Indonesian Journal of Electrical Engineering and Computer Science
Vol. 40, No. 2, November 2025, pp. 941~952
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v40.i2.pp941-952

Unveiling educational enrollment factors in Egypt via ensemble learning

Fahad Kamal Alsheref[1], Mostafa Sayed Mostafa El Misery[2], Mahmoud Mohamed Bahloul[3], Dalia A. Magdi[4], Ibrahim Eldesouky Fattoh[5]
[1]Department of Informatics for Business, College of Business, King Khalid University, Abha, Saudi Arabia
[2]Department of Statistics, Faculty of Economics and Political Science, Cairo University, Egypt
[3]Information Systems Department, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt
[4]Faculty of Computers and Information Specialties, Sadat Academy for Management Sciences, Cairo, Egypt
[5]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt

Article Info

Article history:

Received Nov 5, 2024
Revised Jul 27, 2025
Accepted Oct 14, 2025

Keywords:

Educational enrollment factors
Ensemble learning
Machine learning
Predictive modeling
Socio-economic factors

ABSTRACT

Education plays a vital role in the development of a nation and significantly influences the direction of societies. Understanding the various factors that impact educational enrollment is essential for policymakers and resource allocation strategies. This paper explores the factors impacting educational enrollment in Egypt using predictive modeling and machine learning techniques. The study evaluates six machine learning algorithms and ensemble learning approaches to predict enrollment rates, considering computational efficiency, robustness, and parameter sensitivity. By analyzing socio-economic and demographic indicators from Egyptian educational data, the research examines the interplay of these factors. Results highlight the effectiveness of these methods in elucidating enrollment patterns, with ensemble learning showing promising performance and significant improvements compared to traditional machine learning algorithms. This study offers insights into Egypt's educational landscape that could inform policy formulation and resource allocation strategies.

Corresponding Author:

Ibrahim Eldesouky Fattoh
Department of Computer Science, Faculty of Computers and Artificial Intelligence
Beni-Suef University, Beni-Suef, Egypt
Email: ibrahim_desoky@fcis.bsu.edu.eg

1. INTRODUCTION

Education plays a crucial role in shaping the development of a nation and significantly influences societal trajectories. Understanding the various elements that impact educational participation is essential for policymakers and resource allocation strategies. In Egypt, an evolving demographic landscape and ambitious developmental aims present multiple challenges to the educational system, including economic disparities and cultural norms [1]-[3].

Various factors influence educational enrollment in Egypt, including socio-economic status, gender, and geographical location. In recent years, machine learning techniques have emerged as powerful tools for analyzing complex data patterns and making accurate predictions. These techniques have been successfully applied in various fields, including education, to uncover hidden insights and inform decision-making processes. However, there is limited research on the application of machine learning in understanding educational enrollment factors in Egypt [4], [5]. To further enhance the understanding of educational enrollment factors in Egypt, this study will delve into the ensemble learning approach. Ensemble learning has shown promising performance and significant improvements compared to traditional machine learning

algorithms [6], [7]. By incorporating ensemble methods such as bagging, boosting, and stacking, the analysis aims to capture the complexities and interactions among various factors affecting educational enrollment [1], [8]. This comprehensive approach will provide a robust and nuanced understanding of the educational landscape in Egypt, offering insights that can potentially drive policy formulation and resource allocation strategies. In this research, we make use of a thorough dataset obtained from the Egyptian Demographic and Health Survey (EDHS). This dataset, carefully assembled by EDHS, covers a wide range of factors related to educational enrollment in various regions within Egypt. The dataset includes data on socioeconomic measures, demographic traits, and geographic elements that could impact educational enrollment trends. The collection and curation of the dataset were diligently done by EDHS to ensure its reliability and accuracy for our analysis. It encompasses multiple regions within Egypt, providing a representative sample that takes into account the diverse socio-cultural and economic environment of the country. Additionally, the dataset demonstrates a normal distribution, which enhances the robustness of our analysis efforts. By utilizing this comprehensive and extensive dataset, we aim to carry out an in-depth exploration of the determinants affecting educational enrollment in Egypt. Our analysis aims to reveal underlying patterns, trends, and connections within the data set - offering insight into factors influencing access to education across different areas in Egypt. Through this investigation, we aim to provide valuable insights that can support evidence-based policymaking initiatives focused on promoting equal access to education throughout Egypt.

The rest of this paper is organized as follows: Section 2 describes the Materials and methods employed in this study, including data collection, preprocessing, and the ensemble learning algorithm used. Section 3 introduces the different proposed models suggested in this study to be implemented. Section 4 presents the results of our analysis, including key findings and insights gained from the dataset. Finally, Section 5 concludes the paper with a summary of the main findings and their implications for educational policies in Egypt. The conclusion of this paper underscores the importance of utilizing accurate and comprehensive datasets in research to generate insights that can inform policy-making and resource allocation strategies.

2. METHODS

This section presents the materials and techniques used in our research, establishing the basis for a methodical and meticulous examination of the factors that impact educational enrollment in Egypt.

2.1. Machine learning algorithm

2.1.1. Support vector machine (SVM)

Support vector machines are supervised learning algorithms with specific learning approaches suitable for classification and regression tasks. In simpler terms, they can also be seen as discriminative classifiers that utilize a unique separating hyperplane. This hyperplane is drawn in two-dimensional space to distinguish between two classes, where each side of the hyperplane represents a different class. The main goal extends to drawing a hyperplane in N-dimensional spaces (where N corresponds to the number of variables) for effective data point classification. Hyperplanes are positioned at the maximum distance from data points in order to improve the accuracy of classifying future points. Furthermore, SVM has the ability to perform nonlinear classification using the Kernel trick, which involves mapping data into higher-dimensional feature spaces so that it can be separated by a binary classifier [9]. SVM works by transforming input samples from their original space into a high-dimensional function space while identifying optimal hyperplanes and selecting samples within this process. The most favorable scenario occurs when finding an optimal margin separates classes effectively within this space; here, "margin" refers to the average distance between parallel planes on either side of the optimal plane, without any sample instances, which allows better generalization error based on risk minimization theory, considering wider margins correlate with better failure prediction functions, misclassified samples [10].

