CatBoost_Data_Mining_Light_Gradient_Boosting_Machine_XGBoost.pdf

by Student 3

Submission date: 28-Apr-2024 08:58PM (UTC+0700)
Submission ID: 2225858084
File name: CatBoost_Data_Mining_Light_Gradient_Boosting_Machine_XGBoost.pdf (807.6K)
Word count: 6076
Character count: 35077

Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 10, No. 1, March 2024, pp. 154-165
ISSN: 2338-3070, DOI: 10.26555/jiteki.v10i1.28352

Students Final Academic Score Prediction Using Boosting Regression Algorithms

Dignifo Nauval Muhammady, Haidar Aldy Eka Nugraha, Vinna Rahmayanti Setyaning Nastiti, Christian Sri Kusuma Aditya
Universitas Muhammadiyah Malang, Tegalgondo, Malang, Jawa Timur, 65144, Indonesia

ARTICLE INFO

Article history:

Received February 01, 2024
Revised March 14, 2024
Published March 27, 2024

Keywords:

Academic;
CatBoost;
Data Mining;
Light Gradient Boosting Machine;
Machine Learning;
XGBoost Regressor

ABSTRACT

Academic grades are crucial in education because they assist students in acquiring the knowledge and skills necessary to succeed in school and their future. Accurately predicting students' final academic performance grade score is important for educational decision-makers. However, creating precise prediction models based on students' historical data can be challenging due to the complex nature of academic data. This research analyzes student academic data totaling 649 Portuguese language course student data that has been processed according to data requirements which are then predicted using XGBoost Regressor, Light Gradient Boosting Machine (LGBM), and CatBoost. This research aims to develop a robust prediction model that can effectively predict students' final academic performance. This research offers valuable insights into the factors that influence academic success and provides practical implications for educational institutions looking to improve their decision-making processes. The prediction requires identifying key predictors of academic performance, such as previous grades, attendance records, and socio-economic background. The research makes a contribution by improving the matrix MAE in this research is less than the previous research from 2.2 average each algorithm to 0.22 average, this less MAE means the better model. The research achieved MAE score of 0.22 average. In conclusion, this research is expected to address the challenge of predicting student academic performance through the application of advanced machine learning techniques. The results provide valuable insights for decision-makers in education and highlight the importance of a data-driven approach to improving academic performance. By utilizing machine learning algorithms, educational institutions can effectively support student learning and success.

Corresponding Author:

Vinna Rahmayanti Setyaning Nastiti, Universitas Muhammadiyah Malang, Tegalgondo, Malang, Jawa Timur, 65144, Indonesia
Email: vinastiti@umm.ac.id

1. INTRODUCTION

Educational institutes play a pivotal role in shaping a student's life in various aspects [1], [2]. Education is expected to provide benefits and contribute positively to the lives of individuals and societies as a whole [3]. Education is the development of individual potential to enhance the knowledge and understanding of students. One important aspect of education is academic evaluation [4], [5]. Academic achievement serves as a critical indicator in the field of education, reflecting student abilities and success in achieving learning goals [6], [7]. Academic success is influenced by various circumstances, and individuals' learning abilities vary based on their backgrounds [8].

In an effort to identify factors that influence academic achievement, machine learning approaches have become a highly effective tool [9], [10]. This academic achievement involves the use of machine learning

algorithms and statistical techniques to assist users in interpreting student learning habits, academic performance, and identifying areas for improvement if necessary [11]. Therefore, a new discipline called educational data mining (EDM) has emerged. EDM utilizes data collected from educational environments to reveal valuable and relevant knowledge [12]. Educational Data Mining (EDM) uses data mining (DM) methods to gain a better understanding of student behavior and learning environments [13]. DM enables education practitioners to identify the key factors that affect student performance. This allows for quick action to be taken to assist students, improve educational quality, and optimize school resource management [14], [15].

