparameter optimization. This research provides a practical and scalable predictive tool that can support academic advisors, educators, and policymakers in making informed decisions, promoting equity in education, and guiding students through data-driven interventions. The use of stacking-based ensemble learning and automated hyperparameter optimization via Optuna distinguishes this study from prior research and is a meaningful step forward in the application of predictive analytics in high-risk educational settings.

Cite this article: Mehregan, M. R.; Rezasoltani, A. & Mohammad Khani, A. (2026). Predicting University Entrance Examination Ranks by Developing a Stacking-Based Ensemble Machine Learning Algorithm. *Interdisciplinary Journal of Management Studies* (*IJMS*), 19 (3), 545-566. http//doi.org/10.22059/ijms.2025.385883.677186

 Publisher: University of Tehran Press.
DOI: http//doi.org/10.22059/ijms.2025.385883.677186

1. Introduction

The prediction of students' national rank in university entrance exams remains a crucial issue for educational researchers and policymakers (Ali et al., 2019; Alyahyan & Düştegör, 2020; Chaparro-Cruz et al., 2025; Jafarnejad Chaghoshi et al., 2024). Reliable models can not only influence learning outcomes and promote educational fairness but also significantly influence students' future opportunities, given their impact on the future areas of the students' academic opportunities and even on their career trajectories (Almalawi et al., 2024). Ensemble learning approaches have proven to be powerful techniques for accurately predicting academic performance in this context (Tang et al., 2024). In Iran, the national university entrance exam –known as *Konkur*– plays a pivotal role in the education system. As a highly competitive and centralized annual examination, it determines students' eligibility for admission to universities and higher education institutions. Its significance cannot be overstated, as it serves as the primary criterion for selecting both fields of study and academic institutions. As a result, having the capacity to forecast the outcome of the university entrance exam is extremely important not only for the students and their families but also for the educational consultants and decision-makers in the academic arena. In the case of the Iranian university system, this exam (*Konkur*) is probably one of the most important decisions of the students' career to get into university. The highly charged atmosphere surrounding the *Konkur* exam exerts immense psychological pressure on students and their families, often leading to heightened stress and anxiety. On the other hand, the dark side of the popularization of higher education is that those who already have the convenience to study at a university have now started the tendency to make the university entrance exam (*Konkur*) the only defining factor in university admission. There is evidence suggesting that students may possess valuable skills and abilities that are not adequately reflected in written test scores. As a result, relying solely on such assessments may prevent the country from thoroughly exploiting its human resource potential (Parviz, 2023). Considering such challenges, it is quite natural to build an accurate framework to compute the performance prediction and estimation model so that students' performance can be evaluated in the *Konkur* exam. In this field, ensemble learning models have proven to be the most effective approach recently introduced. Aggregating various machine learning algorithms will further increase the accuracy of the predictions, providing more universities with more insights into how prediction systems work (Khan et al., 2024). Therefore, the outputs of different models are ensembled to minimize errors and ensure more accuracy of the overall predictions (Yan & Liu, 2020).

Ensemble models are recommended by various studies for predicting students' performance based on major exams, such as *Konkur*, to predict scores with high accuracy. According to Adejo and Connolly (2017), heterogeneous ensemble learning models enhance the prediction's manifold accuracy. However, in the case of applying an ensemble of algorithms, such as SVM, Random Forest, and Adaboost, other studies' single-algorithm-based models were less precise or reliable than the group of ensemble models proposed in this study. These findings are recently confirmed by studies that prove higher accuracy of ensembles than other techniques for the students' academic performance prediction. As recent evidence highlights, with more datasets and complexity, the ensemble model will be more applicable (Butt et al., 2023). Such models can be used as tools for policymakers and educational planners to improve prediction tasks using an ensemble of models, such as bagging or boosting methods. In the same vein, Han et al. (2017) revealed that the Adaboost rank algorithm performed much better than decision trees, SVM, and neural networks (Zohrehvandian et al., 2023). Using ensemble learning methods is shown to increase prediction accuracy while reducing systematic errors in student performance measurement. This work applies an ensembled learning model using stacking and optimizing hyperparameters of the model under the OPTUNA framework to improve the prediction of the national ranking of the university entrance exam (*Konkur*). OPTUNA has been demonstrated in previous works to be applicable to a very large space of models in machine learning, including neural networks and tree-based models (Jafarnejad et al., 2025). In the application context of the optimization stock price and heart disease prediction models, the OPTUNA framework produces highly accurate and precise results, improving the prediction accuracy significant (Bishwakarma & Sharma, 2022; Yang et al., 2023). Different base models can be combined to reach a method with the most refined results. Therefore, research evidence shows that stacking different tasks as diverse as electrical load and critical temperature prediction yields improved accuracy compared to other ensemble methods, including bagging and boosting (Massaoudi et al., 2020;

Yu et al., 2023). Unlike the optimization of the base learners' parameters, optimization of the ensemble model parameters is another big challenge; however, OPTUNA presents advanced strategies such as Parzen estimation and pruning techniques. Through parameter selection (Akiba et al., 2019), the best parameters can be selected at lower computational costs, and an improvement in performance is observed for stacking models.

The goals of the current research are to improve predictions for the Iranian students ranking in the university entrance exam (*Konkur*) based on advanced models, such as XGBoost, LightGBM, and CatBoost, within the same stacking framework. These models are further optimized with advanced optimization techniques. The ranking prediction accuracy can be improved by a great margin if the OPTUNA framework for hyperparameter optimization is utilized together with Stacking models. In other words, this is an optimized combination that performs significantly better in terms of accuracy and the computation time involved. The considerably intense reliance on the *Konkur* exam in Iran, along with the psychological and educational pressures surrounding it, necessitate predictive tools that are both efficient and generalizable. There have been numerous studies on using ensemble learning for the prediction of student performance, but only a few have addressed the specific problem of national ranking prediction using pre-exam data. However, as a review literature highlights, no research has used auto hyperparameter optimization in this area, necessitating investigations with a novel stacking-based ensemble model, optimized through the Optuna framework, to predict the students' national ranks based on their preparatory test performance with higher accuracy. This contribution adds to the field of academic prediction with the development of a more precise and practical model than the existing tools, with a major emphasis on the use of a setting such as Iran's national university entrance exam in a high-stakes standardized testing environment. This study seeks to develop an appropriate model with high ranking accuracy and efficiency using Ghalamchi (a famous institute for university entrance exam preparation and simulation) exam data. In this case, OPTUNA is used for hyperparameter tuning of different models for this purpose and introducing the final model performance optimal. As a result, the novelty of this research lies in the provision of an innovative way of university entrance exam ranking prediction with the help of several coupled machine learning models and an advanced hyperparameter optimization method. In addition, the paper carries out some comparisons and analyses, proving that the proposed model is faster and can achieve higher prediction accuracy than the baselines.

The key questions of this research include the following:

- How can the accuracy of student ranking prediction in the university entrance examination be increased through an ensemble learning model using the Stacking method?
- What combination of machine learning models can perform best in predicting university entrance exam ranking?
- How can optimization of hypermeters through OPTUNA improve the final performance of the model?

Based on the students' real data, this work seeks to find solutions for the enhancement of education quality and equal educational opportunities using some advanced ensemble learning methods. This study may form the basis for developing prediction and decision-making systems in the country's educational sector to ensure more appropriate and efficient decisions by policymakers and educational advisors.

2. Literature Review

The proposed study developed an ensemble learning model based on stacking, incorporating the OPTUNA hyperparameter optimization framework, to predict students' national rankings in the university entrance examination (*Konkur*). A review of relevant literature highlights the comparative effectiveness of various ensemble learning models in this domain. These studies consistently demonstrate that ensemble approaches offer superior accuracy and efficiency in predicting exam outcomes. Moreover, such models hold significant potential for integration into educational systems, where they can be used to monitor student performance and support data-driven decision-making in academic planning and guidance. Table 1 presents a summary of the research background.

