International Journal of Advanced Studies in Economics and Public Sector Management | IJASEPSM
p-ISSN: 2354-421X | *e-ISSN: 2354-4228*
Volume 12, Number 2 October, 2024

Comparing Single Classifiers, Bagging, and Voting Ensembles for Predicting Students' Academic Performance

[1]Uduehi I., [2]Ajayi H. I. & [3]Okoroego S. E.
[1&2]Department of Science Laboratory Technology,
Nigerian Building and Road Research Institute (NBRRI), Abuja
[3]Computer Science Department, Federal Polytechnic Mubi, Adamawa State

Article DOI: 10.48028/iiprds/ijasepsm.v12.i2.09

Abstract

Student learning performance is a fundamental aspect of evaluating any educational system. It forms the basis for assessing quality education. This paper highlights the critical role of student academic performance for educational institutions' achievement using educational data mining (EDM). The study analyzes and assesses student performance, proposing a predictive model based on key features such as attendance and grades. It compares various classifiers, including Bayes Network, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree to categorize student performance and predict grades. Additionally, the ensemble methods Voting and Bagging were employed to enhance classifier accuracy. In evaluating the Kaggle online dataset, Voting and Bagging achieved the highest accuracy of 68%. For the local dataset, Voting, Support Vector Machine, Decision Tree, and Random Forest achieved 100% accuracy, while Bagging and Logistic Regression followed with 89% accuracy, 100% precision, and 80% recall. Naïve Bayes had the lowest performance with 56% accuracy. These results demonstrate the effectiveness of ensemble methods in educational settings and suggest their potential for further exploration. The accuracy of these techniques depends on the available data and the nature of the task. The high accuracy of the Voting classifier in predicting academic success can help educators identify at-risk students and provide necessary support, significantly improving educational outcomes. Educators and institutions can use these findings to develop targeted interventions and support systems tailored to individual student needs, ultimately promoting academic success.

Keywords: *Bagging, Machine Learning, Voting, Student Performance, Decision Tree, Random Forest*

Corresponding Author: **Uduehi I.**

Background to the Study

Student learning performance is a fundamental aspect of evaluating any educational system. It forms the basis for assessing quality education. Higher education institutions prioritize academic performance as a key issue in delivering quality education to their students. Universities are currently facing significant challenges in attracting prospective learners, with competition intensifying as more institutions emerge (Olukoya, 2020). Nowadays, the use of advanced computational methods holds significant promise for understanding and predicting the academic performance of secondary school students. The field of educational data mining and learning analytics actively investigates various machine learning techniques to predict student outcomes (Namoun et al., 2020). Previous studies have investigated the efficacy of regression and classification models in this area, with emphasis on variables such as online learning behaviors, assessment outcomes, and the emotional factors affecting academic performance. Researchers have also examined the impact of diverse feature sets, including enrollment data, academic records, attendance records, and demographic information on accurately predicting student academic achievement (Buyrukoğlu, 2022).

One notable strategy identified to enhance prediction accuracy involves the use ensemble learning techniques, particularly bagging and voting. These methods are able combine multiple base classifiers to collectively enhance predictive performance and robustness in predicting academic outcomes. This research aims to compare bagging, voting and single classifiers in predicting academic performance among secondary school students. Our study will assess the predictive accuracy of diverse models and investigate the influence of various feature sets on their performance. By building on insights from previous studies (Namoun et al., 2020; Chui et al., 2020; Buyrukoğlu, 2022), this research seeks to contribute to the ongoing development of data-driven approaches aimed at ameliorating student success.

This research work focused on the task of collecting data connected with academic performance reflecting student grades, attendance records, and students' demographic information from secondary schools and then implementing and contrasting traditional single classifiers such as Decision Trees, Logistic Regression, and Support Vector Machines by applying advanced machine learning techniques and large amounts of data. The primary goal is to develop a precise method for identifying students at risk, enabling teachers to take necessary actions to improve their circumstances. Consequently, ensemble learning techniques and hybrid models demonstrate these capabilities.