Ensemble learning is a powerful statistical modeling approach that combines multiple base models, also known as base learners, to predict a single value [16]. This method leverages the collective intelligence of diverse models, harnessing their individual strengths to achieve superior predictive accuracy and robustness [17]. Ensemble techniques have gained widespread popularity in the realm of machine learning, particularly in regression analysis [18]. Regression analysis algorithms are some of the most popular ensemble learning used in machine learning models [19]. Regression analysis algorithms are fundamental components of ensemble learning frameworks [20]. These are highly effective in capturing intricate relationships between input features and target variables, making them suitable for various prediction tasks [21]. Among the plethora of regression algorithms available, XGBoost, LGBM, and Catboost [22]-[24] are short form for extreme boost gradient regression. Compared to traditional machine learning algorithms, XGBoost, LGBM, and CatBoost offer several advantages, including faster training times, superior predictive accuracy, and enhanced interpretability [25], [26]. They can also handle missing values, categorical features, and noisy data, making them more appealing in practical settings [27]. As a result, these algorithms have become popular choices for data scientists and practitioners who seek cutting-edge solutions to regression problems in various domains [28]. In summary, ensemble learning, specifically through the use of XGBoost, LGBM, and CatBoost algorithms, is a powerful approach to regression analysis in machine learning [29]. These algorithms utilize the collective knowledge of multiple base learners to provide exceptional predictive performance, making them essential tools in modern data-driven applications [30].

The previous researches have extensively explored the efficacy of ensemble learning techniques including regressions techniques [31]. Studies have investigated the benefits of comparing multiple base models to improve predictive score and robustness in various domains. For example, Rangga M. research Final Grade Prediction Model Based On Student's Alcohol Consumption [32] was developed to predict students' academic performance using same Portuguese dataset and various regression such as SVR and Random Forest Regressor with the MAE matrix evaluation [33], [34], with the result MAE of 2,25 on Random Forest Regressor, MAE of 2,24 on SVR. According to this Rangga M. research result, this research has significantly improved the MAE, achieving the lowest score of 0,22.

In addition, a few researchers only take a look for classic regressions and classifiers on predicting student performance. For example, research conducted by Abdelrahman A (2022) [35] aims to predict good/bad the final exam grade. The proposed model was developed by Abdelrahman A, who evaluated a range of popular classification and regression algorithms using data from a data structures and algorithms course (CS2) offered at a large public research university. The regression analysis utilized Random Forest Regression and Multiple Linear Regression [36], [37]. For classification tasks, the following algorithms are considered: Logistic Regression, Decision Tree, Random Forest Classifier, K Nearest Neighbors, and Support Vector Machine [38]. After conducting experiments, he identified the algorithm that demonstrated the most promising performance across both classification and regression tasks. The finding research result in regression analysis using the R2 matrix evaluation, the classic Random Forest Regression model is the top performer with an R2 score of 0.977.

In the research conducted by Singh R (2020) [39] conducted research discusses the analysis of educational data and evaluation of student performance using five different machine learning techniques: Passive Aggressive Classifier (PAC), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Radius Neighbour Classifier (RNC), and Extra Tree (ET). The results obtained from various machine learning algorithms are thoroughly discussed. The Support Vector Machine (SVM) algorithm achieved the highest accuracy, as verified by evaluating various metrics such as sensitivity, specificity, and precision. These findings have practical implications as they can be applied to assess the performance of incoming students and identify those who may require additional support. Identifying underperforming students enables higher educational institutions to allocate resources and interventions effectively, thereby improving student performance [40]. The primary objective of this research is to develop an efficient framework that significantly enhances the accuracy of student performance prediction. Machine learning techniques serve as the cornerstone of this endeavor, offering valuable insights that can inform successful decision-making processes aimed at enhancing student performance [41]. The paper describes the use of various machine

learning techniques, such as PCA, SVM, LDA, RNC, and ET, to evaluate student performance. SVM is found to be the most effective technique with an accuracy of 94.86%, followed closely by LDA with an accuracy of 93.21%.

Although some studies above have made valuable contributions related to the prediction and classification of the Portuguese student dataset, there is a significant research gap. These previous studies only considered only a few prediction regressions, especially only developed with few classic regressions, without adding such modern regressions like gradient boosting regressions [32], [35]. Therefore, this study aims to address this gap by utilizing 3 gradient boosting methods such as XGBoost Regressor, LGBM, and CatBoost Regressor for predicting students' final academic scores (G3) [42]. These 3 regressions have proven to handle categorical numeric target class, according to Injadat (2020) [12] stated that categorial numeric variable can be handle by ensemble learning algorithm such as gradient boosting algorithms. This study's contribution significantly improved the prediction results by reducing the matrix MAE score from 0,22. Furthermore, this study analyze which variables that is the most influential to this prediction on various boosting regression algorithms, providing valuable insights with their effectiveness [43]. By using these gradient boosting in the prediction of final academic scores, this approach not only improves prediction accuracy and robustness on regression predictive, but also results in a more efficient model for handling categorical data. This paper advances our understanding of predictive modeling in educational evaluation and provides practical implications for educational decision-makers seeking to improve student outcomes.

