MACHINE LEARNING AND BIG DATA-DRIVEN DECISION MAKING TO IMPROVE HIGHER EDUCATION INSTITUTION PERFORMANCE FROM THE PERSPECTIVE OF INDUSTRY 5.0

Dr. Tryambak Hiwarkar
Professor, CSE, Sardar Patel University, Balaghat, MP

Sofiya Tabassum
M. Tech Final Year Student, Department of CSE, Sardar Patel University, Balaghat, MP

Abstract
The performance evaluation of the students and their performance improvement are the big concerns for the higher education institution and need to be addressed. If the performance evaluation of the students has been done effectively, a decision can be made for further improvement. For this, decision-making based on the data can be proposed. However, big data and analytics for educational applications are just taking their initial steps and will take some time to grow, but they should not be disregarded. Motivated by the same, the present approach aims to incorporate and test the capabilities of big data analytics for data-driven decision-making in order to improve the performance of higher education institutions. The dataset has been taken to work with the performance of the students in particular exams, and machine learning models such as the random forest classifier, the XGBoost classifier, and the AdaBoost classifier have been simulated on it. for scaling the students and classifying them in various categories according to their performance so that necessary action can be taken before their final evaluation. The results show that the Random Forest Classifier outperformed the AdaBoost Classifier and XGBoost Classifier in decision making, with an accuracy of 90.3413%.

Keyword: Big data-driven decision making, Machine Learning, Performance improvement, Higher education institutions, AdaBoost Classifier and XGBoost Classifier

1. Introduction

A society known as "Industry 5.0" includes both people and automated and artificially intelligent tools. By using cutting-edge tools like big data and Internet-associated objects, automation systems are capable of helping people work more quickly and efficiently. This gives the effectiveness and productivity pillars of Industry 4.0 a more individual touch. Industry 5.0 basically aims to combine human intelligence and ingenuity with the cognitive and computational capacity of office equipment in collaboration as it becomes smarter and much more networked.

For this, "big data" is a technique to gather information from unorganized databases and analyze it methodically if we need to work with enormous or sophisticated data sets using some

preset data-processing application. "Big data" is typically used to refer to data volume, but this is not always the case in practice. On December 1, 2015, Michael Rada used the social media platform LinkedIn to make the first use of the phrase "industry 5.0"[1]. The technologies known as Industry 5.0 are described in the essay by Sc. P. O. Skobelev and P. S. Yu. Borovik [2]. The majority of definitions of big data focus on the data volume in memory, although data volume is not the only characteristic that can be used to define anything precisely. Structured data, unstructured data, and semi-structured data are the different types of datasets [3]. In the education industry, there has been a trend of investigation into big data and sustainability given the transformative significance of big data analytics. However, the majority of this research includes both theoretical and personal examples. There are few empirical studies examining the capability of big data and predictive analytics (BDPA) and how it affects the educational sector for students' performance. But the present study has been focused on testing the capabilities of big data analytics for data-driven decision-making to improve higher education institution performance from the perspective of industry 5.0. Previously, no research work has been focused on this. Machine learning has emerged as a means of developing a solution to the current problem.

2. Literature review

Decision making, based on the data has gained popularity around in 1980s to 1990s which further has growing into the far more complex idea of big data including software techniques typically referred to as analytics. However, big data and analytics for educational applications are just taking their initial steps and will take some time to grow but should not be disregarded. Big data and analytics can be incorporated into administrative and educational responsibilities, even though they are not a silver bullet for all the problems and choices higher education managers confront. Previously, various author has shown their work in it. For example, the author in [4] offered an approach that involved a dual-stage analysis using partial least squares and deep learning, an emerging kind of artificial intelligence models. The designed model can be predicted with an accuracy of 83% using a deep ANN architecture. The results of data-driven choice making from the interaction between big data analytic capability and data-driven decision making towards the performance of HEIs also have discover. Results showed that the association between big data analytic capabilities and the performance of HEIs might positively play a vital role in data-driven decision making. Author of [5] investigated the relationships between the utilization of big data savvy teams' skills, big data-driven activities, and business success, drawing on the resource-based view of the firm and data gathered from big data professionals working in international agrifood networks. The findings from structural equation modeling suggest that the major determinants of big data-driven activities, which ultimately affect business performance, are the abilities of big data-driven activities teams to provide insightful data. According to [6] data-driven decision making has gained popularity in the 1980s and 1990s, is growing into the far more complex idea of big data, which depends on software techniques typically referred to as analytics. Although big data and analytics for educational applications are in their early stages and will take some time to grow, they are being felt and should not be disregarded. Hence, big data and analytics can be incorporated into administrative and educational responsibilities, even though they are not a silver bullet for all the problems and choices higher education managers confront Author [7] considers that attempting to properly define the bounds of data science is not crucial. In order for data science