Table 1. Research Background on Student Performance Prediction

Authors	Article Title	Findings	Model Used
Taher Mazandarani et al. (2025)	Predicting Student Academic Performance: A Machine Learning Approach and Feature Analysis	Positive personality traits, such as interest in studying, quality of homework, contentment, self-regulation, and logical thinking and reasoning, are the most significant predictors of students' academic performance. The CatBoost model achieved high prediction accuracy with $R^2 = 0.87$.	CatBoost Regressor ($R^2 = 0.87$)
Salari et al. (2024)	Predicting Student Performance Using Machine Learning Algorithms and Educational Data Mining (A Case Study of Shahid University)	The decision tree algorithm outperformed other algorithms in predicting students' performance, achieving an accuracy of 84.71% and predicting graduation for 77.88% of honors students.	Decision Tree
Jafarnejad Chaghoshi et al. (2024)	Unleashing the Power of Ensemble Learning: Predicting National Ranks in Iran's University Entrance Examination	XGBoost was the most accurate model and obtained the best results with the least error. LightGBM also had high accuracy and good performance. Random Forest provided decent accuracy but was slower. CatBoost was the weakest-performing model, exhibiting higher error rates compared to the other models evaluated.	XGBoost LightGBM Random Forest CatBoost
Sukhija & Faridi (2024)	Recommending Graduate Admission Using Ensemble Model	Stacked model (DT, KNN, NB) outperformed individual models; applicable for real-time admissions	DT, KNN, Naive Bayes (base), Logistic Regression (meta)
Abiodun & Wreford (2024)	Student's Performance Evaluation Using Ensemble Machine Learning Algorithms	Stacked ensemble achieved RMSE = 0.1768, $R^2 = 0.97$	Random Forest, KNN, XGBoost (base), stacked
Ballaho (2024)	Predicting Student's Success in Programming Courses	Stacking ensemble of SVM, Decision Tree, and Neural Network showed the highest reliability in predicting success in programming-related admissions.	SVM, DT, NN with Stacking
Butt et al. (2023)	Performance Prediction of students in Higher Education using Multi-Model Ensemble Approach	This study implemented different ensemble learning models, such as bagging, boosting, stacking, and voting, for the prediction of student performance, revealing that boosting and stacking outperformed the rest for both small and large datasets.	BaggingBoosting Stacking, Voting
Zub et al. (2023)	Two-Stage PNN–SVM Ensemble for Higher Education Admission Prediction	The two-stage model showed 94% accuracy; better than boosting/bagging	PNN (base), SVM (ensemble), LR (meta)
Saluja et al. (2023)	Designing New Student Performance Prediction Model Using Ensemble ML	Predicts final grade and stream assignment with a 93% accuracy	DT, KNN, NB, SVM (base), One-vs-Rest Ensemble
X. Chen et al. (2022)	A Competition Model for the Prediction of Admission Scores of Colleges and Universities in Chinese College Entrance Examination	The competitive model, along with clustering, could increase the accuracy of predicting acceptance scores by 7.3%.	A competitive model with clustering
Zangooei & Fatemi (2021)	Prediction of Students at Risk of Academic Failure Using Learning Analysis in the Learning Management System	The LSTM network outperformed the SVM in predicting students at risk of academic failure, improving teacher and student performance with a 94% accuracy and 88% efficiency.	LSTM, SVM
Sakri & Saleh (2020)	A Robust Hybrid Ensemble Model for Students' Performance Assessment on Cloud Computing Course	This study introduces a hybrid RHEM model that achieves a 91.7% accuracy by combining naïve Bayes, multilayer perceptron, k-nearest neighbors, decision table, bagging, and random subspace algorithms to predict student performance.	Naïve Bayes, Multilayer Perceptron, k-Nearest Neighbours, Decision Table, Bagging, Random Subspace
Yan & Liu (2020)	An Ensemble Prediction Model for Potential Student Recommendation Using Machine Learning	This study uses ensemble learning models, such as Adaboost, Random Forest, and SVM, to predict student performance, revealing improved accuracy compared to single-algorithm models.	AdaboostRandom ForestSVM
Injadat et al. (2020)	Systematic Ensemble Model Selection Approach for Educational Data Mining	Ensemble learning models, which combine multiple algorithms, provide more accurate and efficient predictions than single-algorithm models, leading to more reliable results.	Random Forest, Gradient Boosting, Adaboost, Bagging
Adejo & Connolly (2017)	Predicting Student Academic Performance Through a Multi-Model Heterogeneous Ensemble Approach (Output Prize)	Combining diverse data sources with collective learning models increases the accuracy and efficiency of prediction, surpassing the effectiveness of a single model and limited data.	Random Forest, Gradient Boosting, Stacking
Han et al. (2017)	Application of Ensemble Algorithm in Students' Performance Prediction	Adaboost algorithm outperformed other collective learning algorithms, such as decision trees, neural networks, and SVM, in terms of prediction accuracy.	Adaboost, Decision Tree, Neural Network, SVM
Wang & Shi (2016)	Prediction of the admission lines of college entrance examination based on machine learning	The Adaboost algorithm, a combination of simpler models, performs better than traditional prediction methods in predicting university entrance exam scores, achieving higher accuracy and reliability.	Adaboost

Previous studies on the use of machine learning algorithms and hybrid models in predicting student academic performance have been widely covered in the literature. However, earlier studies were restricted to predicting the grade in the course, admittance to diploma, or overall student performance evaluation in most cases, failing to focus on the particular problem of predicting a national rank in national examinations. For instance, Salari et al. (2024) used a decision tree as a tool to predict academic performance but did not investigate its application at a national or competitive exam level. Other studies, such as Sakri and Saleh (2020) and Yan and Liu (2020), also have only looked at performance in a certain courses or offered individual suggestions for major choices. Moreover, some researchers, including Injadat et al. (2020) and Butt et al. (2023), used hybrid models such as Bagging, Boosting, and Stacking; however, these studies have mostly compared models using academic data rather than seeking to predict the rankings in competitive exams. Like Abiodun and Wreford (2024) and Ballaho (2024), some studies have applied Stacking models to the admission or success in certain courses. A few studies, such as Jafarnejad Chaghoshi et al. (2024), have focused on predicting college entrance examination outcomes, but the models were not optimized systematically, imposing the hyperparameters and failing to use external tools such as Optuna for the performance improvement. Additionally, model selection in most previous studies tends to be more based on trial and error or default settings, rarely encompassing advanced optimization techniques such as early stopping and adaptive search. In the meantime, with a special focus on predicting the national rank of national entrance exam candidates based on Ghalamchi preparation test data, the present study can cover three main gaps in the existing literature: first, a precise and practical focus on predicting the national rank rather than just the score or acceptance; second, using a combination of three powerful algorithms (XGBoost, LightGBM, and CatBoost) in the form of a Stacking model to employ the advantages of each; and third, implementing automatic hyperparameter optimization under the Optuna framework, which has increased the accuracy and efficiency of the final model. The research innovation lies in combining a robust ensemble of simple yet effective models, using linear regression as the meta-learner, alongside an intelligent optimization method—an approach that has either been entirely overlooked or applied manually and inconsistently in previous studies. Building on this, the current study leverages advanced blended learning and intelligent optimization to develop a novel system for predicting entrance exam scores and supporting decision-making for educational consultants and policymakers. This system offers a reliable, practical, and generalizable solution within the Iranian education context.

3. Research Methodology

This research aims to provide an ensemble learning model for predicting the national rank of students in the university entrance examination (*Konkur*) following the Stacking approach and the OPTUNA hyperparameter optimization framework, using the data of the preparatory tests before the entrance examination. The data used in this research were extracted from the website of the Cultural Center of Education, also known as Ghalamchi (https://www.kanoon.ir), using the web scraping method. One of the most powerful tools in the field of web scraping, octoparse software, was used throughout the data extraction process, enabling the collection of extensive and detailed data for the analysis of ensemble learning models.

3-1. Data Collection:

The data used in this research were extracted from the official website of the Cultural Center of Education (Ghalamchi). This website provides comprehensive and accurate information about the performance of students in the Galamchi (preparatory tests before the entrance examination) and *Konkur* (Iran's university entrance examination), presenting the average score of the Ghalamchi exams, national and regional ranks in the university entrance exam (*Konkur*), fields of study in high school, university of admission, as well as other related details. The extracted data pertain to the period of 2011 to 2021, used for more detailed analysis and prediction of national ranking in Iran's university entrance exam. All scraping was performed under the websites' terms of service, and no login, authentication, or access to restricted content was required. Crucially, no personally identifiable information was collected, and all the data were anonymized and used only for academic research under strict ethical standards for data collection and privacy. Meanwhile, while attempts were made to

collect a suitable dataset with web scraping, there may be some limitations in the representativeness of the data. Particularly, the partial geographical coverage and the possible underrepresentation of some educational regions during the modeling process could produce the bias effect. These limitations should be taken into account when interpreting the results, and future research may employ official or broader data sets to further generalize the results.

3-2. Web Scraping Process

Octoparse software was used to collect data from the website of the Cultural Center of Education. Octoparse is a powerful web scraping tool that allows users to automatically extract data from websites without the need for programming (Kahlon & Singh, 2024). The main stages of the web scraping process in this research were as follows:

- Determining the address of the web pages: The address of the web pages containing information about the students was identified and given to the Octoparse software.
- Data extraction: Various parts of web pages, including name of city, national rank, average test score, focal history, number of tests, field of study, university of admission, and other relevant information, were selected and extracted using Octoparse drag and drop tools.
- Storage: The extracted data were automatically saved in the form of an Excel file and then used for further analyses in the Python environment.

3-3. Dataset Features

3-3-1. Number of Records and Features:

The final dataset contained 73,838 records of students' data who participated in the Ghalamchi tests and were accepted in the national entrance exam. These data pertained to the period from 2011 to 2021. The dataset includes 11 distinct features that were employed for further analysis. Table 2 summarizes the characteristics of the dataset used to predict students' national ranks in the entrance exam.