2.1.2. Random forest (RF)

A random forest acts as a meta-estimator formed from multiple decision trees working together as an ensemble. Each tree is created independently, using bootstrapping and random selection of features, which leads to a diverse set of correlated trees within the forest [11], [12]. The combined prediction of these trees tends to be more dependable than that of any single tree. In a random forest, each tree contributes to predicting the class, and the class with the highest number of votes is chosen as the model's prediction. The power of the random forest model lies in its collaboration among numerous relatively uncorrelated trees, forming a committee that outperforms individual models. Random forests are an important ensemble technique based on the CART algorithm and employ a unique approach to grow their trees. Instead of constructing just one decision tree for prediction, they combine multiple trees into a robust ensemble through

bagging-involving randomly selecting large portions of training data samples and replacing them with equally sized samples. Then, applying a CART-like algorithm to each bootstrap sample constructs a decision tree resulting in variously composed subsets, leading to the creation of distinct collection variations within it [13].

2.1.3. Neural network

The examination of behavioral simulation methods is a key focus in the field of artificial intelligence research. AI has sought to replicate the structure and functions of the human brain since the 1950s, drawing inspiration from brain research. In 1969, Minsky and Papert demonstrated that training algorithms could calculate linearly separable functions. However, it wasn't until 1986 that Rummelhart et al. introduced the error-back propagation algorithm (backpropagation), initially proposed by Werbos in 1974, which revitalized the slow progress of artificial intelligence at that time. Currently, advanced higher-dimensional neural networks such as multilayer perceptrons are being developed for tasks like pattern recognition, image processing, speech recognition, as well as for optimizing processes, controlling systems, and diagnosing and predicting various outcomes [14]. Normally, the multilayer perceptron learns by adjusting the weights, connections, and parameters of a selected activation function, like the logistic function applied to all links. The backpropagation algorithm is used for training this MLP [15]. In essence, the architecture of the MLP consists of 31 neurons in the input layer, 16 neurons in the first hidden fully connected layer, and one neuron in the output layer. The remaining hidden layer neurons are randomly initialized within a range of minus one to one. During the supervised learning phase, using assigned prediction weights and probability estimates based on provided training data#, forecasts student outcomes [15].

$$e_t = \sum_i (\tilde{y}_{it} - y_{it})^2 \quad (1)\ [14]$$

e_t : # Error or loss function determined as the sum of squares of these errors.an error term which measures disparity between actual study outcome from training data and the predicted outcome from neural network.

$\tilde{y}_{it}$: One benefit of supervised learning is assigning a prediction algorithm.

y_{it} : actual outcomes, the process involves initializing the input layer neurons with training data, including determinant variables as external inputs.

The error function has the benefit of being continuously differentiable, making it easier to adjust weights during training. Backpropagation takes advantage of this by optimizing weights to allow the neural network to map inputs to outputs effectively, progressively reducing the error function at each iteration. As a result, backpropagation utilizes error values for computing the gradient of the loss function, which helps in searching for the minimum value of the error function. The resulting weights in a neural network are similar to coefficients in a linear regression model, but there is a considerably larger number of weights compared to coefficients, posing challenges when interpreting them within a neural network [16].

2.1.4. K-nearest neighbor (KNN)

In the field of machine learning and data mining, one widely used method for prediction is the K-nearest neighbor technique. Known for its adaptability and simplicity, KNN can handle diverse types of data in the prediction process [17]. The method was first introduced by E. Fix and J.L. Hodges in an unpublished report for the US Air Force School of Aviation Medicine in 1967, where they formalized its main properties and original concept. KNN operates as a lazy or instance-based method, which means it does nothing the field of machine learning and data mining, the K-nearest neighbor method is widely used for prediction. It is known for its adaptability and simplicity in handling various types of data during the prediction process [18]. This approach was initially introduced by E. Fix and J.L. Hodges in an unpublished report for the US Air Force School of Aviation Medicine back in 1967, where they formalized its main properties and original idea. KNN works as a lazy or instance-based method, meaning it doesn't require building a model to represent the underlying distribution and statistics of the original training data; instead, it directly utilizes the training data along with their actual instances. The formal definition of KNN was later provided by Covert and Hart. The classic KNN algorithm is primarily utilized for classification tasks within supervised machine learning techniques. It uses a flexible parameter denoted as "k," indicating the number of 'nearest neighbors' considered when operating on identifying nearest neighbor(s) from a given query based on proximity within a training dataset [19]. Once these k nearest neighbors are identified, the majority voting rule is employed by this algorithm to determine which class occurs most frequently among them, thus designating this class with the highest frequency as the final classification assigned to that particular query point [20].

2.1.5. Naïve Bayes classifier

A Naïve Bayes classifier is a method that utilizes Bayes' theorem for basic probabilistic classification. It operates under the assumption that the presence or absence of one feature of a class is independent of the presence or absence of any other feature. This approach allows for predicting the probability of a class in future instances, and it shares similarities with decision trees and neural networks. Known for its high accuracy and speed, especially when dealing with large databases, Naïve Bayes has advantages such as being relatively easy to comprehend and implement. Additionally, it excels at generating class predictions more quickly than other classification algorithms while also being effective for training with small datasets [21].

2.1.6. Logistic regression

Logistic regression is a well-established statistical model utilized for binary classification tasks. It involves non-random function variables and represents the class response through a binary random variable with specific probabilities. The success probability, denoted as pp, depends on the feature variables and forms a linear function using the log odds ratio or the logarithm of the odds ratio of predictor variables. Logistic regression incorporates hypothesis testing, diverse evaluations, calculations, fitness measures for each variable's value, and employs variable importance checking for feature selection in classification. Modern computer implementations often include multiple iterations of stepped variable selection to enhance its effectiveness. Due to its mathematical similarity with ordinary multiple regression and automated variable selection ease, logistic regression remains one of the most commonly used data mining methods [22].