to have an effect on business, it is vital to comprehend its linkages to other crucial related topics and (ii) start to establish the basic principles that underlie data science. Author argued on the field's boundaries in an academic context. Authors are able to more clearly comprehend and articulate the benefits of data science once he accepts (ii). Furthermore, authors haven't felt comfortable calling it data science until he embraces (ii). In this essay, we offer a viewpoint that covers all of these ideas. Author concludes by providing a brief list of the underlying core concepts of data science as examples. The goal of the study in [8] is to determine the numerous PMMs used to assess the BDDSC. It is based on a thorough examination of 66 studies. According to the results, there are two groups of PMMs that apply to BDDSC but are not mutually exclusive. The study in [9] set out to empirically investigate the use of artificial intelligence in decision-making processes, organizational sustainability, and automated manufacturing systems in big data-driven smart urban economies. Author has conducted analyses and created projections on the Internet of Things based on data collected in cyber-physical scheme manufacturing and gathered from Algorithmia, Capgemini, Deloitte, Management Events, and PwC. Author of [10] has gives idea, concept and application of big data analytical. According to him data science is referred to in that context as a body of basic ideas that support the learning of information and understanding from data. The methods and tools employed aid in the analysis of crucial data, assisting companies in better understanding their environments and making timely, informed decisions. Further author in [11], author essay advances the theoretical knowledge of the function that big data plays in solving the issues that higher education institutions are currently facing. In order to better understand the sophistication of impacts on student-related outcomes, teaching, and the "what if questions" for study experimentation, the paper draws on emerging literature in big data and describes ways to better incorporate the growing data available from different sources within institutions of higher education. Author in [12] emphasized the ways in which assessments are made in higher education and being put to the test by technological advancements, which include storage for massive volumes of data. According to author most, if not all, stakeholders want more data because they believe it will help them make better decisions. The article'[13] had a goal to investigate big data and learning analytics in blended educational contexts. Author in [14] discussed that big data analytics (BDA) ensures that data may be examined, classified, and changed into knowledge that is beneficial for organizations as well as knowledge connected to big data and effective decision-making processes, enhancing performance. Author of [15] felt that higher education institutions are faced with the issue of continuing to produce high-quality teaching, consulting, and research using virtual learning environments in the wake of the SARS-CoV-2 pandemic. As a result of the massive amount of data being produced in this situation, technologies like big data analytics are required to open up new prospects for innovation activities.

According to author of [16] big data's effects are starting to be felt, especially in the higher education industry. Higher academic quality and improved student and staff experiences would result from the strategic use of big data and its uses. This study examines and describes various instances of big data analytics in UK higher education institutions and uses the output from JISC's BI projects. Author of [17] felt that higher education institutions are working in a more difficult and competitive environment. This essay discusses current issues that higher education

institutions around the world are facing and considers how big data might help. The article's different sections had discussed several potential and practical difficulties related to the use of big data in higher education. Author in [18] research on operations and industrial management and focused heavily on the value of big data and predictive analytics. Big data and predictive analytics have been reported to enhance supply chains and organizational effectiveness, but little has been written about the influence of outside institutional influences on an organization's resources to develop big data capabilities.

According to author of [19], today's learning management systems (LMS) are widely employed in the education sector and have advanced. It produced a vast and varied amount of data from the pupils' online content usage behaviors. It produces a lot of useless and discarded data, and standard learning analytics are unable to process and assist with these problems. Author of [20], emphasized on big data's effects are starting to be felt, especially in the higher education industry. Higher educational quality and improved student and staff experiences would result from the strategic use of big data and its applications. This paper examines and summarizes eleven instances of big data analytics at UK higher education institutions using the results of JISC's business intelligence projects.

However, big data analytics has been applied in various industrial applications for data-driven decision making, but it has not been reported in the educational industry for higher education institutions' performance assessments. Hence, the proposed work has been conducted to improve higher education institution performance from the perspective of Industry 5.0 based on machine learning and big data-driven decisions.