Table 2. Dataset Characteristics for Predicting National Rank in the University Entrance Exam

Feature	Time to determine the result	Data type	Description
City	Before the national exam	Categorical	The city where the student lived or studied
National rank in university entrance exam (*Konkur*)	After the national exam	Numerical	The final rank of the student at the national level in the entrance exam
Regional Rank in university entrance exam (*Konkur*)	After the national exam	Numerical	The students' ranking in the region where they took the exam (regions 1, 2, or 3)
Region	Before the national exam	Categorical	Determining the students' study area based on the place of residence (regions 1, 2, or 3)
Number of years participating in the preparation tests	Before the national exam	Numerical	The number of years or the perior of students participation in Ghalamchi tests
Average preparation test score	Before the national exam	Numerical	The average score of the students in the Ghalamchi tests
Number of preparation tests	Before the national exam	Numerical	The number of tests that the students have participated in the cultural center of education (Ghalamchi)
Acceptance field	After the national exam	Categorical	The field of study in which the students have been accepted in the national entrance exam
University of admission	After the national exam	Categorical	The university where the students have been accepted to continue their studies
High school major	Before the national exam	Categorical	The field that the students studied in high school (mathematics, empirical science, humanities, etc.)
Year of participating in Iran's nationwide university entrance exam	Before the national exam	Numerical	The year students participated in the entrance exam and were accepted

This research selected six main characteristics, all related to the preparation tests before the national entrance exam, to predict the national rank of students in the national entrance exam (which is determined after the entrance exam). These characteristics specifically include the average preparation test score, the number of years participating in the preparation tests, the number of preparation tests, high school major, region, and the year of participating in Iran's nationwide university entrance examination. These features were selected due to their direct influence on students' performance in the entrance exam and their potential to improve the accuracy of national rank prediction. Among them, the study region stands out as a key factor, determined based on the city in which students reside or attend school. It classifies candidates into three categories, including Regions 1, 2, and 3, according to the availability of educational resources and the overall quality of educational conditions. Considering the region feature rather than the city is justified for the following reasons:

- **Effect of educational facilities**: Zoning policies based on access to educational and social resources have a direct impact on students' academic performance. Empirical studies, examining geographic disparities in educational infrastructure, have revealed significant variations in entrance exam outcomes across different regions (Daniele, 2020).
- **Generalizability and better categorization**: Segmenting cities into three broader categories based on training contexts allows for more effective generalization and categorization of the data. In contrast, treating each city as a separate unit of analysis may introduce too much specificity for the model and could detract from the generalizability of the results and their interpretation (Gibson & Webb, 2015).

The variables related to ultimate outcomes on entrance exams and field selection, including the field of admission or the university of admission, were excluded from this study. The rationale for excluding these variables is as follows:

- **Preventing the impact of output data:** Incorporating intrinsic features that are directly related to the final outcomes of the entrance exam, such as the chosen field of study or university placement, can inadvertently introduce bias into the model. These outcome-related variables are determined after the exam and are often influenced by the student's performance or the chosen field. These features are considered output data after the test rather than features effective in predicting; therefore, they are not going to have any impact on improving the accuracy of predictions (Navarro et al., 2021).
- **Focusing on academic performance before the entrance exam:** This research aims to predict the students' ranking based on academic performance and preparation tests before the entrance exam. Therefore, the characteristics related to post-concert results, such as field or university of admission, are known as posteriori information and cannot be used when predicting the national rank. These features may increase the complexity of the model and decrease the prediction accuracy, as they are unavailable at the moment of prediction and created after the test (Yağcı, 2022).

Table 3 summarizes the descriptive statistics for the numerical features used in this study. Only the features selected for the model development were included in the descriptive analysis, while features not utilized in the modeling process were excluded. This approach ensures that the descriptive statistics directly reflect the variables relevant to the predictive modeling tasks. The table includes the mean, standard deviation, minimum, and maximum values for each feature. These statistics provide an overview of the data distribution and variability across different attributes, providing deeper insights into the characteristics of the dataset and ensuring appropriate preprocessing steps before model training.

Table 3. Descriptive Statistics of the Numerical Features Used in the Model Development

Feature	Mean	Std	Min	Max
Number of years participating in preparation tests	2.0195	1.2853	1.0	13.0
Average preparation test score	6235.5941	1162.6476	3888.0	8693.0
Number of preparation tests	35.2456	22.9175	1.0	239.0
Year of participating in entrance exam	1394.8828	2.8837	1390.0	1400.0
National rank in the university enterance exam (*Konkur*)	10093.5632	5168.7058	1.0	18255.0

3-4. Data Preprocessing:

This study used the Ghalamchi test data. Variables under study included the average preparation test score, the number of years of participating in the preparation tests, the number of preparation tests, high school major, region, and year of participation in Iran's nationwide university entrance exam. During data preparation, the raw data were first inspected and cleansed. The subsequent steps in the preprocessing workflow included the procedures below (Furkat et al., 2024):

- **Removal of missing values and anomalies:** Data with missing values or anomalies that adversely affected the modeling were removed.
- **Encoding of category features**: For categorical features such as high school major and region, label encoding was applied to convert the data into numerical format, making it suitable for use in machine learning models.
- **Standardization of numerical features**: To scale the features, StandardScaler was used to standardize numerical features such as mean focal balance, focal history, number of tests, and year, ensuring that all the features had a mean of zero and a variance of one, which could help improve the efficiency and accuracy of the models.

3-5. Data splitting

After preprocessing, the data were split into training and testing. As suggested by Han et al. (2011), in a good model, 80% of the data could be used for training and the rest as test data. To make the evaluation performance of the models even more realistic, K-fold cross-validation with five repetitions was performed, giving a more realistic generalizability check on the models (Collins et al., 2024). K-fold cross-validation, in other words, is a resampling method that performs the division of data into K parts. In one run, everything but K-1 segments will be considered training data and the rest as test data (Teodorescu & Brașoveanu, 2025). To ensure that every data segment serves once as test data, it repeats K times, enabling us to test the model on all the data by the end and take the average of these repetitions as a mean approximate measure for the model's final performance (T r et al., 2023). The random distribution in different partitions will reduce the variance, making model evaluation more accurate and fair. It should be noted that rather than a repeated cross-validation approach, a standard 5-fold cross-validation was employed. Each fold division was performed randomly without stratification, ensuring that every sample had an equal chance of being included in each partition. After cross-validation, the final models were retrained on the entire dataset using the best-found hyperparameters to fully utilize the available data for final evaluation.

3-6. Correlation Analysis

To examine the interrelationships among the selected features and assess the potential risk of multicollinearity, a Kendall rank correlation matrix was computed. Kendall's Tau was chosen due to its reduced sensitivity to outliers and its greater reliability in cases where the data deviate from normality, making it a more appropriate measure than Pearson's correlation in this context (Akoglu, 2018). As shown in Figure 1, no dyadic pair of features exhibits a correlation coefficient greater than 0.8, indicating a lack of multicollinearity and supporting the independence of the selected features for model training. Furthermore, only features available prior to the university entrance examination were incorporated into the model. This ensures the predictive validity of the model, as it relies solely on exogenous, independent variables that precede the outcome. Figure 1 presents the correlation coefficients in the form of the heatmap, with the color density and color shade reflecting the intensity of the correlation and its direction, respectively: dark red indicates a strong positive correlation, dark blue represents a strong negative relationship, and lighter tones denote weak or even unimportant relationships. Such ease of visual interpretation helps to quickly find a couple of possible problematic features that should be avoided; checking that chosen features are as minimally redundant as possible and help provide unique information to make a convenient prediction.

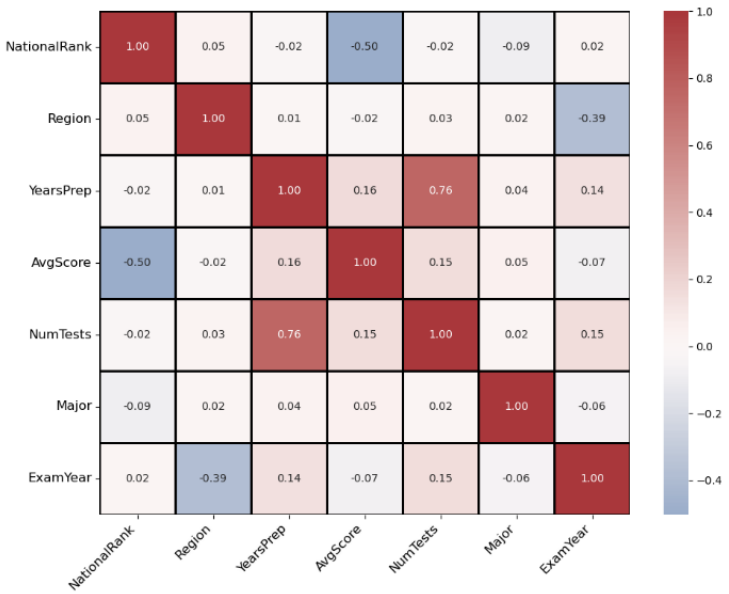

Fig. 1. Correlation Matrix

4. Stacking

The current study used various machine learning models in combination with a stacking ensemble model to predict the rank of students. In essence, ensemble learning comprises a variety of learning wherein the idea lies in integrating several machine learning models to enhance the overall performance and raise the accuracy of prediction (Balaji et al., 2021). This approach helps to improve the performance by reducing the variance, bias, and errors of the base models. One of the important ensemble learning methods is stacking, in which different models are trained in parallel and their outputs are combined by a final model, known as a meta-model (Dey & Mathur, 2023). Stacking is intended to serve the final model, covering the weaknesses and errors the base models might have produced, finally providing stronger and more accurate predictions. Several base models are independently trained in this approach, whose combination predicts the output, which is fed as an input to the meta-model. Meta-modeling will hence learn the pattern in the output of base models and combine them in such a way that minimizes the final error (Aboneh et al., 2022).