Related Works

Accurately predicting student performance in secondary education is crucial for identifying students at risk and applying timely interventions. Recent studies in data mining techniques have showed significant accuracy of predictive models. This review explores the effectiveness of single classifiers, voting and Bagging, an ensemble learning technique in this context. Several studies have investigated the use of single and ensemble classifiers to predict secondary school student performance. In the research that done by Siddique (2021) and Jalota (2023) both found that ensemble methods, especially MultiBoost and LogitBoost, outperformed single classifiers, achieving high accuracy of 98% in predicting student performance. Olukoya

(2020) also emphasized the importance of ensemble methods. Olukoya (2020) found that REP Tree and its ensemble attained the highest accuracy, while voting ensembles performed slightly better than bagging and boosting homogeneous ensembles. Collectively, these studies indicate that ensemble methods, particularly MultiBoost and LogitBoost, are very effective in predicting secondary school student academic performance.

Adejo and Connolly (2017) demonstrated that ensemble learning using multiple data sources significantly improves prediction accuracy related to single classifiers with a single data source, aiding in the identification of at-risk students. According to Razak (2021) found that Boosted Decision Trees did better than other models, while Eleyan (2022) reported that classification trees and logistic regression were the most effective. Jalota (2023) extended these findings by employing ensemble classification techniques, obtaining accuracy of 99.8%. Additionally, Joshi (2020) demonstrates the use of data mining and machine learning for predicting the academic performance of secondary students using Naïve Bayes, Naive Bayes, Decision Tree, and Logistic Regression algorithms to predict and analyze student performance

In another study, Hasib (2022) and Singh (2020) both applied different machine learning algorithms. Hasib in his investigation discovered that Support Vector Machine (SVM) outperformed other algorithms, while Singh identified bagging as the most effective ensemble technique. Ragab (2021) focused on the impact of different factors on performance, highlighting the influence of a history of grades. These studies collectively stressed the potential of machine learning in predicting student performance, with SVM and bagging emerging as particularly effective methods. Injadat et al. (2020), Miguéis et al. (2018), Buyrukoğlu (2022), and Chui et al. (2020) investigated multiple machine learning methodologies aimed at improving the precision of forecasting student performance. One study used enrollment data and academic records to examine the effectiveness of different feature sets in predicting student performance, finding that academic, behavioral, demographic, student attendance, and family-related features were all influential (Buyrukoğlu, 2022). Another study proposed a gradient boosting machine algorithm to predict student performance at the end of the academic year. It considered factors such as age, school, neighborhood, absence, and grades, achieving accuracies of 86% and 89%. Additionally, Chui et al. (2020) systematic review highlighted the frequent use of regression and supervised machine learning models for classifying student performance. It identified student online learning activities, term assessment grades, and student academic emotions as the most significant predictors of learning outcomes. (Namoun & Alshanqiti, 2020).

Furthermore, Al-Hagery et al. (2020) demonstrated that the ability to predict student performance can assist stakeholders make good decisions and take actions that benefit higher education institutions, particularly in curriculum development and enhancing instructor effectiveness. These impacts can stem from personal, social, environmental, and psychological factors. While these studies have provided valuable insights into the factors influencing student performance, there remains a need for a more comprehensive understanding of the relative importance of different predictors and the potential of ensemble methods, such as voting, bagging, in improving prediction accuracy.

Data Mining Concept

Data mining is a complex process that involves aligning observed data with real-world phenomena (Smith, 2001). One approach to achieving this alignment is through the utilization of visualization techniques, which can be automated using fuzzy set theory to handle high-dimensional datasets effectively (Last, 1999). Another important concept in data mining is information granulation, which enhances interpretability and streamlines computational processes. In the context of investment, data mining methods such as customer clustering are employed to identify and quantify investors' perceptions and preferences (Batra, 2012). Knowledge Discovery in Databases (KDD) is a systematic process for extracting insights and patterns from large datasets. Here are the key steps involved:

1. **Data Cleaning and Preprocessing: Data cleaning and preprocessing are important steps in data analysis processes** (Kotsiantis, 2007).
2. **Data Integration:** Combining data from diverse sources into a unified dataset to facilitate comprehensive analysis and ensure all relevant information is available.
3. **Data Selection:** Data selection is a critical aspect of various fields, including language technologies, signal processing, and machine learning (Clark, 2008**).**
4. **Data Transformation:** Data transformation is a fundamental step in data preprocessing, with the potential to significantly impact the quality of visualization and user task performance (Wen, 2008).
5. **Data Mining:** Employing algorithms such as classification, clustering, association rule mining, and anomaly detection to uncover meaningful patterns, trends, and relationships within the dataset.
6. **Pattern Evaluation:** Assessing discovered patterns for significance, validity, and relevance to original research objectives using metrics like accuracy and interpretability.
7. **Knowledge Presentation:** Communicating extracted insights in formats like visualizations, reports, or summaries that are understandable and actionable for domain experts.
8. **Knowledge Utilization:** Applying discovered knowledge to inform decisions, enhance processes, or develop predictive models that drive advancements in the relevant field of study.

Classification Algorithms in Data Mining

Classification involves determining the category to which a new observation belongs based on a training set of data with known category memberships. For example, classifying an email as "spam" or "non-spam," or diagnosing a patient based on observed characteristics like gender, blood pressure, and symptoms. In machine learning terminology, classification is a type of supervised learning where a training set of correctly labeled observations is provided (Bardab, 2021).

Decision Tree

A Decision Tree is a widely used and intuitive approach in machine learning for both classification and regression tasks. It constructs decisions and their potential outcomes, forming a hierarchical structure where each internal node tests an attribute, each branch

indicates the outcome of the test, and each leaf node represents a class label or a continuous value. Kesavaraj and Sukumaran (2013) developed a model and argued that the model is capable of predicting the target variable's value based on multiple input variables. Decision trees are typically simple yet capable of managing intricate datasets, offering transparent insights into decision processes. They are adept at handling noisy data but may have limited effectiveness with large datasets.

Support Vector Machine

Support Vector Machines (SVMs) are advanced tools in machine learning that are particularly effective for classification tasks requiring distinct separation between classes. They demonstrate a complete performance across various fields yet achieving optimal results hinges on meticulous parameter tuning and a deep understanding of kernel functions. SVMs represent a modern approach in machine learning rooted in statistical learning theory (Boswell, 2002). Essentially, they aim to identify a hyperplane that perfectly divides d-dimensional data into its respective classes and have proven to be very successful in practical applications. Additionally, SVMs have been extended to tackle regression tasks, where the objective is to predict numerical values rather than simply classifying outcomes into "Yes" or "No".

Logistic Regression

Logistic regression is a powerful statistical technique used in modelling the relationship between binary response variables and explanatory factors (Dereck, 2006). The fact it is versatile makes it capable of handling both continuous and categorical explanatory variables. The primary objective of logistic regression is to develop a model that fits the data well and provides a clear interpretation of the relationships (Hosmer, 2005). Recent advancements in logistic regression, such as penalized regression methods, have been introduced to effectively handle situations with many predictor variables (Makalic, 2010).

Bayesian networks

A Bayesian Network is a graphical representation that shows the probabilistic relationships among variables and their dependencies through a Directed Acyclic Graph (DAG). Phyu (2009) described Bayesian networks as graphical models used to reason under uncertainty, with nodes representing variables (whether discrete or continuous) and arcs indicating direct connections between them.