The major contribution of the present research work is

1. The present research work has been implemented the big data analytics techniques for data-driven decision- making to improve higher education institution performance from the perspective of Industry 5.0.
2. Different machine learning models has been simulated on the dataset for the appropriate outcome in term of effective decision making.

3. Material and methodology

For the proposed research work the material and methodology used has been described in the following sections.

3.1 Material

The proposed approach aims to develop a data driven decision making in education sector for performance assessment. The publicly available benchmark dataset for weather has been gathered to verify the suggested methodology. Machine learning (ML) methods have been used to simulate this dataset in order to identify the actual environmental conditions. The dataset has been collected from https://www.kaggle.com/, which has been publically available.

The present dataset has consisted of 16 variables and 393537 values for each variable. The dataset variables are code_module, code_presentation, id_student, gender, region, highest_education, imd_band, age_band, num_of_prev_attempts ,credits, disability,final_result assessment_type, score, activity_type_visits and date_visits.

The dataset variable's visualization has been given below.

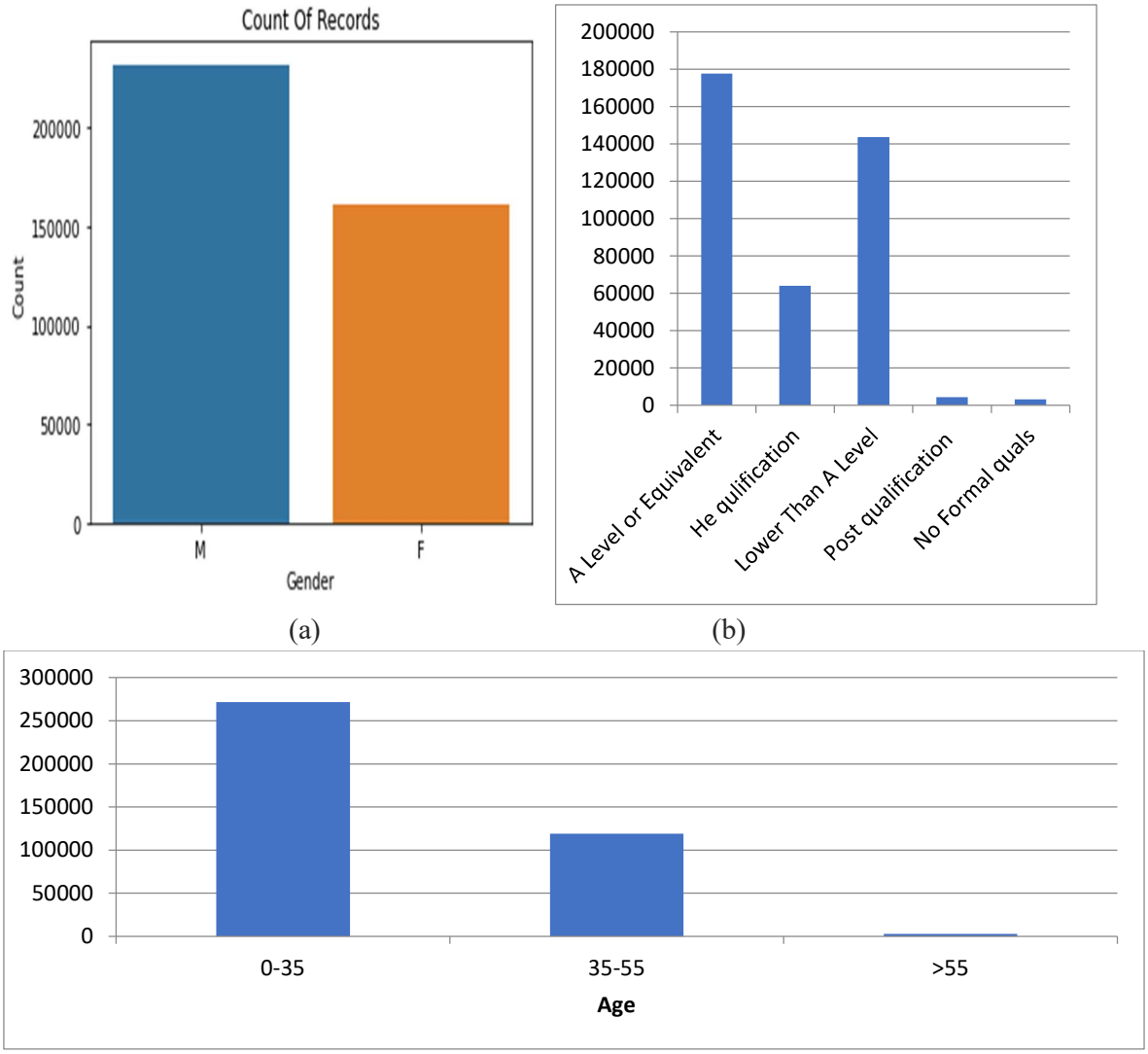

Figure 1: Dataset variable's visualization (a) Gender count in dataset (b) Qualification level of the students (c) Age visualization of student

3.2 Dataset splitting

The ML model, as is well known, is a learning-based approach on a given dataset. Hence, for the efficient learning of the same, the given dataset needed to be split into two sets. The first set is a training set, and the second set is a testing set. A 2 split of dataset often involves testing or statistical analysis of the data in one part and training the algorithm in the other. Data splitting is a crucial component in data science, especially when building models from data For the present research