The meta-model plays a very important role in stacking, as it decides how to best combine the output of the base models (Zhang et al., 2024). This research has used linear regression as a meta-model because linear regression can easily learn the relationship between the outputs of the base models and the target variable, functioning as an effective combiner. Linear regression was chosen due to its simplicity and high speed, making the meta-model training process more quickly and accurately. This research has used three basic models, XGBoost, LightGBM, and CatBoost, as underlying models, while a linear regression was responsible for combining the outputs of the basic models. These models were chosen because each had certain advantages as follows: XGBoost showed significant speed and high performance in big data, LightGBM exhibited low memory consumption and high training speed, and CatBoost could handle classified data without the need for complex pre-processing. Linear regression was used as a meta-model to simply and effectively combine the outputs of the basic models and make the final prediction. Figure 2 illustrates the structure of the proposed ensemble learning model, highlighting the integration of multiple base learners through a meta-model for enhanced predictive performance.

The performance of each of these methods is described in detail below:

XGBoost (Extreme Gradient Boosting) is an approach to decision tree-based machine learning algorithms that implements the gradient boosting method to create prediction models (Jafarnejad et al., 2025). In the algorithm, the models are constructed sequentially, and each new model attempts to reduce the previous model's errors (Bentéjac et al., 2020).

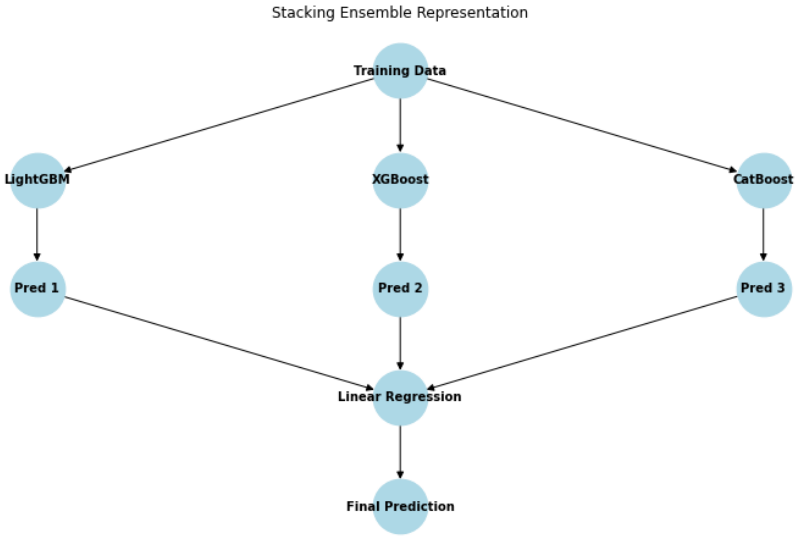

Fig. 2. The Proposed Ensemble Learning Model

The main goal of XGBoost is to minimize the cost function, which usually includes a loss function, L, and a penalty term to avoid overfitting. The total cost function is as follows:

$$L(\theta) = \sum_{i=1}^{n} L\left(Y_i, \hat{Y}_i\right) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

In which:

y_i: the actual value of the i-th sample.

$\hat{y}_i$: the predicted value for the i-th sample.

$L(y_i, \hat{y}_i)$: loss function, for example, mean squared error (MSE) for regression or cross-entropy for classification.

$\Omega(f_k)$: penalty term for model complexity defined as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j} \omega_j^2 \quad (2)$$

In which:

T: the number of leaves of the decision tree.

w_j: weights of leaf nodes.

γ and λ: hyperparameters that control the complexity of the model.

- **Updating the model at each step:**

In each step of the XGBoost algorithm, a second-order approximation to the loss function is used instead of directly minimizing the cost function. This optimization is conducted using gradient and Hessian:

$$parent L^{(t)} \approx \sum_{j=1}^{n} \left[g_j f_t(x_j) + \frac{1}{2} h_j f_t^2(x_i) \right] + \Omega(f_t) \quad (3)$$

where, g_j is the gradient loss function, and h_j represents the Hessian loss function (T. Chen & Guestrin, 2016).

LightGBM is a gradient-boosting algorithm that uses advanced optimization methods in design to make the system fast and efficient, especially for large and complex datasets (Rizkallah, 2025). The novelty of this algorithm is leaf-wise growth with histogram-based splitting when building decision trees (Huang & Wang, 2023). Similar to XGBoost, LightGBM tries to minimize an objective function that consists of a loss function and a penalty term on model complexity. The major difference it has from other algorithms is its use of histogram segmentation or the process of reducing computational time by quantizing continuous features into discrete bins and then performing calculations on these bins. In this way, the method drastically reduces the computational complexity and hence is much faster.

- **Updating the model at each step:**

Like in XGBoost, LightGBM also uses the second-order approximation of the loss function. These are possible optimizations used by LightGBM in building decision trees, enabling it to give high accuracy with improved performance (Ke et al., 2017).

CatBoost is one of the most significant algorithms based on gradient boosting, developed expressly for batch data. Unlike most machine learning counterparts, which require the input data to be pre-transformed into batches before they can be processed, CatBoost operates directly on batch data, offering a major benefit. Like other gradient boosting models, including XGBoost and LightGBM, the underlying method of CatBoost is also the optimization of a cost function that contains two parts: the loss function and the penalty term. However, more exciting than all other features is the use of the Ordered Boosting method in CatBoost. While classical gradient boosting methods train all the models on all the data, it can lead to overfitting. But in CatBoost, data is used in a certain order such that new models use only those data that previous models never saw. It reduces overfitting and improves the final model.

- **Ordered Boosting in CatBoost:**

One of the most distinguishable differences we look for when using CatBoost versus other gradient-boosting algorithms is the use of Ordered Boosting over Standard Boosting. In Ordered Boosting, the data are sorted in a specific order, and each new model is trained using the value based on the model (and tree) before it. This technique prevents over fitting and reduces data leaking (Hancock & Khoshgoftaar, 2020).

- **Updating the model at each step:**

Similar to any other gradient boosting algorithm, at each iteration of the CatBoost updating process, the model is updated using gradient information and Hessian information. The key difference about CatBoost is to introduce batch data and apply ordered boosting (Prokhorenkova et al., 2017).

Linear regression: Linear regression is one of the simplest and most effective machine learning methods for predicting continuous variables. This model assumes a linear relationship between the independent variables (inputs) and the dependent variable (outputs), seeking to find the best fitting line or surface that describes the data. The general formula of linear regression is as follows (Khani et al., 2022):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \quad (4)$$

Here, the response variable or outcome is comprised of a set of predictive or return variables, a set of unknown parameters, and a random error called a linear regression. Linear regression seeks to fit a hypothesized function to the data by finding the values of coefficients that minimize the sum of the squared errors (ϵ values) so as to maximize the model'sprediction ability. This research has used linear regression as a meta-model in Stacking to combine the predictions of the base models and make the final prediction. The meta-model in the stacking ensemble is designed to integrate and optimize the predictive outputs of the base learners. Instead of using the original input features, the meta-model takes the predicted values generated by each base model as an input for the given instances. It then learns how to assign appropriate weights to these predictions, capturing complementary strengths and mitigating individual weaknesses of the base models. The meta-model trains on the outputs rather than the raw features, effectively blending the diverse perspectives of the base learners and ensuring improved generalization and overall predictive performance on unseen data.

- **Hypermeters optimization**

The OPTUNA hypermeter optimization framework was used for hypermeter optimization of the models used in stacking (Akiba et al., 2019). Optuna is an open-source hypermeter optimization framework that automatically searches for the correct hypermeter combinations based on advanced methods, including adaptive search and Bayesian optimization (Srinivas & Katarya, 2022). The system has many features, including pruning, which reduces the time taken to train models, automatically stopping the training if the model is identified to be under-performing. Optuna explores the hypermeter search space in the best way possible, as it uses a sequential optimization framework to return the best possible combination of hypermeters (Rimal et al., 2024). Traditional methods, such as Grid Search and Random Search, have been widely used for hyperparameter tuning (Petro & Pavlo,

2019). Grid Search exhaustively explores all possible combinations within a predefined range, which can be computationally expensive and impractical for large parameter spaces. Random Search samples parameter combinations randomly and is often faster but may miss optimal regions. In contrast, Optuna is a modern optimization framework based on sequential model-based optimization (SMBO), which intelligently explores the search space. This framework uses techniques such as early stopping (pruning) and adaptive sampling, leading to faster convergence and better results with fewer evaluations. Optuna was chosen for this study due to its efficiency, scalability, and ability to find better hyperparameters with fewer trials. The optimization process utilizing Optuna is implemented by initially defining the search space for various hypermeters. Then, based on intelligent search algorithms like TPE (Tree-structured Parzen Estimator), the framework studies different points of the search space, searching for points that have a better chance to improve the dv/bs. For the purposes of this research, hypermeters for CatBoost, like number of iterations and number of trees (n_estimators) for LightGBM and XGBoost, were optimized using Optuna (Cai et al., 2024).

One of the salient features of Optuna is its pruning system, in which the poorly performing model in the early stages of training is detected through the system and stopped, saving computation time and increasing the efficiency of optimization. Additionally, Optuna allows the user to interface nicely for optimization purposes and visualize search results. Lastly, we could find an optimal combination of hypermeters and increase the performance of the Stacking model using Optuna. Below, are some optimized hypermeters discovered through this method: number of CatBoost iterations, number of trees, and n_estimators in LightGBM and XGBoost. Table 4 presents the optimal value for each hyperparameter.