2. METHODS

This research specifically focuses on the application of boosting regression algorithms such as XGBoost, LightGBM, and CatBoost, which are known for their effectiveness in handling complex datasets and delivering high predictive R2 score [22], [25], [27]. The research flows shown on Fig. 1, including data preprocessing such feature encoding and scaling, data split into train and test, Modeling, each step will be explained in the next section to provide an understanding of the research process.

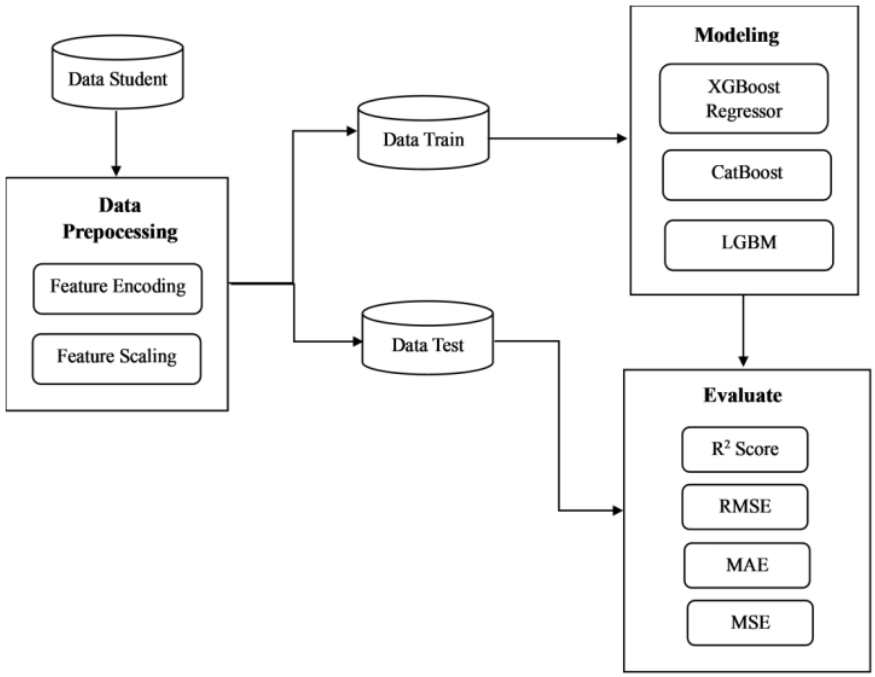

Fig. 1. Regression flowchart

2.1. Data Collection

The dataset is taken from website https://www.kaggle.com/datasets/uciml/student-alcohol-consumption. A Portuguese language course in high school students. It consists of 649 data with 33 attributes. Where all

the data and attributes are used with G3 (final grade) as the target output [44]. The Important attributes used can be seen in Table 1.

Table 1. Students important attributes

Atribute	Description
Absences	number of school absences (numeric: from 0 to 93)
Age	student's age (numeric: from 15 to 22)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
Freetime	free time after school (numeric: from 1 - very low to 5 - very high)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
Health	current health status (numeric: from 1 - very bad to 5 - very good)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
School	student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho

2.2. Pre-Processing Data

In this research apply several preprocessing steps to the data to prepare it before using it in model development. Firstly, encoding categorical variables such as gender into binary variables, where the value 0 represents male and the value 1 represents female. This allows machine learning algorithms to handle categorical variables more efficiently [12]. Next section, the numerical data is normalized to ensure a consistent scale. This is critical for machine learning algorithms to work effectively and not be affected by the different scales of each variable. In the end splitting the data into training and test sets in appropriate proportions, such as 80% for the training set and 20% for the test set. This data splitting is done randomly to ensure balanced representation in both sets and allows for an objective evaluation of the performance of our predictive model. These preprocessing steps are essential to ensure that the data used in our predictive model development is optimally prepared.

2.3. XGBoost Regressor

Fig. 2 shows Extreme Gradient Boosting (XGBoost), an algorithm created by Chen and Guestrin in 2016. XGBoost is a powerful tool among data science researchers due to its effectiveness as a tree-based ensemble learning algorithm [22]. XGBoost is a regression tree that enables analysts to measure the impact of covariates on target variables during each iteration of the boosting process. The decision rules of XGBoost are the same as those of the classic decision tree [29]. The regressive tree uses nodes to represent attribute testing values, with the leaf node and its score representing the final result. The final result is the sum of all scores predicted by the tree K, as demonstrated.

$$y = \sum_{k=i}^{K} f_k(U_i), f_k \epsilon F \quad (1)$$

$$\overline{y_i} = y_i^0 + n \sum_{k=i}^{n} f_k(U_i) \quad (2)$$

Where, y_i shows the predicted output for iht data with a parameter vector U_i; n marks the number of estimators corresponding to the independent tree structure for each f_k (which is k 1 to n); and y_i shows a primary hypothesis, which is actually the average of the original parameters in the training data. η represents the level of learning associated with improving the performance of the model. XGBoost has the ability to enhance the tree boosting approach to process almost any type of data quickly and accurately. This model is built using scikit-learn XGBoost compatibility.