work, a 4:1 ratio has been employed for splitting the dataset into training and testing sets. That means 80% of total data entries have been employed in the training set for improving the learning of the model, and the remaining 20% of data entries have been employed in the testing set for evaluating the learning of the model As a result, the training set consists of 299088entries (row) for 16 different parameters from the dataset, while the testing

set consists of the remaining 74772 entries (row).

3.3 Results and analysis

The results of proposed methodology have been divided into two parts for both of the machine leaning models known as training results and testing results. For machine learning models the detailed results in term of training results and testing results has been presented below.

3.3.1 Training results of AdaBoostClassifier

The AdaBoostClassifier has been tested for its performance based on accuracy, precision, recall, and F1 score. The training results of AdaBoostClassifier have been presented in Table 1.

Table 1: AdaBoost training results

Parameters	Value
Accuracy	64.66%
Precision	0.6466
Recall	0.6466
F1 score	0.6466

It has been analyzed from the table 1 that, the AdaBoostClassifier has been achived an accuracy of 64.66% when used on the given dataset for performance assessment of the students. Hence it can be conclude that the performance of AdaBoostClassifier can be consider as moderate on the given dataset. The confusion matrix for the same has been given below in figure 2

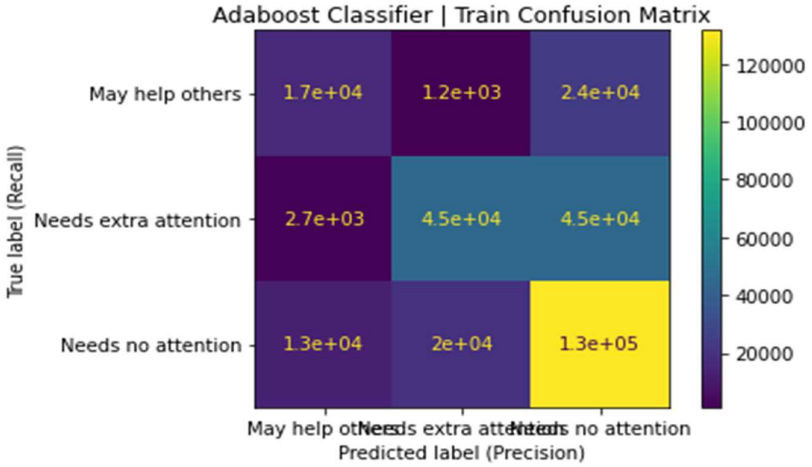

Figure 2: Confusion matrix for Adaboost at training set

From the confusion matrix, it can be analyzed that the model has been making a decision in regards to the students performance within three category of decision including may help other for active students, need extra attention for week students and no need attention for moderate students.

3.3.2 Testing results of AdaBoostClassifier

In this model has been tested for 16 variable input for its 74772 values. The results for testing set with Adaboost have been presented below in table 2.