Table 4. Optimal Hyperparameters Using OPTUNA

Model	Hypermeter	Optimized value
CatBoost	Number of iterations	106
LightGBM	Number of trees (n_estimators)	178
XGBoost	Number of trees (n_estimators)	164

Once the optimal hyperparameters were identified using the Optuna framework, each base model (XGBoost, LightGBM, CatBoost) was retrained on the entire dataset using these best-found parameter values. This approach maximizes the use of available data and enhances the generalization capability of the models before integrating them into the final stacking ensemble.

To evaluate the effectiveness of the stacking model, several machine learning models, including linear regression, random forest, gradient boosting, and MLP regressors, were also used to predict students' national rank. For each model, the hyperparameters were optimized using the grid search method to obtain the best possible performance, and then their performance was compared with the stacking model. Below is a description of each of the basic models:

Random Forest: This decision tree-based model increases prediction accuracy by randomly combining multiple trees. In Random Forest, multiple decision trees are built from random samples of the training data (Bootstrap Sample). The final prediction is made based on the predictions from these trees. The final prediction is calculated as follows (Lee et al., 2019; Salmanpoursohi et al., 2024):

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{5}$$

$\hat{f}(x)$: the final prediction.

$f_b(x)$: the prediction of the b^{th} tree for sample x.

B: the number of trees in Random Forest.

For regression, the average of the predictions of all trees is calculated:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \tag{6}$$

For classification, majority voting is conducted on tree predictions:

$$y = \text{mode}\,(f_1(x), f_2(x), \ldots, f_B(x)) \tag{7}$$

- **Formula for error calculation in Random Forest:**

One of the important aspects of Random Forest is the ability to measure out-of-bag errors (OOB errors). This method uses data not used in building a specific tree and to evaluate the performance of the model without the need for a separate test data set. The formula for the OOB error calculation is as follows:

$$\text{OOB Error} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{y}_{ooB,i}) \quad (8)$$

In which:

N: number of samples.

y_i: real value.

$\hat{y}_{OOB,i}$: prediction using trees that have not seen this sample.

L: error function, such as MSE for regression or Cross-Entropy for classification (Breiman, 2001).

Gradient Boosting: This model utilizes a collection of decision trees, the difference being that it improves the mistakes produced from the earlier models. Every time it increases one model, it innovates one model that minimizes the mistakes of the prior model. This model was included in the simple model, as it maintained excellent accuracy and the ability to model complex relations (Biau & Cadre, 2021).

MLP Regressor (Multilayer Neural Network): This artificial neural network includes hidden layers and can learn complex nonlinear relationships between features and the target variable. This model was used as one of the basic models due to its ability to learn complex patterns (Lazcano et al., 2024).

Linear Regression: This is the easiest and fastest kind of machine learning model, in which prediction is for continuous variable(s). A linear regression model relies on the assumption of a linear relationship between the independent and dependent variable(s). Despite limitations, linear regression has been chosen as one of the basic models due to simplicity in analysis and interpretation (Sarker, 2021).

- **Evaluation of models:**

NRMSE (normalized root mean squared error) and R^2 were employed to assess how well the models performed. These attributes were calculated for both the experimental and cross-validation data. NRMSE provides a view of the model's error normalized between a scale of zero to one, and R^2 represents the amount of variance explained (Botchkarev, 2019).

NRMSE (normalized root mean squared error): NRMSE is among the performance evaluation criteria of prediction models, transforming root mean squared error (RMSE) into a standardized score based on the variance of the RMSE across aggregate data sets. NRMSE gives a better representation of the error rate in a way that can be contrasted between multiple datasets. Using the range of the dependent variable, NRMSE is obtained by dividing RMSE into the difference between the maximum and minimum values of the dependent variable. NRMSE is typically between 0-1 (Botchkarev, 2019).

NRMSE helps us illustrate how well the model is scaled (normalized), making it possible to evaluate the performance of the model compared to other models, regardless of the scale of the variables. The lower the NRMSE value, the better the performance of the model. This criterion is calculated as follows:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{y_{max} - y_{min}} \quad (9)$$

In which:

n: number of samples

y_i: true value of sample i

$\hat{y}_i$: predicted value for sample i

And y_{max} and y_{min} are the maximum and minimum of the target variable, respectively.

R-squared (R^2): R^2 is a common measure to evaluate the explanatory power of the model, indicating what percentage of the variance of the dependent values is explained by the model. The formula for R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (\bar{y} - \hat{y}_i)^2} \quad (10)$$

Here, $\bar{y}$ is the mean of the actual values and the denominator is the variance fraction of the actual values. The R^2 value is between zero and 1, where 1 indicates a perfect model that explains all the variance, and zero indicates a model that explains no variance.

- **Implementation and Tools**

This research was implemented using Python language and sci-kit-learn, Optuna, LightGBM, XGBoost, and CatBoost libraries. Besides, Pandas was used for data management and Numpy for numerical calculations. All models and data preprocessing steps were implemented in a reproducible manner to ensure reliable research results for future use and further development.

- **Analysis and results**

The comparison results of different models for students' national rank prediction in the national university entrance exam showed that the stacking ensemble learning model outperformed basic models of linear regression, random forest, gradient boosting, and MLP regressor. Moreover, NRMSE and R^2 measures were calculated for all models to evaluate their performance precisely. As previously mentioned, these two metrics provide two different perspectives of evaluating model performance. NRMSE provides an average assessment of how large the prediction errors are in relation to the scale of the target variable. It is an intuitive measure of how closely the predictions match the actual values of the target variable. A lower value of NRMSE occurs when the model's prediction is made with less deviation, indicating more accuracy in absolute terms. R^2 measures the proportion of the variability in the dependent variable that can be explained by the predictions made by the model. A larger R^2 indicates that the model successfully captured the underlying patterns or relationships in the data and has explanatory and generalization power concerning the variability of the target variable. Both NRMSE and R^2 must be evaluated together to fully understand the performance of a model. While NRMSE assesses the magnitude of prediction error, R^2 accounts for how well the model explanation accounts for the variability from the target variable. Simultaneously evaluating the two metrics allows for either the variability explained by the model or the predictive accuracy of the model to be weighed evenly. All models were run on a system with an Intel Core i5-7200U processor, 8GB of RAM and Python 3.12.

Stacking model: The stacking model could achieve the best performance by using a combination of three basic models, XGBoost, LightGBM, and CatBoost, and linear regression as a meta-model. The NRMSE of the model in cross-validation was 0.0659 and R^2 yielded a value of 0.7735, highlighting the high accuracy and ability of this model to accurately predict country ranks.

We also compared the performance of basic models with the stacking model. The results indicated that the random forest model and gradient boosting outperformed the stacking model, although with a higher error rate. Random forest, with respective NRMSE and R^2 values of 0.0940 and 0.7390, could provide excellent accuracy. The gradient boosting model also had the same accuracy as the random forest with an NRMSE of 0.0970 and an R^2 of 0.7224 but performed slightly worse. The results of MLP regression, which had an NRMSE of 0.1033 and an R^2 of 0.6849, and linear regression, with an NRMSE of 0.1414 and an R^2 of 0.4098, were weaker than those of other models, indicating their inability to effectively model the complexities of the data. Table 5 presents a summary of the results obtained from comparing the models:

Table 5. Comparison Results of Models

Model	NRMSE (cross-validation)	R^2 (cross-validation)
Linear Regression	0/1414	0/4098
Random Forest	0/0940	0/7390
Gradient Boosting	0/0970	0/7224
MLP Regressor	0/1033	0/6849
Stacking	0/0659	0/7735

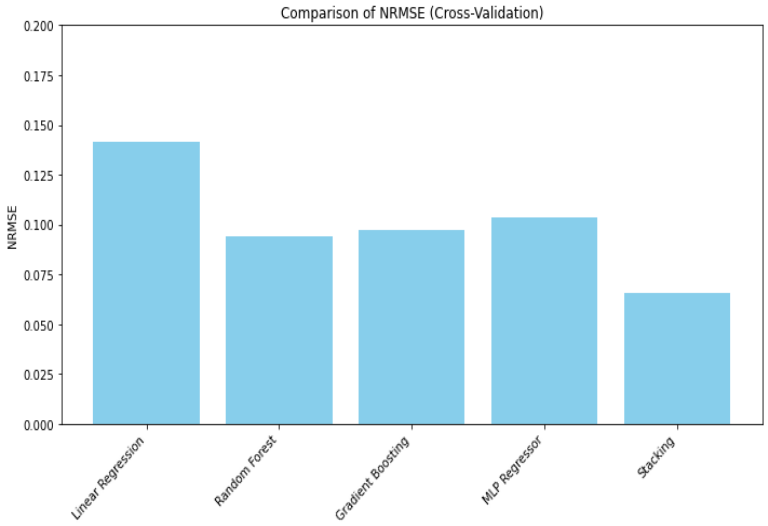

Fig. 3. Comparison of NRMSE Results (Cross-Validation)