Methods

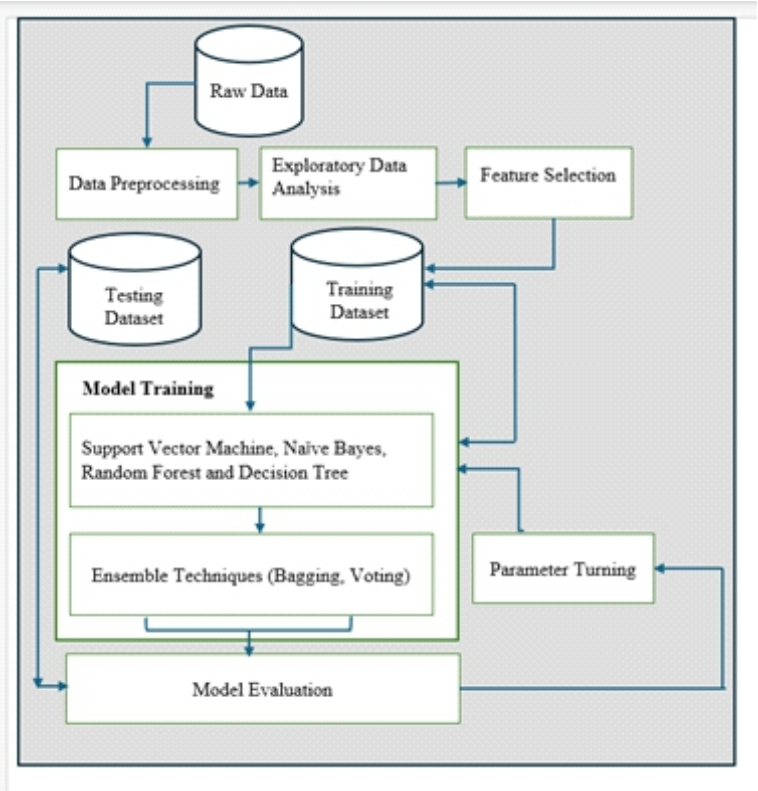

Figure 1: Proposed Model

Proposed Model Learning Activities.

The research was conducted using Python 3.11 as the programming language within an open-source Jupiter notebook. Several data analysis tools were used in this study. This includes Pandas, Seaborn, and Scikit-learn libraries. Several stages were involved in the data gathering process. This includes data preprocessing, exploratory data analysis, feature selection, model selection, model training, testing, and evaluation. The dataset for this study was sourced from two main places: the Kaggle online learning repository, which includes 1,044 instances with 31 attributes, and a local dataset obtained through a questionnaire administered in a Secondary School in Abuja, Nigeria, consisting of 45 instances and 19 attributes (Cortez et al., 2018). The attributes of the dataset are categorized into demographic, behavioral, and other relevant types.

Model Evaluation and Measurement Terms

In the field of machine learning, model performance is evaluated using a test dataset. Vijayalakshmi et al., (2019) highlights that essential evaluation metrics for classification tasks include accuracy, precision, recall, F1-Score and precision and recall (PR) curves, confusion matrix and specificity. The choice of a specific metric depends on the nature of the problem.

Table 1: Confusion Matrix

		Measured	
		Positive	Negative
Actual	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

Accuracy indicates the percentage of correct predictions out of all predictions made. Precision calculates the ratio of correctly classified instances to the total number of both correctly and incorrectly classified cases. Recall measures the ratio of correctly classified instances to the total number of both unclassified and correctly classified cases. Additionally, the F-measure combines precision and recall providing a balanced assessment of their relationship. The ROC Area, obtained by plotting the true positive rate against the false positive rate across various thresholds, serves as another important metric. Additional information is at times included in the table that includes the True Positive, True Negative, False Positive and False Negative. The terms can be explained as follows:

$$\text{Accuracy} = \frac{TP + TN}{TN + FN + FP + TP} \quad \ldots\ldots (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad \ldots\ldots (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \ldots\ldots (3)$$

$$\text{F. Measure} = \frac{\text{Precision} \times \text{Recall}_c}{\text{Precision} + \text{Recall}_c} \quad \ldots\ldots (4)$$

Results and Discussion

The following presents the findings obtained from the analysis of student performance.

Table 2: The Results of the Kaggle Online Dataset on single classifiers performance and ensemble methods.

Name of the classifier	*Accuracy*	*Precision*	*Recall*	*F-Measure*
Bagging	*0.68*	*0.62*	*0.84*	*0.71*
Voting	*0.68*	*0.64*	*0.83*	*0.72*
Logistic Regression	*0.67*	*0.64*	*0.84*	*0.72*
Naïve Bayes	*0.66*	*0.61*	*0.85*	*0.71*
Support Vector Machine	*0.64*	*0.60*	*0.92*	*0.72*
Decision tree	*0.65*	*0.62*	*0.76*	*0.69*
Random Forest	*0.67*	*0.63*	*0.89*	*0.73*

The classification results above in table 2 reveals varying strengths among the classifiers. Firstly, Bagging, Voting, and Logistic Regression models obtained 68% accuracy. SVM

demonstrates the highest recall at 92%. In terms of the F-Measure, the model achieved 73% with Random Forest. In contrast, both Naïve Bayes and Decision Tree resulted in slightly lower metrics. Finally, both Voting and Logistic Regression demonstrated the highest precision of 92% and 91% respectively. Despite their lower recall, these two classifiers are effective in applications where accurate prediction of positive instances is critical.