Table 2: Adaboost testing set results

Parameters	Value

Accuracy	64.60%
Precision	0.6460
Recall	0.6460
F1 score	0.6460

It has been analyzed from the table 2 that, the AdaBoostClassifier has been achived an accuracy of 64.60% when used on the given dataset for performance assessment of the students. Hence it can be conclude that the performance of AdaBoostClassifier can be consider as moderate on the given dataset.The confusion matrix for the same has been given below in figure 3

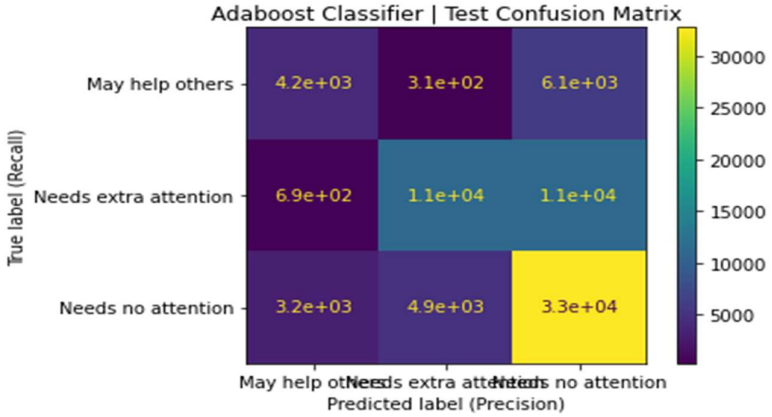

Figure 3: Confusion matrix for Adaboost at testing set

3.3.3 Training Results of XGBoostClassifier

The XGboostClassifier has been tested for its performance based on accuracy, precision, recall, and F1 score. The training results of XGboostClassifier have been presented in Table 3.

Table 3: XGBoost training results

Parameters	Value
Accuracy	67.09%
Precision	0. 6709
Recall	0. 6709
F1 score	0. 6709

It has been analyzed from the table 3 that, the XGClassifier has been achieved an accuracy of 67.09% when used on the given dataset for performance assessment of the students. Hence it can be conclude that the performance of XGClassifier can be consider as best on the given dataset. The confusion matrix for the same has been given below in figure 4.

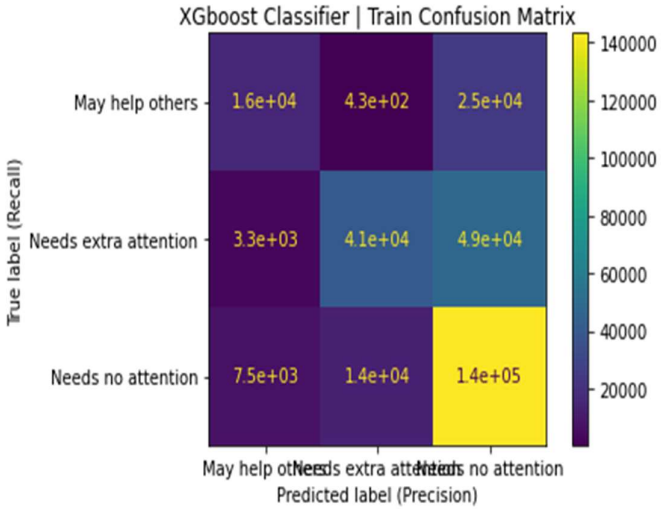

Figure 4: Confusion matrix for XGboost at training set

From the confusion matrix, it can be analyzed that the model has been making a decision in regards to the students performance within three category of decision including may help other for active students, need extra attention for week students and no need attention for moderate students.

3.3.4 Testing results of XGBoostClassifier

In this model has been tested for 16 variable input for it's 74772 values. The results for testing set with XGboost has been presented below in table 5.

Table 5: XGboost testing set results

Parameters	Value
Accuracy	64.84%
Precision	0.6484
Recall	0.6484
F1 score	0.6484

It has been analyzed from the table 4.2 that, the AdaBoostClassifier has been achived an accuracy of 64.84% when used on the given dataset for performance assessment of the students. Hence it can be concluding that the performance of XGBoostClassifier can be consider as moderate on the given dataset.

The confusion matrix for the same has been given below in figure 6

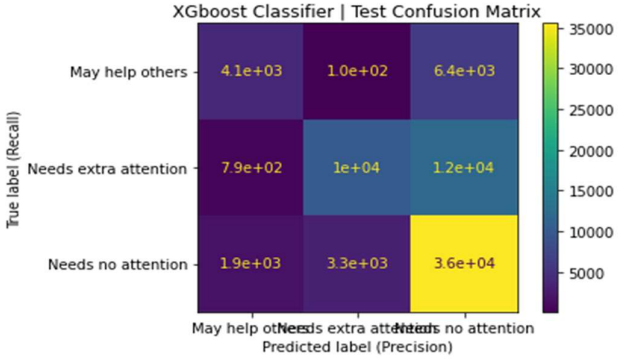

Figure 6: Confusion matrix for XGboost at testing set

.