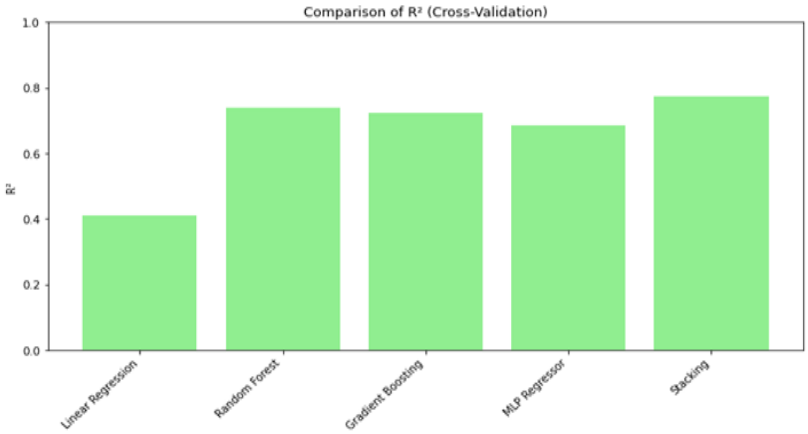

Fig. 4. Comparison of R^2 Results (Cross-Validation)

Figures 3 and 4 present the comparisons of model performances based on NRMSE and R^2 results, respectively, obtained through cross-validation. These results show that by combining advanced models, the Stacking model has achieved the best performance in predicting students' national rank. The use of NRMSE and R^2 criteria and cross-validation provide a more accurate performance assessment of the models and allow a more fair comparison between them. Lower NRMSE and higher R^2 values of this model indicate its greater ability to accurately predict national ranks than other models. Figure 5 illustrates the comparative analysis of model performances based on both NRMSE and R^2 results.

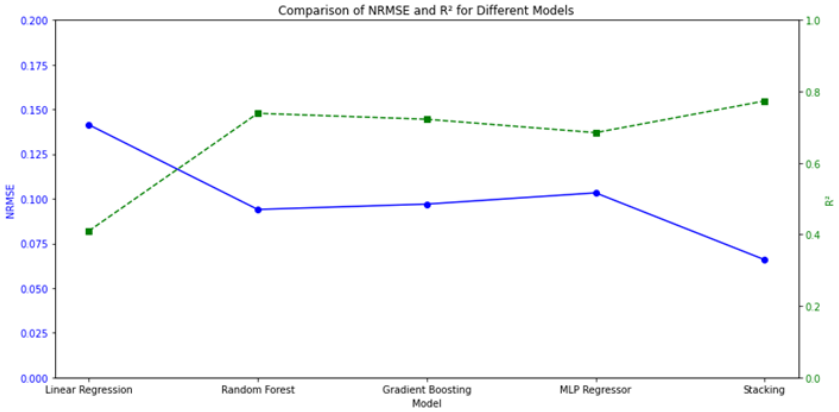

Fig. 5. Comparison of NRMSE and R^2 Results

An informative index of the NRMSE provides some hints on the model accuracy and a comparison among different models. The NRMSE is one of the main advantages because it normalizes the error relative to the data range and enables comparisons between models of various scales. A model's accuracy in predicting actual values is greater if the NRMSE is smaller, which indicates that the amount of error is smaller. The NRMSE value from stacking was the best for predicting students' national ranking at 0.0659, while that of linear regression was the weakest on the same task at 0.1414. By enabling the model's comparability with simpler ones, Random Forest and Gradient Boosting models obtained NRMSE values of 0.0940 and 0.0970, respectively, higher than simple models but less accurate than the stacking model. R^2 is another measure referred to as the coefficient of determination, which denotes the percentage of variance accounted for by a model, implying how accurate the model's predictions are to actual data. The R^2 is bounded between 0 and 1 inclusive, with closer R^2 values to 1 indicating a better explanation of the dependent variable variance, and closer R^2 values to 0 revealing a worse explanation of the variance within the data. Taking the current study as an example, Stacking had the best performance of 0.7735 and outperformed Linear Regression (the best R^2 value of 0.4098), which could not explain the variance of data X.

Overall, NRMSE and R^2 are respective metrics that contribute crucially to evaluating the performance of models; therefore, an overlapping use of these metrics can yield a more professional viewpoint of how well the models can describe the data. R^2 stands for the model's capacity to explain the variance of a target variable, and NRMSE measures only the amount of prediction error. In general, predicting the data and adequately explaining its variation is usually achieved by models with low NRMSE and high R^2. The following paragraphs compare the performance of different models by illustrating the prediction graphs of the linear regression (LR), random forest (RF), gradient boosting (GB), and MLP regressor (MLPR) models with real values of the students' national ranking (as inputs) in Figure 6. For each of these graphs, the horizontal axis is the actual values and the vertical axis is the predicted ones. The prediction is perfect and error-free, as shown by the red line. An analysis of these charts is as follows:

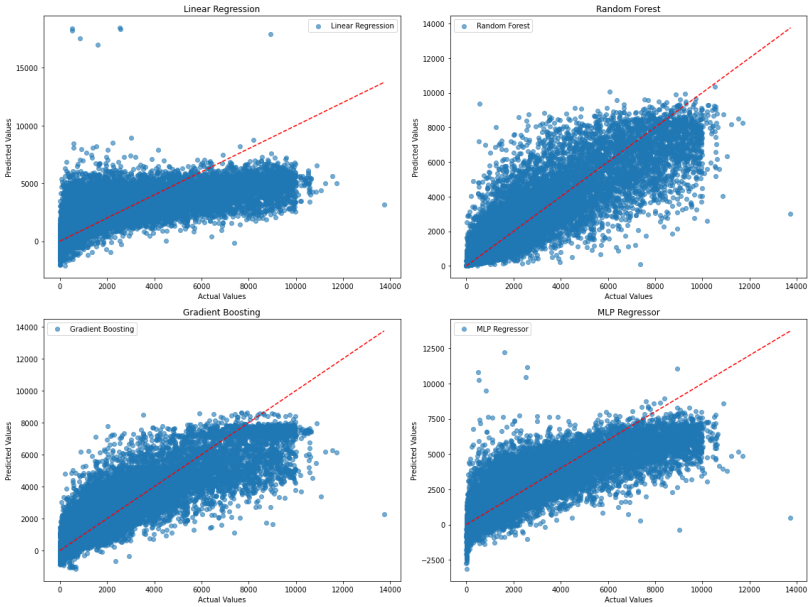

Fig. 6. Comparison of the Predicted and Actual Values for Four Base Models

Linear regression: The predictions of the linear regression model have a high dispersion compared to the full prediction line, highlighting its poor performance in modeling complex relationships and inability to accurately predict national ranks.

Random forest: The random forest model performs better than linear regression and is less scattered than the full prediction line, but there are still scattered points around the line, indicating prediction error in some cases.

Gradient Boosting: The gradient boosting model has the same accuracy as the random forest, and the predictions are closer to the full prediction line, but there is still some scatter.

MLP Regressor: The predictions of the MLP Regressor model also exhibit a significant scatter, indicating that it is not as capable of modeling the complexities of the data as more advanced models.

Figure 7 presents the performance of the stacking model against the real values of the students' national ranking. Therefore, such a model could combine the three basic models of XGBoost, LightGBM, and CatBoost using a meta-model based on linear regression to make better predictions. This shows a very low dispersion of the points in the full prediction line of this model compared to other models, reflecting higher accuracy in predicting country ranks. In the meantime, the stacking model was the best among all models, with an NRMSE value equal to 0.0659, while this graph shows how its predictions are closer to the actual value, revealing the capability to learn and combine various features.

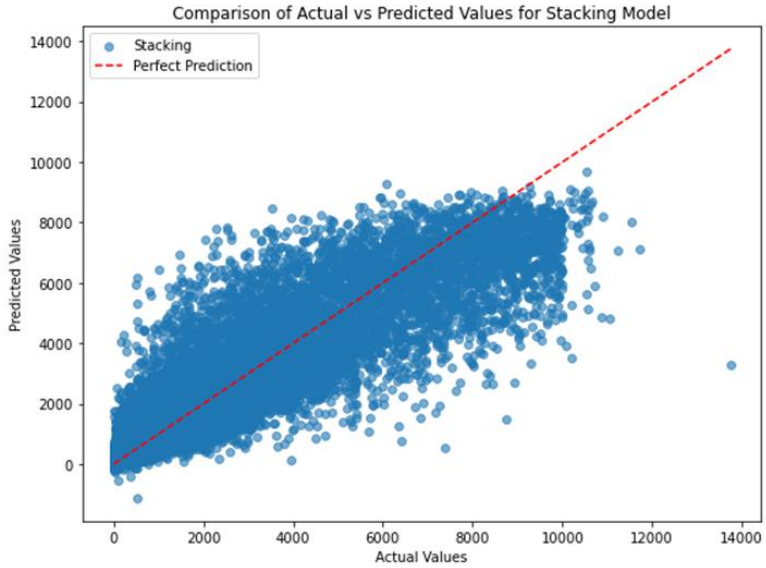

Fig. 7. Comparison of the Predicted and Actual Values for the Stacking Model

In general, the research findings indicated the superior performance of Stacking in forecasting students' national ranks using the strengths of several models instead of single models and with the help of a meta-model. Accordingly, this method can be a practical solution in educational guidance systems to perform better estimations and assist students. The stacking model achieved better predictive performance than the individual models; however, it requires more computational resources. In addition, we need to train several base models in isolation and then train the meta-model on their predictions, leading to more training time and memory consumption. Nevertheless, the increased computational overhead is justified in practical applications by the higher prediction accuracy and model robustness. Stacking offers a useful trade-off in scenarios where predictive performance is highly valued and computational resources are available.