2.2. Ensemble techniques

Ensemble methods utilize multiple machine learning models to enhance overall performance by leveraging the diverse perspectives of individual models. This approach helps mitigate biases and errors, leading to more resilient and precise predictions. Techniques like bagging, boosting, and stacking are successful in capturing intricate data relationships and improving predictive abilities. By combining outputs from different models, ensembles can achieve superior performance compared to using a single model alone. These widely adopted techniques have applications across domains such as classification, regression, and anomaly detection for achieving cutting-edge outcomes [23]. Ensemble methods, such as bagging, voting, and stacking, are widely used in machine learning. Bagging involves training multiple models in parallel on different subsets of the training data and then combining their predictions through averaging or voting [24].

2.2.1. Voting

Voting ensemble methods in machine learning in the field of machine learning, voting ensemble techniques are powerful methods used to integrate the predictions from multiple base models for making a final decision, as shown in Figure 1. These methods are particularly effective in classification assignments. The ensemble typically comprises diverse base models, each trained on a subset of training data or using different algorithms. During inference, each base model independently forecasts the class label for a given input, and the ultimate prediction is made through a voting mechanism. There exist various types of voting strategies, such as majority voting - where the most frequently predicted class label is chosen, and weighted voting - where each model's prediction carries weight based on its performance or confidence level. Voting ensemble techniques leverage individual models' collective knowledge to enhance predictive accuracy and resilience by mitigating biases and errors inherent in any single model. They find wide application across different domains and have shown considerable success in enhancing classification performance [25].

2.2.2. Stacking

Stacking ensemble methods, also referred to as stacked generalization, present a robust approach in ensemble learning by amalgamating predictions from multiple base models. In contrast to conventional voting techniques, stacking entails training a meta-learner (blender or combiner) that effectively integrates the outputs of the base models shown in Figure 2. The process involves dividing the dataset into training and validation sets. Subsequently, base models are trained on the training set to generate predictions for the validation set. These predictions function as features for the meta-learner's input [26]. Following this, the meta-learner is trained on the validation set using these predicted values from various base models along with true labels as targets. The role of this learner lies in understanding how to weigh and integrate such model predictions adeptly while maximizing their strengths and minimizing weaknesses. During inference stages, new data undergoes prediction generation via base models; these are then aggregated by a trained meta-learner to derive an ultimate prediction [24-26]. Notably beneficial attributes of stacking ensemble methods encompass their capability to capture intricate data relationships and potentially surpass individual basic model performances. Nonetheless, they necessitate meticulous tuning and entail higher computational costs

compared to simpler voting approaches. Nevertheless, stacking has demonstrated high effectiveness in enhancing predictive outcomes across diverse machine learning tasks and domains [27].

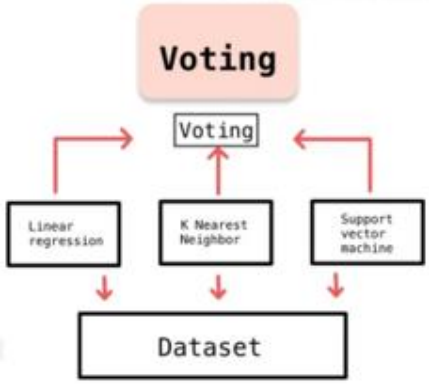


Figure 1. Ensemble (Voting) [25]

2.2.3. Bagging

Ensemble Technique: Bagging, also known as Bootstrap Aggregating, is a robust method in ensemble learning that is employed to enhance predictive accuracy and mitigate overfitting. In this technique, several individual base models are trained on distinct bootstrap samples of the training dataset. These samples are created by randomly selecting data points with replacement [23]. During the training process, each base model is trained to make predictions on a specific bootstrap sample, which introduces variability among the models. This variation aids in minimizing the variance in individual models and lowers the potential for overfitting to the training data, as shown in Figure 3 [6]. The bagging ensemble combines predictions from the base models using averaging or voting to produce the final prediction, thereby enhancing the overall accuracy and resilience of the ensemble model [8]. Bagging is most beneficial when paired with high-variance, low-bias models like decision trees. It involves training numerous trees on diverse subsets of the data and then averaging their predictions. This process results in a more dependable and precise model compared to using a single decision tree [23].

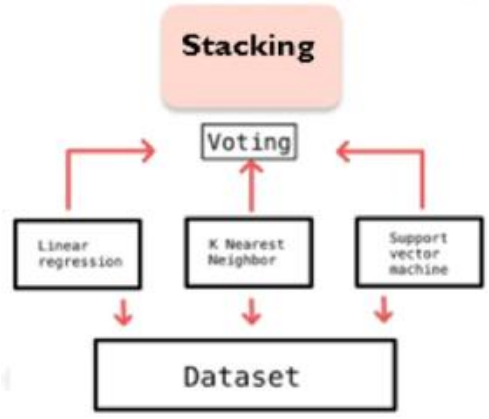


Figure 2. Ensemble (Stacking) [26]

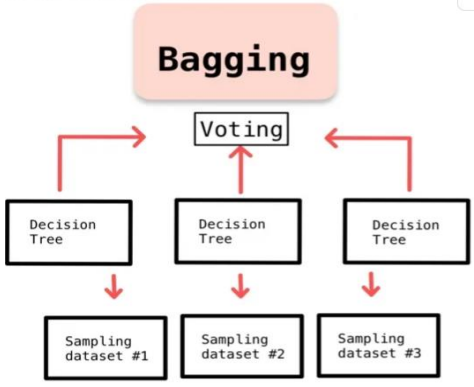


Figure 3. Ensemble (Bagging) [23]