3.3.5 Training results of RandomforestClassifier

The RandomforestClassifier has been tested for its performance based on accuracy, precision, recall, and F1 score. The training results of RandomforestClassifier have been presented in Table 6.

Table 6: Randomforest training results

Parameters	Value
Accuracy	90.34%
Precision	0.9034
Recall	0. 9034
F1 score	0. 9034

It has been analyzed from the table 6 that, the RandomforestClassifier has been achieved an accuracy of 90.34% when used on the given dataset for performance assessment of the students. Hence it can be conclude that the performance of RandomforestClassifier can be consider as moderate on the given dataset. The confusion matrix for the same has been given below in figure 7.

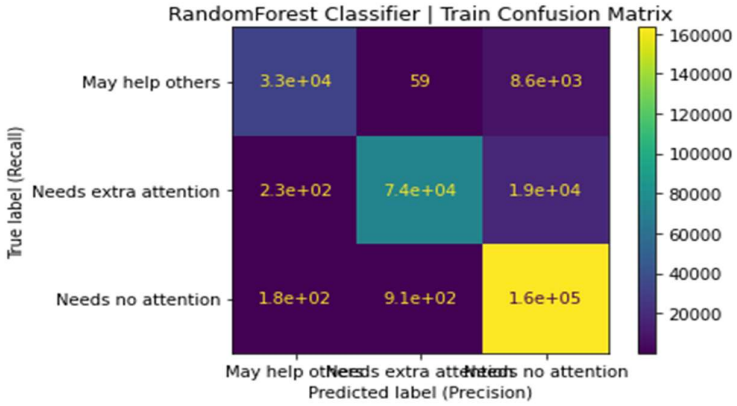

Figure 7: Confusion matrix for Randomforest at training set

From the confusion matrix, it can be analyzed that the model has been making a decision in regards to the students performance within three category of decision including may help other for active students, need extra attention for week students and no need attention for moderate students.

3.3.6 Testing results of RandomforestClassifier

In this model has been tested for 16 variable input for it's 74772 values. The results for testing set with Randomforest has been presented below in table 7.

Table 7: Randomforest testing set results

Parameters	Value
Accuracy	84.18%
Precision	0. 8418
Recall	0. 8418

F1 score	0. 8418

It has been analyzed from the table 4.6 that, the RandomforestClassifier has been achived an accuracy of 84.18% when used on the given dataset for performance assessment of the students. Hence it can be conclude that the performance of RandomforestClassifier can be consider as moderate on the given dataset. The confusion matrix for the same has been given below in figure 8

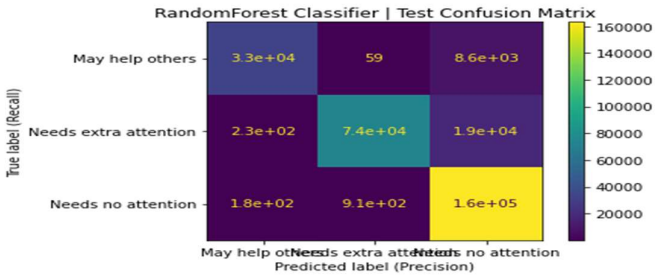

Figure 8: Confusion matrix for Randomforest at testing set

It has been analyzed that the Randomforest classifier outperformed the other two classifiers.

3.4 Conclusion

The current work, aims to incorporate and evaluate big data analytics' capabilities for data-driven decision making in order to enhance the performance of higher education institutions, with particular emphasis on Industry 5.0. The dataset was used for the study, which is publically available, to examine how well students performed in specific exams, and machine learning models like the random forest classifier, XGBoost classifier, and AdaBoost classifier were simulate on it . This allowed for the necessary action to be taken before the students' final evaluation. The findings indicate that the Random Forest Classifier, with an accuracy of 90.3413%, outperformed the AdaBoost Classifier and the XGBoost Classifier in decision-making.