Table 3: The Result of the Local Dataset on Single classifiers and Ensemble methods

Name of the classifier	Accuracy	Precision	Recall	F-Measure
Bagging	*0.89*	*1.00*	*0.80*	*0.89*
Voting	*1.00*	*1.00*	*1.00*	*1.00*
Logistic Regression	*0.89*	*1.00*	*0.80*	*0.89*
Support Vector machine	*1.00*	*1.00*	*1.00*	*1.00*
Decision Tree	*1.00*	*1.00*	*1.00*	*1.00*
Random Forest	*1.00*	*1.00*	*1.00*	*1.00*
Naïve Bayes	*0.56*	*1.00*	*0.20*	*0.33*

As shown in Table 3 above, the validation process was conducted using 10-fold cross-validation. The results indicated that the proposed model, utilizing the Voting, Support Vector Machine, Decision Tree, and Random Forest classifiers, achieved perfect scores with 100% accuracy, precision, recall, and F-measure. Bagging and Logistic Regression closely followed, both achieving 89% accuracy, 100% precision, 80% recall, and an F-measure of 0.89. Naïve Bayes had the lowest performance, with 56% accuracy. This study compared individual classifiers (Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Naïve Bayes) with ensemble methods (Bagging and Voting). The results indicate that while single classifiers can achieve good performance, ensemble methods generally excel, particularly in attaining higher accuracy and maintaining balanced precision, recall, and F-measure scores.

Conclusion

The research primarily aimed to compare the results of bagging and voting classifiers with their single classifiers counterpart. Machine learning has proven to be very effective in analyzing and predicting student learning. Ensemble classifiers, including Decision Tree, Random Forest, and Support Vector Machine, were used on both a Kaggle online dataset and a local dataset acquired through questionnaires. From the local dataset evaluation, Bagging and Logistic Regression showed the highest performance with 89% Accuracy and F1-Score, 100% Precision. While Naïve Bayes had the lowest performance with 56% accuracy. Conversely, Voting and Logistic Regression achieved the highest accuracy of 68% on the Kaggle online dataset. The validation revealed that ensemble methods are effective in educational settings and warrant further investigation.

This research's findings can help pinpoint underperforming students and give more attention to them to ameliorate their learning. With this, there will be quality in terms of learning at the secondary level of education while helping the potential of higher education. The accuracy of these techniques depends on the available data and the nature of the problem. Machine

learning can play a vital role in improving the educational system by predicting and improving students' academic progress.

References

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.

Al-Hagery, M. A., Alzaid, M. A., Alharbi, T. S., & Alhanaya, M. A. (2020). Data mining methods for detecting the most significant factors affecting students' performance, *International Journal of Information Technology and Computer Science, 12*(5), 1-13.

Bardab, S. N., Ahmed, T. M., & Mohammed, T. A. A. (2021). Data mining classification algorithms: An overview, *Int J Adv Appl Sci, 8*(2), 1-5.

Batra, G., & Gupta, A. (2012). Mining the investor's perception about different investment options using clustering analysis.

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression, *Critical Care, 9,* 112 - 118.

Boswell, D., (2002). *Introduction to support vector machines*.

Buyrukoğlu, S., & Akbaş, A. (2022). Efficiency of ensemble learning algorithms in the analysis of effects of Covid-19 pandemic on electricity consumption in Turkey, *Inspiring Technologies and Innovations*, *1*(1), 9-15.

Cortez, P., & Silva, A. M. G. (2008). *Using data mining to predict secondary school student performance.*

Clark, J., Frederking, R. E., & Levin, L. S. (2008). *Toward active learning in data selection: Automatic discovery of language features during elicitation*, International Conference on Language Resources and Evaluation.

Chu, D. K., Akl, E. A., Duda, S., Solo, K., Yaacoub, S., Schünemann, H. J. & Reinap, M. (2020). Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. *The Lancet*, *395*(10242), 1973-1987.