3. PROPOSED MODEL

In this section, we present the findings of a series of experiments carried out to assess different machine learning approaches for solving a specific task. Five separate proposed models were conducted to explore various methodologies and techniques. Initially, traditional ML algorithms were utilized to establish a baseline performance. Then, ensemble learning techniques such as model voting, bagging, and stacking were investigated in the second experiment to evaluate their effectiveness in improving predictive accuracy. The third and fourth proposed models focused on analyzing the individual performances of bagging and

voting methods, respectively. Lastly, the fifth experiment is a multi-stage ensemble combining bagging and voting techniques in a novel hybrid approach with the objective of leveraging both paradigms' strengths. These models are tested to provide insights into the effectiveness and comparative benefits of these ML methodologies for addressing our task at hand. The proposed models are described in Figures 4-8 as shown below:

3.1. Traditional ML proposed model

In this proposed model, we executed each algorithm separately and assessed the metrics for each one to identify the factors that influence learning participation. We executed each algorithm separately in order to assess the metrics for each one and identify the factors that influence learning participation. Furthermore, we conducted statistical analyses to determine the significance of these factors in predicting learning enrollment and performance.

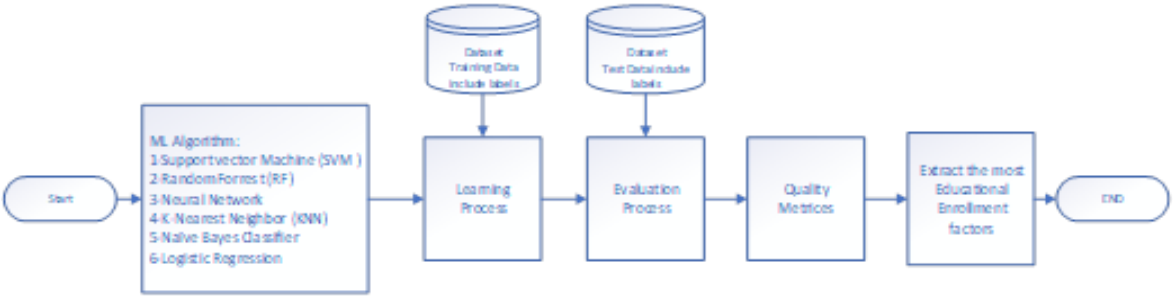


Figure 4. Traditional ML model for determining learning enrollment in Egypt

3.2. Ensemble (Voting) proposed model

In this proposed model, we focused on ensemble voting for ML algorithms and analyzed its performance measures to uncover the variables that impact participation in learning. We compared the performance metrics with those of other proposed models to determine if there are enhancements between traditional methods and ensemble methods. Additionally, we identified important factors influencing involvement in learning activities.

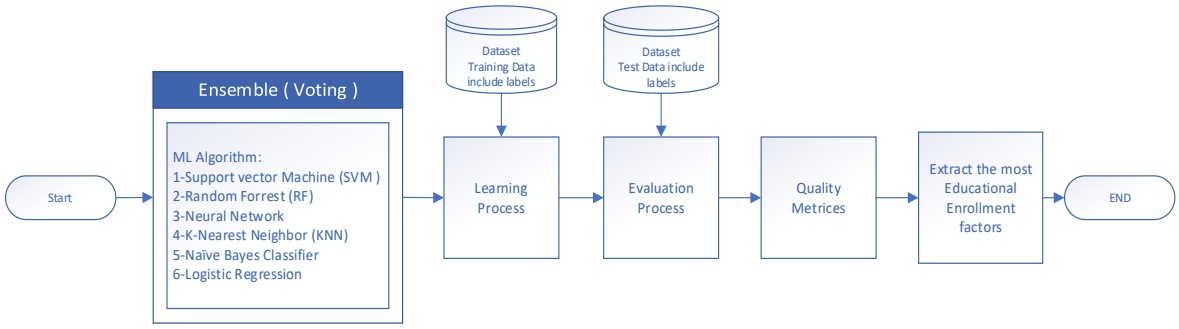


Figure 5. Ensemble (Voting) model for determining learning enrollment in Egypt

3.3. Ensemble (Bagging) proposed model

In this proposed model, we utilized bagging ensemble techniques for each individual machine learning algorithm and evaluated the performance measures to determine the factors affecting engagement in learning. Each algorithm was run with a bagging ensemble separately to analyze the performance metrics and identify the determinants of participation in learning activities. A comparison was made between the performance metrics of this proposed model and others to ascertain any improvements gained from using traditional methods versus ensemble methods. Furthermore, significant factors impacting involvement in learning activities were identified during the analysis.

3.4. Ensemble (Stacking) proposed model

In this study, we employed stacking ensemble methods to merge predictions from multiple machine learning algorithms and assessed their combined performance metrics in order to identify the factors influencing engagement in learning. The ensemble model, created by combining outputs from various base learners, was trained to comprehend the intricate relationships within the data and generate precise predictions. Through this method, we evaluated the overall effectiveness of using stacking ensembles to enhance learning participation. Additionally, we compared the performance metrics of the stacking ensemble with those of conventional methods to determine any notable improvements. Our analysis also aimed at uncovering factors that affect participation in learning activities using the stacking ensemble methodology.