2.4. LGBM

LightGBM (Light Gradient Boosting) is a widely used gradient boosting technique in the ensemble method. It utilizes decision trees. Meanwhile, the formula of the LGBM regressor on Fig. 3. Where $F_t(x)$ is a prediction on iteration t, $F_{t-1}(x)$ (x) is a prediction on the previous iteration, η is learning rate, J_t is number of leaves on the tree t, $\omega_{j,t}$ is the weight of the leaf j on iteration t, I $(x \epsilon R_{j,t})$ is a function indicator with a value of 1 if the x data falls on the sheet of j on the iteration of t [20], [45].

$$F_t(x) = F_{t-1}(x) + \sum_{j=1}^{J_t} \eta \cdot \omega_{j,t} \cdot I\,(x \in R_{j,t}) \quad (3)$$

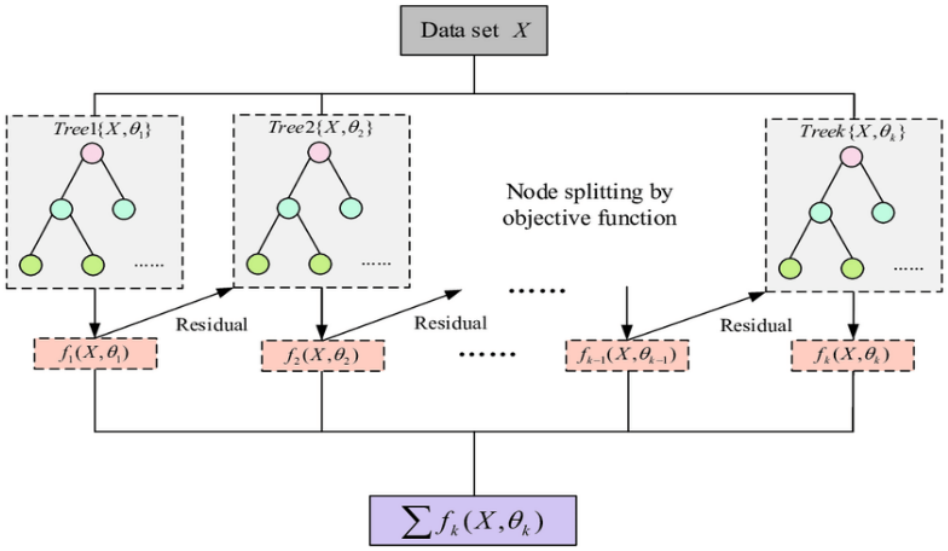

Fig. 2. XGBoost model

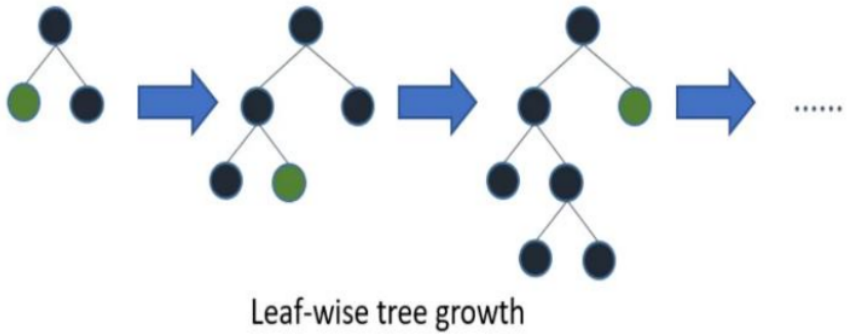

Fig. 3. LGBM model

2.5. CatBoost

Each decision tree f_j feature vector mapping results x_i to the prediction value. This process involves separating data based on features in trees, as well as selecting predictive values on each tree leaf. During training, CatBoost optimizes these trees and tries to minimize the function of the specified loss (e.g., Root Mean Squared Error, RMSE) between the model prediction and the actual target value in training data [46]. Finally, CatBoost regression in Fig. 4 can be modeled as a set of decision trees used to map features x_i to the y_i regression predict value. These trees collectively produce final regression predictions based on combinations of results from each tree. The Catboost regression model is formulated as follows.