References

1. Goicoechea, A., 2007. Architectures and Digital Administration. *Planning, Design, and Assessment*.
2. Skobelev, P.O. and Borovik, S.Y., 2017. On the way from Industry 4.0 to Industry 5.0: From digital manufacturing to digital society. *Industry 4.0*, *2*(6), pp.307-311.
3. Bay Atlantic University June 1 and University, B.A. (2022) *Characteristics of big data: Types, & examples*, *Bay Atlantic University - Washington, D.C.* Available at: https://bau.edu/blog/characteristics-of-big-data/ (Accessed: December 14, 2022).
4. Ashaari, M.A., Singh, K.S.D., Abbasi, G.A., Amran, A. and Liebana-Cabanillas, F.J., 2021. Big data analytics capability for improved performance of higher education institutions in the Era of IR 4.0: A multi-analytical SEM & ANN perspective. *Technological Forecasting and Social Change*, *173*, p.121119.
5. Akhtar, P., Frynas, J.G., Mellahi, K. and Ullah, S., 2019. Big data-savvy teams' skills, big data-driven actions and business performance. *British Journal of Management*, *30*(2), pp.252-271.
6. Picciano, A.G., 2012. The evolution of big data and learning analytics in American higher education. *Journal of asynchronous learning networks*, *16*(3), pp.9-20.

7. Provost, F. and Fawcett, T., 2013. Data science and its relationship to big data and data-driven decision making. *Big data*, *1*(1), pp.51-59.
8. Kamble, S.S. and Gunasekaran, A., 2020. Big data-driven supply chain performance measurement system: a review and framework for implementation. *International Journal of Production Research*, *58*(1), pp.65-86.
9. Cunningham, E., 2021. Artificial intelligence-based decision-making algorithms, sustainable organizational performance, and automated production systems in big data-driven smart urban economy. *Journal of Self-Governance and Management Economics*, *9*(1), pp.31-41.
10. Vassakis, K., Petrakis, E. and Kopanakis, I., 2018. Big data analytics: applications, prospects and challenges. *Mobile big data*, pp.3-20.
11. Daniel, B. and Butson, R., 2014, September. Foundations of big data and analytics in higher education. In *International conference on analytics driven solutions: ICAS2014* (pp. 39-47).
12. Webber, K.L. and Zheng, H., 2020. Data analytics and the imperatives for data-informed decision-making in higher education. *Big data on campus: Data analytics and decision making in higher education (Part 1, 1)*.
13. Picciano, A.G., 2014. Big data and learning analytics in blended learning environments: Benefits and concerns. *IJIMAI*, *2*(7), pp.35-43.
14. Ferraris, A., Mazzoleni, A., Devalle, A. and Couturier, J., 2018. Big data analytics capabilities and knowledge management: impact on firm performance. *Management Decision*.
15. Marchena Sekli, G.F. and De La Vega, I., 2021. Adoption of big data Analytics and its impact on organizational performance in higher education mediated by knowledge management. *Journal of Open Innovation: Technology, Market, and Complexity*, *7*(4), p.221.
16. Ong, V.K., 2015, July. Big data and its research implications for higher education: Cases from UK higher education institutions. In *2015 IIAI 4th International Congress on Advanced Applied Informatics* (pp. 487-491). IEEE.
17. Daniel, B., 2015. B ig D ata and analytics in higher education: Opportunities and challenges. *British journal of educational technology*, *46*(5), pp.904-920.
18. Dubey, R., Gunasekaran, A., Childe, S.J., Blome, C. and Papadopoulos, T., 2019. Big data and predictive analytics and manufacturing performance: integrating institutional theory, resource-based view and big data culture. *British Journal of Management*, *30*(2), pp.341-361.
19. Adam, K., Bakar, N.A.A., Fakhreldin, M.A.I. and Majid, M.A., 2018. Big data and learning analytics: a big potential to improve e-learning. *Advanced Science Letters*, *24*(10), pp.7838-7843.
20. Ong, V.K., 2016. Business intelligence and big data analytics for higher education: Cases from UK higher education institutions. *Information Engineering Express*, *2*(1), pp.65-75.
21. Saini, A. (2022) *AdaBoost algorithm - A complete guide for beginners*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/ (Accessed: December 15, 2022).

22. guest_blog (2020) *XGBoost algorithm: XGBoost in machine learning*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/ (Accessed: December 15, 2022).