Figure 6. Ensemble (Bagging) model for determining learning enrollment in Egypt

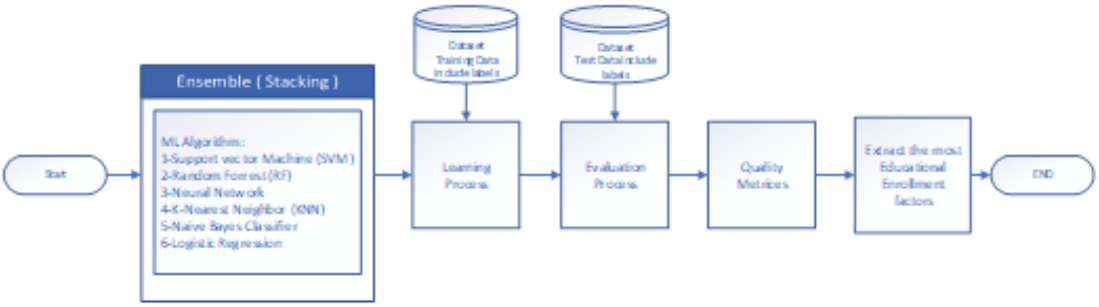


Figure 7. Ensemble (Stacking) model for determining learning enrollment in Egypt

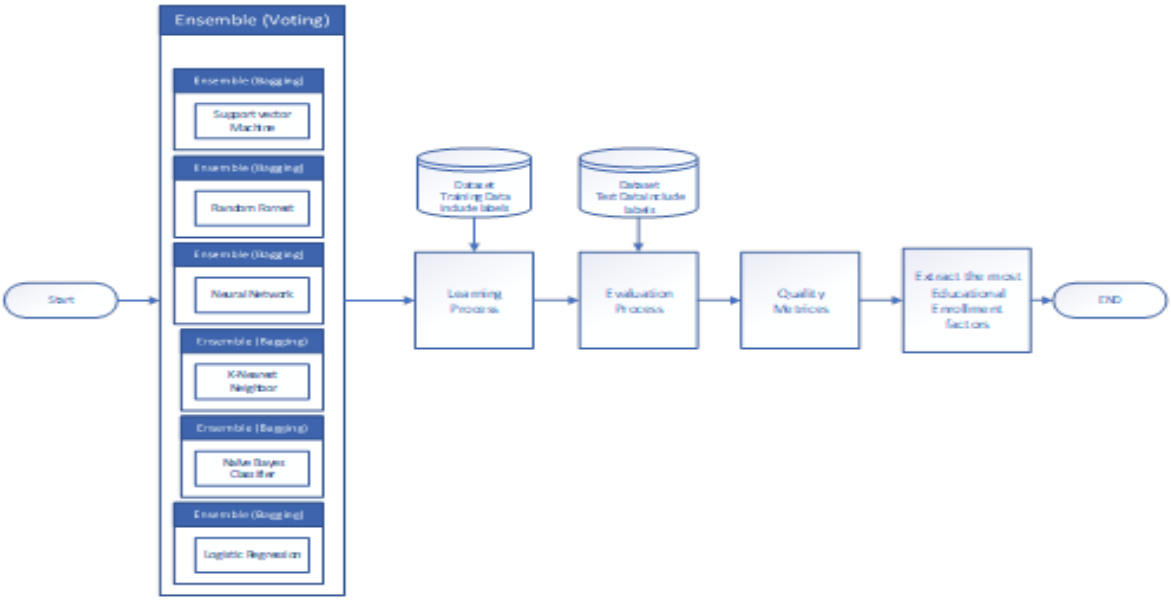


Figure 8. Ensemble (Bagging and Voting) model for determining learning enrollment in Egypt

3.5. Multi-stage ensemble (Voting and Bagging) proposed model

In this proposed model, we utilized a multi-stage ensemble strategy that incorporated bagging for each individual machine learning algorithm, followed by a voting mechanism. Initially, bagging was used independently for each algorithm to generate diverse sets of models trained on different subsets of the data. Subsequently, these bagged models were combined using a voting scheme to produce collective predictions. We aimed to take advantage of both bagging and voting techniques to improve predictive performance and robustness. Our assessment focused on analyzing the combined impact of bagging and voting in enhancing learning engagement. Additionally, we conducted a comparative analysis to evaluate the performance metrics of this ensemble approach against conventional methods, providing insights into its effectiveness in identifying factors influencing learning participation.

4. RESULTS AND DISCUSSION

We showcase the outcomes of our suggested models, starting with executing each machine learning algorithm separately and subsequently employing ensemble learning methods like bagging and stacking for various ML algorithms. Following the integration of bagging for each specific algorithm, we introduced a voting system to consolidate predictions. Finally, we analyzed the influential elements impacting enrollment in education extracted from the proposed models.

4.1. Experiment results of machine learning algorithm models

Table 1 presents the performance of six machine learning algorithms in predicting educational enrollment rates in Egypt. SVM achieved 77.95% accuracy, with 80.72% precision and 82.66% recall. Random forest attained 78.08% accuracy, with 75.87% precision and a notably high 92.06% recall. Neural network models showed promise with 79.87% accuracy, 83.30% precision, and 82.2% recall. KNN reached 80.27% accuracy, 80.37% precision, and 87.87% recall. Naïve Bayes classifier and logistic regression displayed similar accuracies (79.08% and 80.10%, respectively) and closely matched precision and recall scores. Based on the provided metrics, it appears that KNN demonstrates the most balanced performance among the evaluated machine learning algorithms for predicting educational enrollment rates in Egypt. With an accuracy of 80.27%, KNN achieves a competitive performance in correctly classifying enrollment statuses. Additionally, KNN exhibits balanced precision (80.37%) and recall (87.87%), suggesting its ability to minimize both false positives and false negatives effectively.