$$y_i = \sum_{j=1}^{J} f_j\,(x_i) \quad (4)$$

3. RESULTS AND DISCUSSION

The analysis results emphasize the versatility of machine learning techniques, such as XGBoost, LGBM, and CatBoost, in handling both continuous and discrete variables as inputs. These algorithms can accommodate various types of data, including numerical and categorical features, making them suitable for a wide range of predictive modeling tasks [47], [48]. However, it is important to note that the output variables produced by these algorithms must be discrete, including binary variables, to ensure compatibility with the underlying models. This paper evaluated the predictive performance using scatter plots to visually assess the relationship between the actual and predicted data [49]. The scatter plot's linear relationship indicates that the model effectively captures the underlying patterns in the data. Additionally, the nature of this relationship, whether positive or negative, provides insights into the direction and strength of the association between the variables being considered [50]. The use of scatter plots as a visualization tool provides valuable insights into the predictive accuracy and performance of the machine learning models used in our analysis. Examining the patterns and trends depicted in these plots enhances our understanding of the predictive capabilities of the algorithms and their ability to capture the underlying structure of the data. The Comparison between actual data and prediction data is shown in Fig. 5, Fig. 6, Fig. 7, the three models produce a good linear line. The actuals data and predicted data show a positive relationship.

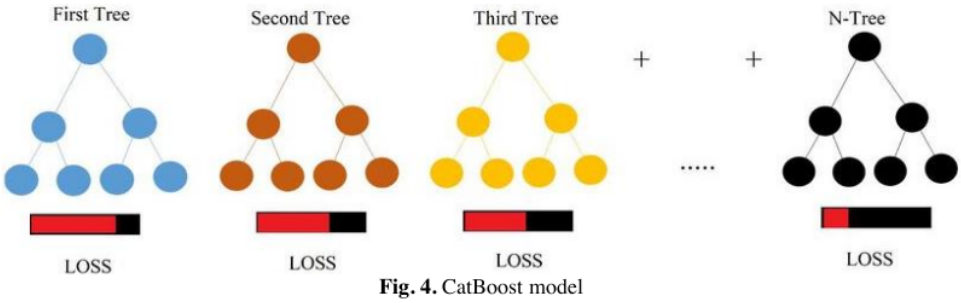

Fig. 4. CatBoost model

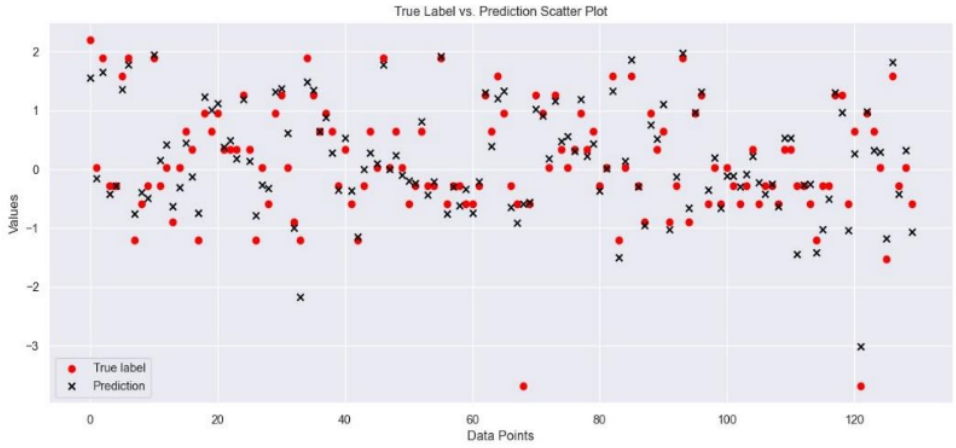

Fig. 5. XGBoost scatterplot prediction

This graph in Fig. 5 is used to compare two sets of data i.e. the actual values (actual labels) versus the predicted values obtained from the XGBoost model. The distribution of the dots looks somewhat random. However, it can be seen that there are some red and black dots that overlap or are close together, indicating that the predictions are close to or match the true labels. The model performs reasonably well for predicting values that are close to the true values, but performs poorly for predicting values that are far from the true values, This plot also shows that the XGBoost model also has points where there is a visible difference between the true and predicted labels, with some significantly separated points.