5. Conclusions

This study proposed an ensemble learning model following the Stacking approach for predicting the national rank of students in the national university entrance exam, known as *Konkur*. Advanced XGBoost, LightGBM, and CatBoost were combined with a linear regression meta-model. We used the optimization framework of hypermeters with Optuna, a component of our paper's proposed method, to enhance the models' performance and determine the optimal settings for them. The results showed that the stacking ensemble learning model, combining advanced XGBoost, LightGBM, CatBoost, and linear regression models, had the best performance in predicting national rankings, providing high accuracy, with NRMSE and R^2 values equal to 0.0659 and 0.7735, respectively. On the other hand, single basic models such as linear regression, random forest, gradient boosting, and MLP regressors could not adequately model the complexities in the data and, as a result, had lower accuracy. Butt et al.

(2023) depicted the best performance in small and large datasets of student performance prediction in the Boosting and Stacking models. Similarly, the same stacking model had the best outcome in the current research with NRMSE and an R^2 values of 0.0659 and 0.7735, respectively, confirming the results obtained by Butt et al. (2023) regarding the superior performance of Stacking. Both studies thus, showed improved prediction accuracy with combined models, particularly when the simpler models failed to return satisfactory accuracy. X. Chen et al. (2022) used a competitive clustering model for grade prediction, recording a 7.3% improvement in prediction accuracy. Although competitive clustering used in their study totally differed from Stacking, the idea behind it resembles that in this study, focusing on the prediction accuracy improvement by combining various methods. Still, depending on several strong models and a meta-model as the final decision-maker, Stacking has been more accurate than the competitive model proposed by X. Chen et al. (2022). One such example of a hybrid model is the application introduced by Sakri and Saleh (2020), who combined several algorithms, including Naïve Bayes and Multilayer Perceptron, reaching a 91.7% accuracy. As shown in the current research, combining several advanced models, such as XGBoost, LightGBM, and CatBoost, within the Stacking model contributed to an increase in the accuracy of the national ranking predictions. Yan and Liu (2020) used Adaboost, Random Forest, and SVM models and achieved superior prediction accuracy than single algorithm models. In the present study, Random Forest generated good results with NRMSE and R^2 values equal to 0.0940 and 0.7390, respectively, while the Stacking model proved to be better than Random Forest. Comparison with these simpler collective models (such as Adaboost or Random Forest) indicates that they are typically unable to compete with the combination of models (like stacking). Regarding multi-model heterogeneous collective learning models (e.g., Random Forest and Gradient Boosting), Adejo and Connolly (2017) found that the diversity of data sources could improve prediction accuracy and efficiency. The findings of the current research support the accuracy of the heterogeneous combination of various models, such as XGBoost, LightGBoost, and CatBoost, within the stacking model over basic models like Random Forest. Han et al. (2017) presented Adaboost as an ideal algorithm to predict students' performance. Nevertheless, our current findings indicate that the more sophisticated ensemble technique—stacking—outperforms other single-model approaches in predictive accuracy. This indicates that higher-complexity problems can be predicted more accurately by combining several more advanced models.

The contributions and results of this study include the creation of a model for the stacking ensemble of base learners, which are XGBoost, LightGBM, and CatBoost, as well as a meta-model for linear regression that demonstrates high accuracy in forecasting the students' national ranks. Additionally, the model's efficiency and predictive performance are increased through Optuna's involvement in automatically optimizing hyperparameters. The research also presents a viable implementation of machine learning in a controlled evaluation environment, such as a national examination, using real-world student data of more than 73,000 records..

The theoretical implication of the current research relies in developing an application of ensemble learning in educational prediction. The study substantiates the superiority of stacking, an approach to ensembling, over a combination of single models and smaller ensembles of models, thereby adding to the body of methodological knowledge about education and data science. It also indicates the usefulness of the automated optimization of hyperparameters, which can potentially take the place of manual tuning, implementing more complex and data-informed methods, including Bayesian optimization. This study also implies a theoretical undertaking to the integration of explainability and generalizability into educational AI.

The Practical implications show that educational management systems can embrace the proposed model to predict student rankings and initiate the process of intervention, individual academic plans, and resource allocation as early as possible. The tool can help academic counselors and school administrators to find students with difficulties and offer them extra help that would increase equity in education. In addition, the fact that the model is lightweight and also fits into any Python-based platform makes it fit into the existing educational software, even in resource-limited environments. Meanwhile, it has been established that constant retraining and data quality tracking are required to further develop the relevance and reliability of the model in the long term.

Reference

Abiodun, O. J., & Wreford, A. I. (2024). Student's performance evaluation using ensemble machine learning algorithms. *Engineering and Technology Journal, 09*(08). https://doi.org/10.47191/etj/v9i08.23

Aboneh, T., Rorissa, A., & Srinivasagan, R. (2022). Stacking-Based ensemble learning method for multi-spectral image classification. *Technologies, 10*(1), 17. https://doi.org/10.3390/technologies10010017

Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education, 10*(1), 61–75. https://doi.org/10.1108/jarhe-09-2017-0113

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A next-generation hyperparameter optimization framework*. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3292500.3330701

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine, 18*(3), 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Ali, R., Ali, S. K., & Afzal, A. (2019). Predictive validity of a Uniform Entrance Test for the health professionals. *Pakistan Journal of Medical Sciences, 35*(2). https://doi.org/10.12669/pjms.35.2.334

Almalawi, A., Soh, B., Li, A., & Samra, H. (2024). Predictive models for educational purposes: A systematic review. *Big Data and Cognitive Computing, 8*(12), 187. https://doi.org/10.3390/bdcc8120187

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*(1). https://doi.org/10.1186/s41239-020-0177-7

Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of machine learning models towards student academic performance prediction: A systematic review. *Applied Sciences, 11*(21), 10007. https://doi.org/10.3390/app112110007

Ballaho, J. C. (2024). *Predicting student's success in programming courses: A decision support system for admission in computer science and information technology programs*. In *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)* (pp. 60–64). Kota Kinabalu, Malaysia. https://doi.org/10.1109/IICAIET62352.2024.10729909

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Biau, G., & Cadre, B. (2021). *Optimization by gradient boosting*. In M. Lovric (Ed.), *Springer eBooks* (pp. 23–44). https://doi.org/10.1007/978-3-030-73249-3_2

Bishwakarma, S. T., & Sharma, G. (2022). Automated hyperparameter optimization in machine learning for stock prediction. *2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS)*, 1-6. https://doi.org/10.1109/icngis54955.2022.10079816

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information Knowledge and Management, 14*, 045–076. https://doi.org/10.28945/4184

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Butt, N. A., Mahmood, Z., Shakeel, K., Alfarhood, S., Safran, M., & Ashraf, I. (2023). Performance prediction of students in higher education using multi-model ensemble approach. *IEEE Access, 11*, 136091–136108. https://doi.org/10.1109/access.2023.3336987

Cai, Y., Feng, J., Wang, Y., Ding, Y., Hu, Y., & Fang, H. (2024). The optuna–lightgbm–xgboost model: A novel approach for estimating carbon emissions based on the electricity–Carbon nexus. *Applied Sciences, 14*(11), 4632. https://doi.org/10.3390/app14114632

Chaparro-Cruz, I. N., Huertas-Condori, L. N., Cabana-Yupanqui, S. B., & Chaparro-Guerra, A. (2025). Relationship between entrance exam scores, academic performance, and student dropout rates: A longitudinal case study. *International Journal of Learning, Teaching and Educational Research, 24*(3), 216–243. https://doi.org/10.26803/ijlter.24.3.11

Chen, T., & Guestrin, C. (2016, August). *XGBoost: A scalable tree boosting system*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Francisco, California, USA) (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Chen, X., Peng, Y., Gao, Y., & Cai, S. (2022). A competition model for prediction of admission scores of colleges and universities in Chinese college entrance examination. *PLoS ONE, 17*(10), e0274221. https://doi.org/10.1371/journal.pone.0274221

Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Calster, B. V., Harrell, F. E., Martin, G. P., Moons, K. G. M., Smeden, M. van, Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): From development to external validation. *BMJ, 384*, e074819. https://doi.org/10.1136/bmj-2023-074819

Daniele, V. (2021). Socioeconomic inequality and regional disparities in educational achievement: The role of relative poverty. *Intelligence, 84*, 101515. https://doi.org/10.1016/j.intell.2020.101515

Dey, R., & Mathur, R. (2023). *Ensemble learning method using stacking with base learner: A comparison*. In *Lecture Notes in Networks and Systems* (Singapore) (pp. 159–169). https://doi.org/10.1007/978-981-99-3878-0_14

Gibson, D. C., & Webb, M. E. (2015). Data science in educational assessment. *Education and Information Technologies, 20*, 697-713. https://doi.org/10.1007/s10639-015-9411-7

Furkat, B., Nasimov, R., Rashidov, A., Akhmedov, F., & Cho, Y.-I. (2024). Effective methods of categorical data encoding for artificial intelligence algorithms. *Mathematics, 12*(16), 2553–2553. https://doi.org/10.3390/math12162553

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Han, M., Tong, M., Chen, M., Liu, J., & Liu, C. (2017). *Application of ensemble algorithm in students' performance prediction*. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (Hamamatsu, Japan) (pp. 735–740).