Table 1. Results of traditioal machine learning algorithms

Model	Accuracy	Precision	Recall	F Measure	AUC	Classification error	Sensitivity	Standard deviation
1-SVM	**77.95%**	**80.72%**	**82.66%**	**81.67%**	**80.54%**	**22.05%**	**82.66%**	0.70%
2-RF	78.08%	75.87%	92.06%	83.19%	86.76%	21.92%	92.06%	0.59%
3-Neural network	79.87%	83.30%	82.22%	82.76%	87.68%	20.13%	82.22%	0.59%
4-KNN	80.27%	80.37%	87.87%	83.95%	88.13%	19.73%	87.87%	0.57%
5-Naïve Bayes classifier	79.08%	79.89%	86.15%	82.90%	85.83%	20.92%	86.15%	0.47%
6-Logistic regression	80.10%	79.68%	88.88%	84.03%	87.65%	19.90%	88.88%	0.70%

4.2. Experiment results of the ensemble voting model

The ensemble voting model presents compelling results in predicting educational enrollment rates in Egypt. Table 2 shows an accuracy of 82.13%; the model demonstrates its capability to accurately classify enrollment statuses. Additionally, achieving a precision of 79.22% indicates its proficiency in correctly identifying enrolled students among the predicted positives. The model's recall score of 76.11% suggests its effectiveness in capturing a significant proportion of enrolled students out of all actual enrolled students, thereby minimizing false negatives. These outcomes highlight the ensemble voting model's robustness and balanced performance, offering valuable insights into educational enrollment dynamics and providing a foundation for informed policymaking in Egypt.

4.3. Experiment results of the ensemble bagging model

The ensemble Bagging Model achieves an accuracy of 82.75% as shown in Table 3, with a precision of 85.25%, indicating its proficiency in accurately identifying enrolled students. While its recall score of 70.42% suggests some missed enrollments, it maintains a balanced F-measure of 77.06%. With an AUC score of 85.10% and low classification error of 17.25%, the model demonstrates its effectiveness in

distinguishing between positive and negative cases. Its high sensitivity score of 81.89% and low standard deviation of 0.63% further affirm its reliability and consistency in predicting enrollment patterns.

Table 2. Results of the ensemble voting model

Model	Accuracy	Precision	Recall	F Measure	AUC	Classification error	Sensitivity	Standard deviation
Ensemble voting model	82.13%	79.22%	76.11%	77.50%	83.80%	17.87%	82.89%	0.60%

Table 3. Results of the ensemble bagging model

Model	Accuracy	Precision	Recall	F Measure	AUC	Classification error	Sensitivity	Standard deviation
Ensemble bagging model	82.75%	85.25%	70.42%	77.06%	85.10%	17.25%	81.89%	0.63%

4.4. Experiment results of ensemble stacking model

The ensemble stacking model's strong performance metrics, as shown in Table 4, highlight its proficiency in predicting educational enrollment rates in Egypt. Achieving an accuracy of 84.26% and a precision of 85.72%, the model demonstrates its capability in accurately identifying enrolled students. Despite a recall score of 70.42%, the model maintains a balanced F-measure of 77.34%. With an AUC score of 85.70% and a low classification error of 15.74%, it effectively distinguishes between positive and negative cases. Additionally, the model's sensitivity score of 82.87% and low standard deviation of 0.64% further affirm its reliability and consistency in predicting enrollment patterns.

Table 4. Results of the ensemble stacking model

Model	Accuracy	Precision	Recall	F Measure	AUC	Classification error	Sensitivity	Standard deviation
Ensemble stacking model	84.26%	85.72%	70.42%	77.34%	85.70%	15.74%	82.87%	0.64%

4.5. Experiment results of multi-stage ensemble model

The multi-stage ensemble model, incorporating both Voting and Bagging techniques, as depicted in Table 5, emerges as a standout performer in predicting educational enrollment rates in Egypt. With an impressive accuracy of 85.25% and precision of 86.52%, the model excels in accurately identifying enrolled students, surpassing previous models' performance. Notably, despite achieving a recall score of 73.42%, the model maintains a balanced F-measure of 79.45%, indicative of its effectiveness in capturing enrolled students while minimizing false positives. These results underscore the superiority of the Multi-stage Ensemble model over previous models, highlighting its efficacy in providing valuable insights for educational policymakers and stakeholders in Egypt.

Table 5. Results of ensemble multi-stage model

Model	Accuracy	Precision	Recall	F Measure	AUC	Classification error	Sensitivity	Standard deviation
Multi-stage ensemble (Voting and Bagging)	85.25%	86.52%	73.42%	79.45%	84.23%	14.75%	84.02%	0.60%

4.6. Influential factors affecting education enrollment

The analysis conducted using the multi-stage ensemble model reveals the most influential factors affecting educational enrollment rates in Egypt. Among the attributes considered, husband or partner's education level emerges as the most significant factor, accounting for 13.55% of the overall predictive power. Following closely behind are the total number of children ever born and the wealth index, with weights of 9.27% and 7.95% respectively. Additionally, the age of the respondent at first birth, main floor material, type of toilet facility, and household telephone ownership are identified as notable factors, each contributing to the predictive power with weights ranging from 3.61% to 6.32%. In Figure 9, these findings shed light on the multifaceted nature of the socio-economic and demographic factors influencing educational enrollment,

providing valuable insights for policymakers and stakeholders in devising targeted interventions and resource allocation strategies to enhance educational access and equity in Egypt.