This graph in Fig. 6 resembles the previously presented graph for XGBoost, where the distribution of dots looks somewhat random. However, the difference lies in the larger number of overlapping or adjacent red and black dots in this graph, indicating that the predictions are closer to or match the actual labels. Thus, it can be concluded that this LGBM model is superior to the XGBoost model. However, just like the XGBoost model, the LGBM model is also poor at predicting values that are far from the true value.

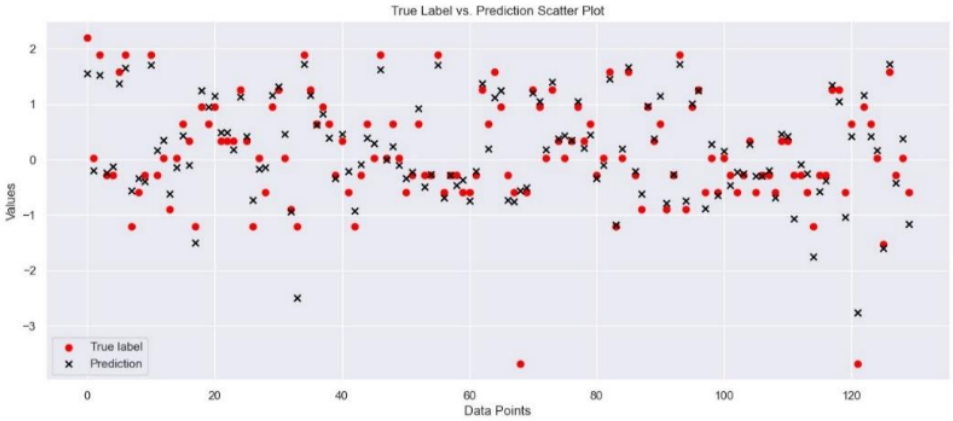

Fig. 6. LGBM scatterplot prediction

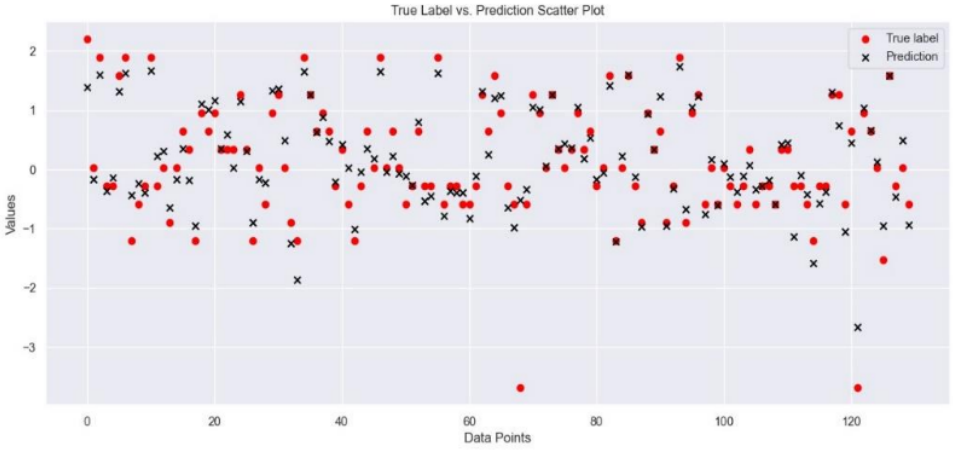

Fig. 7. CatBoost scatterplot prediction

This graph in Fig. 7 displays a similar pattern to the previous two graphs, XGBoost and LGBM, where the distribution of dots looks somewhat random. However, the difference lies in the number of red and black dots that overlap or are close to each other, although not as many as in LGBM, but quite better compared to XGBoost. As with the XGBoost and LGBM models, the Catboost model also shows poor performance in predicting values that are far from the true value.To measure all the gradient boosting regression matrix evaluation scores with XGBoost Regressor, LGBM, and CatBoost Regresor for comparison, containing MAE, MSE, RMSE, and R2 Scores are shown in Table 2, Table 3, and Table 4.

In Table 2 Before tuning, all three algorithms, XGBoost, LGBM, and CatBoost, showed strong performance in the regression task. However, LGBM was consistently superior to the other two algorithms in most metrics, suggesting that LGBM may have captured the underlying patterns in the data more effectively than XGBoost and CatBoost.