Hancock, J. T., & Khoshgoftaar, T. M. (2020). *CatBoost for big data: An interdisciplinary review*. *Journal of Big Data, 7*(1), 1–45. https://doi.org/10.1186/s40537-020-00369-8

Huang, B., & Wang, C. (2023). *Research on data analysis of efficient innovation and entrepreneurship practice teaching based on LightGBM classification algorithm*. *International Journal of Computational Intelligence Systems, 16*(1), 1–13. https://doi.org/10.1007/s44196-023-00324-4

Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems, 200*, 105992. https://doi.org/10.1016/j.knosys.2020.105992

Jafarnejad, A., Rezasoltani, A., Khani, A. M., & Sayedeh Hoda, H. (2025). *A hybrid feature selection and classification framework for predicting entrepreneurial competency using machine learning and binary Grey Wolf Optimizer*. *Journal of Systems Thinking in Practice, 4*(4), 129–155. https://doi.org/10.22067/jstinp.2025.94947.1172

Jafarnejad, A., Rezasoltani, A., & Khani, A. M. (2025). Cost-sensitive machine learning for predicting production defects: A novel approach based on MetaCost. *Research in Production and Operations Management*, *16*(2), 73–94. https://doi.org/10.22108/pom.2025.144489.1610

Jafarnejad, A., Rezasoltani, A., & Khani, A. M. (2025). Predicting heart disease using automated machine learning based on genetic algorithms. *Journal of Information Technology Management, 17*(2), 91–122. https://doi.org/10.22059/jitm.2024.382556.3829

Kahlon, N. K., & Singh, W. (2024). Comparative analysis of web scraping tools for low-resource language text. *International Journal of Engineering Trends and Technology, 72*(1), 284–299. https://doi.org/10.14445/22315381/ijett-v72i1p128

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. *Advances in Neural Information Processing Systems, 30*, 3146–3154. https://doi.org/10.5555/3294996.3295074

Khan, Z., Ali, A., Khan, D.M. *et al.* Regularized ensemble learning for prediction and risk factors assessment of students at risk in the post-COVID era. *Sci Rep* 14, 16200 (2024). https://doi.org/10.1038/s41598-024-66894-1

Khani, A. M., Kazazi, A., & Taqhavi Fard, M. T. (2022). *Evaluating the quality of services of the cultural and social deputy of Tehran municipality in the field of culture and art*. *Social Development & Welfare Planning, 13*(50), 205–250. https://doi.org/10.22054/qjsd.2021.58035.2110

Lazcano, A., Jaramillo-Morán, M. A., & Sandubete, J. E. (2024). Back to basics: The power of the multilayer perceptron in financial time series forecasting. *Mathematics, 12*(12), 1920. https://doi.org/10.3390/math12121920

Lee, T., Ullah, A., & Wang, R. (2019). *Bootstrap aggregating and random forest*. In *Advanced Studies in Theoretical and Applied Econometrics* (Cham, Switzerland) (pp. 389–429). https://doi.org/10.1007/978-3-030-31150-6_13

Massaoudi, M., Refaat, S. S., Chihi, I., Trabelsi, M., Oueslati, F. S., & Abu-Rub, H. (2020). A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting. *Energy, 214*, 118874. https://doi.org/10.1016/j.energy.2020.118874

Navarro, C. L. A., Damen, J. A., Takada, T., Nijman, S. W., Dhiman, P., Ma, J., ... & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: Systematic review. *bmj, 375*. https://doi.org/10.1136/bmj.n2281

Parviz, M. Reflecting on the consequences of the Iranian university entrance examination: a systematic-narrative hybrid literature review. *Discov Educ* 2, 22 (2023). https://doi.org/10.1007/s44217-023-00046-x

Petro, L., & Pavlo, L. (2019). Grid search, random search, genetic algorithm: A big comparison for NAS. *ArXiv*. https://doi.org/10.48550/arXiv.1912.06059

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: Unbiased boosting with categorical features*. *Advances in Neural Information Processing Systems*, *31*, 6638–6648. https://doi.org/10.48550/arXiv.1706.09516

Rimal, Y., Sharma, N. & Alsadoon, A. The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. *Multimed Tools Appl* 83, 74349–74364 (2024). https://doi.org/10.1007/s11042-024-18426-2

Rizkallah, L. W. (2025). *Enhancing the performance of gradient boosting trees on regression problems*. *Journal of Big Data*, *12*, Article 35. https://doi.org/10.1186/s40537-025-01071-3

Sakri, S., & Saleh, A. (2020). *RHEM: A robust hybrid ensemble model for students' performance assessment on cloud computing course*. *International Journal of Advanced Computer Science and Applications*, *11*(11), 761–767. https://doi.org/10.14569/ijacsa.2020.0111150

Salari, M., Radfar, R., & Faghihi, M. (2024). Predicting students' performance using machine learning algorithms and educational data mining (A case study of Shahed University). *Business Intelligence Management Studies*, *12*(47), 315-366. https://doi.org/10.22054/ims.2023.75523.2375

Salmanpoursohi, B., Daneshvar, A., Salmanpoursohi, S., Pourghader Chobar, A., & Salahi, F. (2024). Cancer detection from textual data using a combination of machine learning approach. *Interdisciplinary Journal of Management Studies*, *17*(3), 1001–1014. https://doi.org/10.22059/ijms.2023.362252.676037

Saluja, R., Rai, M., & Saluja, R. (2023). Designing new student performance prediction model using ensemble machine learning. *Journal of Autonomous Intelligence*, *6*(1), 583–583. https://doi.org/10.32629/jai.v6i1.583

Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* **2**, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

Srinivas, P., & Katarya, R. (2022). hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost. *Biomedical Signal Processing and Control*, *73*, 103456. https://doi.org/10.1016/j.bspc.2021.103456

Sukhija, N., & Faridi, M. (2024). *Recommending graduate admission using ensemble model*. In *2024 International Conference on Computational Intelligence and Computing Applications (ICCICA)* (India) (pp. 526–530). https://doi.org/10.1109/iccica60014.2024.10584593

Taher Mazandarani, M., Zand, Z., Khodabandelou, M. H., Mozaffari, F., & Sohrabi, B. (2025). Predicting student academic performance: A machine learning approach and feature analysis. *Interdisciplinary Journal of Management Studies*, *18*(3), 425–440. https://doi.org/10.22059/ijms.2025.362506.676053

Tang, B., Li, S., & Zhao, C. (2024). Predicting the performance of students using deep ensemble learning. *Journal of Intelligence*, *12*(12), 124–124. https://doi.org/10.3390/jintelligence12120124

Teodorescu, V., & Obreja Brașoveanu, L. (2025). Assessing the Validity of k-Fold Cross-Validation for Model Selection: Evidence from Bankruptcy Prediction Using Random Forest and XGBoost. *Computation*, *13*(5), 127. https://doi.org/10.3390/computation13050127

T r, M., V, V. K., V, D. K., Geman, O., Margala, M., & Guduri, M. (2023). The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification. *Healthcare Analytics*, *4*, 100247. https://doi.org/10.1016/j.health.2023.100247

Wang, N. Z., & Shi, N. Y. (2016). *Prediction of the admission lines of college entrance examination based on machine learning*. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)* (Chengdu, China) (pp. 332–335). https://doi.org/10.1109/compcomm.2016.7924718

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Yan, L., & Liu, Y. (2020). An ensemble prediction model for potential student recommendation using machine learning. *Symmetry*, *12*(5), 728. https://doi.org/10.3390/sym12050728

Yang, H., Chen, Z., Yang, H., & Tian, M. (2023). Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison. *IEEE Access*, *11*, 23366–23380. https://doi.org/10.1109/access.2023.3253885

Yu, J., Zhao, Y., Pan, R., Zhou, X., & Wei, Z. (2023). Prediction of the critical temperature of superconductors based on two-layer feature selection and the optuna-stacking ensemble learning model. *ACS Omega*, *8*(3), 3078–3090. https://doi.org/10.1021/acsomega.2c06324

Zangooei, H., & Fatemi, O. (2021). Predicting students at risk of academic failure using learning analytics in the learning management system. *Quarterly of Iranian Distance Education Journal*, *3*(2), 32-44. https://doi.org/10.30473/idej.2022.63913.1104

Zhang, H. W., Wang, Y. R., Hu, B., et al. (2024). *Using machine learning to develop a stacking ensemble learning model for the CT radiomics classification of brain metastases*. *Scientific Reports*, *14*, 28575. https://doi.org/10.1038/s41598-024-80210-x

Zohrehvandian, K., Ghaffarian, H., & Mahmoudi, A. (2023). Predicting the level of salesperson's performance in encouraging customers to use appropriate shopping strategies in sports clubs. *Interdisciplinary Journal of Management Studies*, *17*(1), 169–183. https://doi.org/10.22059/ijms.2023.342973.675100

Zub, K., Pavlo Zhezhnych, & Strauss, C. (2023). Two-Stage PNN–SVM ensemble for higher education admission prediction. *Big Data and Cognitive Computing*, *7*(2), 83–83. https://doi.org/10.3390/bdcc7020083