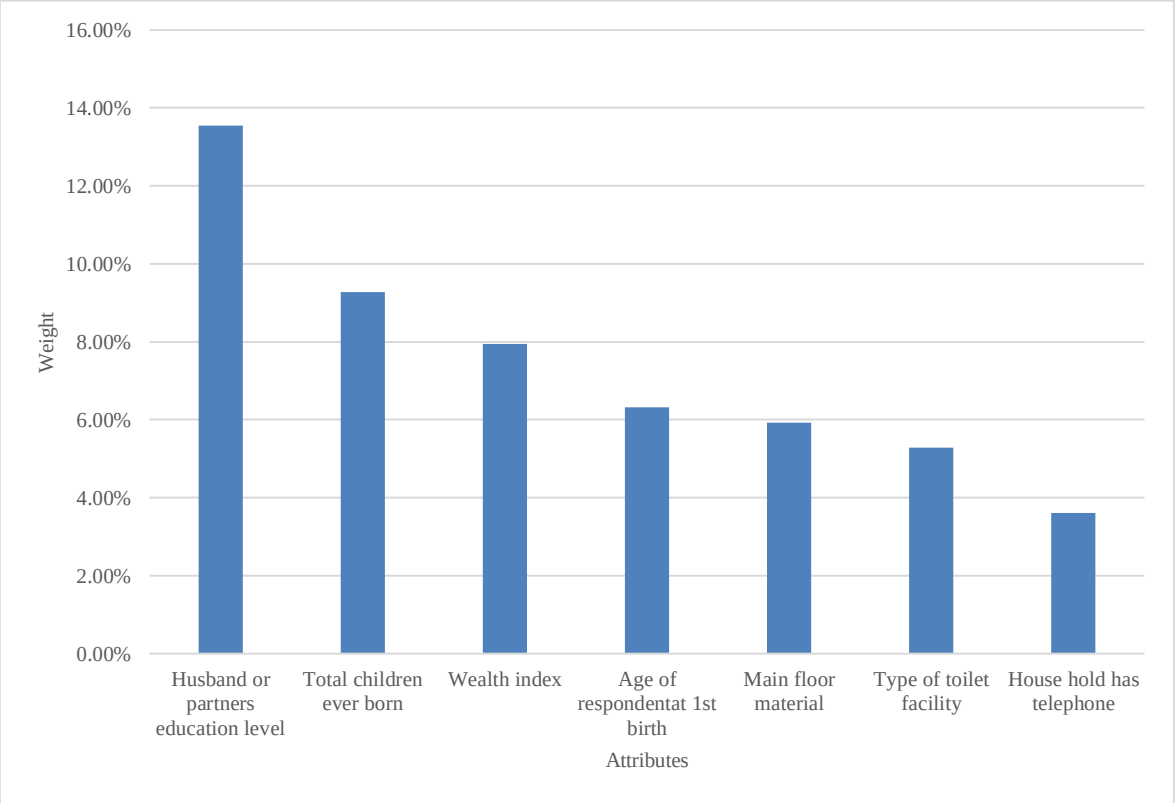


Figure 9. Influential factors affecting education enrollment

5. CONCLUSION

In conclusion, our research comprehensively examined the factors influencing educational enrollment in Egypt using advanced predictive modeling techniques. Through rigorous analysis of various machine learning algorithms and ensemble methods, we identified the most effective models and factors impacting enrollment rates. The Multi-stage Ensemble model emerged as the top performer, highlighting the significance of socio-economic indicators such as partner's education level, total children ever born, and wealth index. These findings offer valuable insights for policymakers, guiding efforts towards enhancing educational access and equity in Egypt.

FUNDING INFORMATION

The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Small Research Project under grant number RGP1/38/46.

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] M. A. Z. Ewiss, F. Abdelgawad, dan A. Elgendy, "School educational policy in Egypt: societal assessment perspective," *Journal of Humanities and Applied Social Sciences*, vol. 1, no. 1, pp. 55–68, Jun. 2019, doi: 10.1108/jhass-05-2019-004.

[2] K. S. Selim dan S. S. Rezk, "On predicting school dropouts in Egypt: A machine learning approach," *Education and Information Technologies*, vol. 28, no. 7, pp. 9235–9266, Jan. 2023, doi: 10.1007/s10639-022-11571-x.

[3] S. S. Al-Qudsi, "Family background, school enrollments and wastage: evidence from Arab countries," *Economics of Education Review*, vol. 22, no. 6, pp. 567–580, Dec. 2003, doi: 10.1016/s0272-7757(03)00028-1.

[4] Tansel dan A. D. Güngör, "Schooling investments and gender gap in schooling in MENA countries: An International Perspective," *Social Science Research Network*, Jan. 2001, doi: 10.2139/ssrn.267050.

[5] Alshanqiti dan A. Namoun, "Predicting student performance and its influential factors using hybrid regression and multi-label classification," *IEEE Access*, vol. 8, pp. 203827–203844, Jan. 2020, doi: 10.1109/access.2020.3036572.

[6] M. Injadat, A. Moubayed, A. B. Nassif, dan A. Shami, "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," *Applied Intelligence*, vol. 50, no. 12, pp. 4506–4528, Jul. 2020, doi: 10.1007/s10489-020-01776-3.

[7] H. Hassan, N. B. Ahmad, dan S. Anuar, "Improved students' performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining," *Journal of Physics: Conference Series*, vol. 1529, no. 5, p. 052041, May 2020, doi: 10.1088/1742-6596/1529/5/052041.

[8] E. A. Amrieh, T. Hamtini, dan I. Aljarah, "Mining educational data to predict student's academic performance using ensemble methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119–136, Aug. 2016, doi: 10.14257/ijdta.2016.9.8.13.

[9] D. Nurfadzilah dan A. Kesumawati, "Classification of student grade based on academic records using support vector machine," 2020, doi: 10.2991/assehr.k.201010.029.

[10] S. Aksenova, D. Zhang, dan M. Lu, "Enrollment prediction through data mining," Sep. 2006, doi: 10.1109/iri.2006.252466.

[11] Statnikov, L. Wang, dan C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC Bioinformatics*, vol. 9, no. 1, 2008, doi: 10.1186/1471-2105-9-319.

[12] X. Chen dan H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, Jun. 2012, doi: 10.1016/j.ygeno.2012.04.003.

[13] "Random Forest," Wikipedia, 07-Jul-2018. [Online]. Available: https://wikipediaquality.com/wiki/Random_forest.

[14] D. E. Rumelhart, G. E. Hinton, dan R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[15] Almási, S. Woźniak, V. Cristea, Y. Leblebici, dan T. Engbersen, "Review of advances in neural networks: Neural design technology stack," *Neurocomputing*, vol. 174, pp. 31–41, Jan. 2016, doi: 10.1016/j.neucom.2015.02.092.

[16] "Backpropagation neural networks," ScienceDirect. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016974399380052J.

[17] Triguero, J. Maillo, J. Luengo, S. García, dan F. Herrera, "From big data to smart data with the k-nearest neighbours algorithm," 1 Dec. 2016, doi: 10.1109/ithings-greencom-cpscom-smartdata.2016.177.