Dierckx, G., (2006). Logistic regression model.

Eleyan, N., Al Akasheh, M., Malik, E. F., & Hujran, O. (2022). Predicting student performance using educational data mining. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (1-7). IEEE.

Hasib, A. (2022). *Atlas Fast Simulation-from classical to deep learning* (No. Atl-Soft-Slide-2022-006). ATL-COM-SOFT-2022-006.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2005). Introduction to the Logistic Regression Model.

Injadat, M., Moubayed, A., Nassif, A. B., & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining, *Knowledge-Based Systems*, *200*, 105992.

Jalota, H., Mandal, P. K., Thakur, M., & Mittal, G. (2023). A novel approach to incorporate investor's preference in fuzzy multi-objective portfolio selection problem using credibility measure, *Expert Systems with Applications, 212*, 118583.

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)* (1-7). IEEE.

Kleindorfer, D. O., Towfighi, A., Chaturvedi, S., Cockroft, K. M., Gutierrez, J., Lombardi-Hill, D., ... & Williams, L. S. (2021). 2*021 guideline for the prevention of stroke in patients with stroke and transient ischemic attack: A guideline from the American Heart Association / American Stroke Association. Stroke*.

Kotsiantis, S. B., Kanellopoulos, D. N., & Pintelas, P. E. (2007). Data preprocessing for supervised Leaning. World academy of science, engineering and technology, *International Journal of Computer, Electrical, Automation, Control and Information Engineering, 1,* 4104-4109.

Last, M., & Kandel, A. (1999). *Automated perceptions in data mining. FUZZ-IEEE'99. 1999 IEEE international fuzzy systems*. Conference Proceedings (Cat. No.99CH36315), 1, 190-197 1.

Ling, L. (2011). A review of classification algorithms in data mining, *Journal of Chongqing Normal University.*

Makalic, E., & Schmidt, D. F. (2010). *Review of modern logistic regression methods with application to small and medium sample size problems*. Australasian Conference on Artificial Intelligence.

Mining, K. D. T. D. (1996). *What is knowledge discovery*, Tandem Computers Inc, 253.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review, *Applied Sciences, 11*(1), 237.

Nguyen, L. H., Drew, D. A., Graham, M. S., Joshi, A. D., Guo, C. G., Ma, W., & Zhang, F. (2020). Risk of COVID-19 among front-line health-care workers and the general community: A prospective cohort study. *The Lancet Public Health*, *5*(9), e475-e483.

Olukoya, B. M. (2020). Comparison of feature selection techniques for predicting student€™ s academic performance. *International Journal of Research and Scientific Innovation*, *7*(8), 97-101.

Phyu, T. N., (2009). *Survey of classification techniques in data mining. In the international multi conference of engineers and computer scientists*, Hong Kong, China, 1, 1-5.

Razak, F. A., & Zamzuri, Z. H. (2021). Modelling heterogeneity and super spreaders of the COVID-19 spread through Malaysian networks, *Symmetry*, *13*(10), 1954.

Saddique, R., Zeng, W., Zhao, P., & Awan, A. (2023). Understanding multidimensional poverty in pakistan: implications for regional and demographic-specific policies, *Environmental Science and Pollution Research*, 1-16.

Smith, M. H., & Pedrycz, W. (2001). Perception issues in data mining. 2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236), 4, 2553 vol.4-.

Singh, R. P., & Chauhan, A. (2020). Impact of lockdown on air quality in India during COVID-19 pandemic, *Air Quality, Atmosphere & Health*, *13*, 921-928.

Wen, Z., & Zhou, M. X. (2008). Evaluating the use of data transformation for information visualization. *IEEE Transactions on Visualization and Computer Graphics, 14.*

Wright, P. M., & Boswell, W. R. (2002). Desegregating HRM: A review and synthesis of micro and macro human resource management research, *Journal of Management*, *28*(3), 247-276.

Williams, L., McGraw, G., & Migues, S. (2018). Engineering security vulnerability prevention, detection, and response, *IEEE Software*, *35*(5), 76-80.

Vijayalakshmi et al. (2019). *Deep neural network for multi-class prediction of student performance in educational data.*