Table 2. Regressions results before tuning

	MAE	MSE	RMSE	R2
XGBoost	0.250	0.159	0.399	0.829
LGBM	0.239	0.157	0.397	0.831
CatBoost	0.242	0.160	0.400	0.828

$$R^2 = 1 - \frac{RSS}{TSS} \tag{5}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

Table 3 shows the result after Tunning, all three algorithms show improvements in their performance metrics compared to the version without tunning in Table 2. LGBM continues to perform better than XGBoost and CatBoost in most metrics, demonstrating its superiority and effectiveness in the regression task, even after hyperparameter tuning. In addition, LGBM also achieved the lowest values for other metrics, such as MAE, MSE, and RMSE, indicating that the predictions generated by LGBM have less error compared to XGBoost and CatBoost after tuning.

In result comparison with the previous research that shown in Table 4, It shows the MAE score of this research is significantly better than the previous research with LGBM model. The result show LGBM is the perfect regression model for this portuguese dataset, because Portuguese dataset is a light dataset and the other models, Catboost and XGBoost are designed to handle bigger dataset. This research produces a model to predict factors that determine student academic results. From several factors that determine student academic results, it can be seen that the three gradient boosting algorithms have 10 same importance feature such as "G2, school, absences, famsize, G1, failure, guardian, Fjob, freetime, health". G2 as the most influential feature among them on student academic results, which can be known that G2 is the second period value in numerical form. The feature importance is shown in Fig. 8, Fig. 9, Fig. 10.

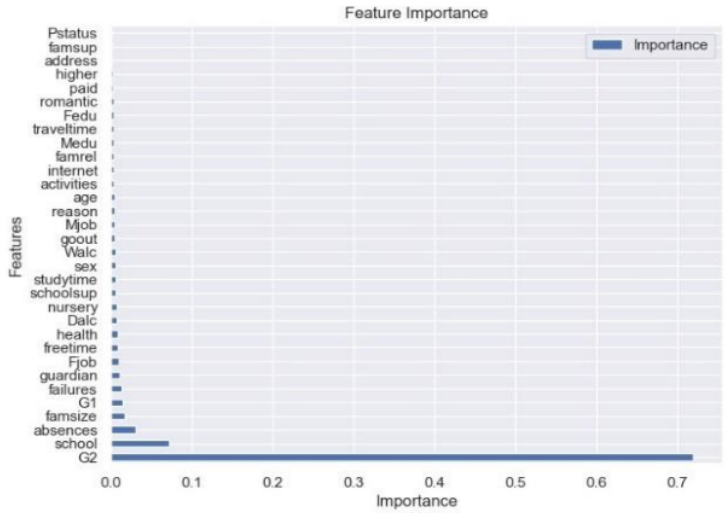

Fig. 8. Feature importance on XGBoost regressor

Table 3. Regressions results after tuning

	MAE	MSE	RMSE	R2
XGBoost	0.227	0.148	0.387	0.8416
LGBM	0.223	0.145	0.381	0.8442
CatBoost	0.233	0.153	0.391	0.8361

Table 4. Comparison best MAE score

Model	MAE
SVR [32]	2.24
LGBM	0.223

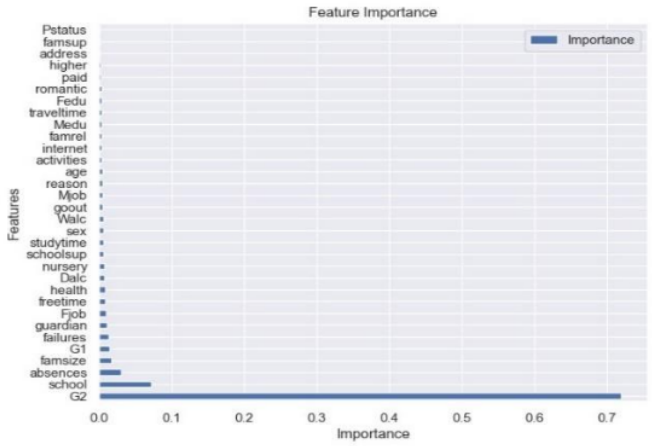

Fig. 9. Feature importance on LGBM

This study shows good results in prediction with MAE score using Gradient Boosting of 0.227, XGBoost of 0.223, and LGBM of 0.233, this study focuses more on the context of regression which has MAE value as an evaluation matrix with the lowest MAE score of 0.223 in LGBM. In this study, although the dataset used is limited, this study has created a model of gradient boosting regression that is good enough to predict the factors that affect student academic results with the LGBM method with the lowest MAE score.