[18] M. Huang, R. Lin, S. Huang, dan T. Xing, "A novel approach for precipitation forecast via improved K-nearest neighbor algorithm," *Advanced Engineering Informatics*, vol. 33, pp. 89–95, Aug. 2017, doi: 10.1016/j.aei.2017.05.003.

[19] "Data mining algorithms in R/Classification/kNN," Wikibooks, 09-Jan-2016. [Online]. Available: https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/kNN.

[20] L. Jiang, H. Zhang, dan J. Su, "Learning k-nearest neighbor naive bayes for ranking," *Lecture Notes in Computer Science*, pp. 175–185, Jan. 2005, doi: 10.1007/11527503_21.

[21] King, D. Radev, dan S. Abney, "Experiments in sentence language identification with groups of similar languages," 2014, doi: 10.3115/v1/w14-5317.

[22] W. Hosmer dan S. Lemeshow, *Applied Logistic Regression*, 2000, doi: 10.1002/0471722146.

[23] T. G. Dietterich, "Ensemble methods in machine learning," *Lecture Notes in Computer Science*, pp. 1–15, Jan. 2000, doi: 10.1007/3-540-45014-9_1.

[24] Tuv, "Ensemble Learning," in *Springer*, Jan. 2006, doi: 10.1007/978-3-540-35488-8_8.

[25] K. Seewald dan J. Fürnkranz, "An Evaluation of Grading Classifiers," *Lecture Notes in Computer Science*, pp. 115–124, Jan. 2001, doi: 10.1007/3-540-44816-0_12.

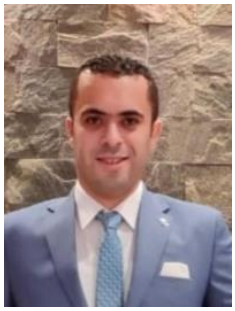
[26] O. Steinki dan Z. Mohammad, "Introduction to ensemble learning," *Social Science Research Network*, Jan. 2015, doi: 10.2139/ssrn.2634092.

[27] Jurek, Y. Bi, S. Wu, dan C. Nugent, "A survey of commonly used ensemble-based classification techniques," *The Knowledge Engineering Review*, vol. 29, no. 5, pp. 551–581, May 2013, doi: 10.1017/s0269888913000155.

BIOGRAPHIES OF AUTHORS

Assoc. Prof. Fahad Kamal Alsheref he was born in Sohag, Egypt, in 1983. He received his B.Sc. degree from the Faculty of Computers and Information, Assiut University, Egypt, in 2005. He later obtained both his M.Sc. and Ph.D. degrees in Information Systems from the Faculty of Computers and Information, Helwan University, Cairo, Egypt, in 2011 and 2012, respectively. Dr. Alsheref began his academic career at the Information Systems Department, Faculty of Computers and Information, Beni-Suef University, Egypt, where he served as an Associate Professor. During his time there, he authored and co-authored over 22 research publications in reputable peer-reviewed journals. His research spans multiple domains, including social informatics, health information systems, machine learning, and data mining. Currently, Dr. Fahad Kamal Alsheref serves as an Associate Professor at King Khalid University. He can be contacted at email: drfahad@fcis.bsu.edu.eg

Mostafa Sayed Mostafa El-Misery he was born in Cairo, Egypt, in 1981. He received his B.Sc. degree from Department of Statistics, Faculty of Economics and Political Science, Cairo University, Egypt, in 2002. He later obtained both his M.Sc. and Ph.D. degrees in Statistics from Department of Statistics, Faculty of Economics and Political Science, Cairo University, Egypt, in 2006 and 2016, respectively. He began his academic career at Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt, where he served as an Assistant Professor. During his time there, he authored and co-authored over 12 research publications in reputable peer-reviewed journals. His research spans multiple domains, including Demography, simultaneous processes, Applied statistics, and Machine learning. Currently, he serves as an assistant Professor at Cairo University. He can be contacted at email: mostafa.sayed@feps.edu.eg.



Mahmoud Mohamed Bahloul is a Lecturer at the Faculty of Commerce and Business Administration, Business Information Systems Department, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt. Deputy director of Business Information. Systems (BIS) program. Research interests: Artificial Intelligence, Data Mining, Text Mining, Business Intelligence. He can be contacted at email: mahmoud.bahloul@commerce.helwan.edu.eg.

Assoc. Prof. Dalia A. Magdi she was Vice dean of Computer Science School at the Canadian International College. She is the chair of Internet of things application and future international conference. She acted as Dean of Faculty of Management and Information Systems, she was Head of Information System Department, Faculty of Management and Information Systems, Vice-Director of CRI (Centre de Recherche Informatique), x-Coordinator of Univrsite de Paris-Sud, French University in Egypt. X Coordinator of University of New Brunswick program at Sadat Academy for Management Sciences. Member of the Editorial Board of many International journals and Reviewer of many International journals such as SCIREA Journal of Information Science, SCIREA Journal of Computer sciences, Internet of Things and Cloud Computing (IOTCC) Journal, Horizon Journal of Library and Information Science, and Journal of Computer Science and Security (JCSS) in the areas of Computer Science and Security. She published many books internationally such as a proposed enhanced model for adaptive multi-agent negotiation applied on e-commerce. She can be contacted at email: daliamagdi@gmail.com.

Assoc. Prof. Ibrahim Eldesouky Fattoh he is an Associate Professor of Computer Science at the Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt. He received his Ph.D. in Computer Science from Helwan University in 2015. His research interests include Natural Language Processing (NLP), Artificial Intelligence (AI), Machine Learning, and Large Language Models (LLMs). He has contributed to several research projects focusing on Arabic language technologies and intelligent systems. Dr. Eldesouky actively supervises graduate research and participates in academic peer review and conference organization. He can be contacted at email: ibrahim_desoky@fcis.bsu.edu.eg.