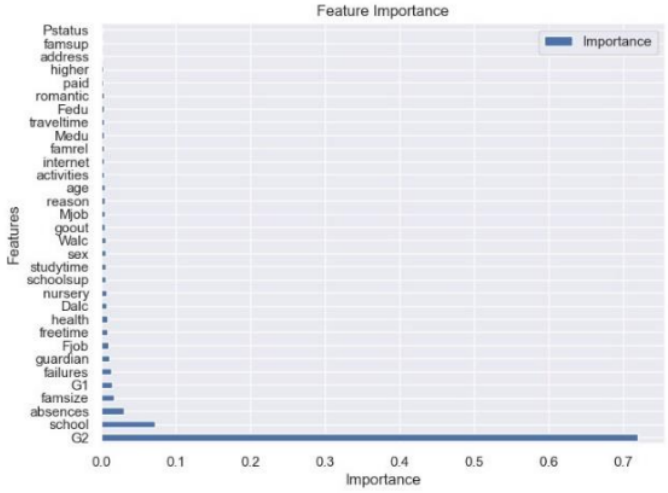

Fig. 10. Feature importance on CatBoost

4. CONCLUSION

This study finds that XGBoost Regressor, LGBM Regressor and CatBoost Regressor have the same ability to predict the final rank (G3) in Portuguese language classes, with the highest MAE value being 0.223 in the LGBM model. The LGBM Regressor model stands out from the other two models due to its ability to adapt to the Portuguese language dataset, which has a limited amount of data. These results can help predict students' final grades and identify students who need additional support to achieve good results [51]. In addition, this study conducted an analysis of the main factors influencing final grades, especially G2, which was identified as a factor that strongly influences students' academic performance. It should be noted that not all schools have access to enough data to make accurate predictions, so coordination is needed to collect accurate data. Nevertheless, this experiment shows that it is essential to consider the prediction of final grades in each school due to its high level of accuracy and reliability. This research successfully addresses the challenge of predicting students' academic performance by applying advanced machine learning techniques. The results provide valuable insights for educational decision-makers and highlight the importance of a data-driven approach to improving academic performance. By using machine learning algorithms, educational institutions can effectively support student learning and success and optimise the use of resources and interventions. Future research is recommended to evaluate the use of newer datasets and explore deep learning methods, such as neural networks, to improve predictive outcomes with the ability to capture more complex patterns.

REFERENCES

[1] H. A. Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," *IEEE Access*, vol. 8, pp. 55462-55470, 2020, https://doi.org/10.1109/ACCESS.2020.2981905.

[2] M. S. Siraj and M. A. R. Ahad, "A Hybrid Deep Learning Framework using CNN and GRU-based RNN for Recognition of Pairwise Similar Activities," *2020 Joint 9th International Conference on Informatics, Electronics & Vision (ICIEV) and 2020 4th International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, pp. 1-7, 2020, https://doi.org/10.1109/ICIEVicIVPR48672.2020.9306630.

[3] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Systematic Ensemble Model Selection Approach for Educational Data Mining," *Knowledge-Based Systems*, vol. 200, p. 105992, 2020, https://doi.org/10.1016/j.knosys.2020.105992.

[4] W. Abdullahi, M. O. Kenneth, and M. Olalere, "Student Performance Prediction Using A Cascaded Bi-level Feature Selection Approach," *Journal of Computer Science Research*, vol. 3, no. 3, pp. 16–28, 2021, https://doi.org/10.30564/jcsr.v3i3.3534.

[5] M. Bikienga, O. Bombiri, and E. Sawadogo, "Design of a machine learning based model for academic performance prediction," *Proceedings of the 5th edition of the Computer Science Research Days*, p. 113, 2023, http://dx.doi.org/10.4108/eai.24-11-2022.2329809.

[6] M. C. Mihaescu and P. S. Popescu, "Review on publicly available datasets for educational data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 3, 2021, p. e1403, https://doi.org/10.1002/widm.1403.

[7] N. P. Dahal and S. Shakya, "An Analysis of Prediction of Students' Results Using Deep Learning," *Computing Open*, vol. 01, 2023, https://doi.org/10.1142/S2972370123500010.

[8] K. T. Chui, R. W. Liu, M. Zhao and P. O. De Pablos, "Predicting Students' Performance With School and Family Tutoring Using Generative Adversarial Network-Based Deep Support Vector Machine," *IEEE Access*, vol. 8, pp. 86745-86752, 2020, https://doi.org/10.1109/ACCESS.2020.2992869.

[9] S. Rai, K. A. Shastry, S. Pratap, S. Kishore, P. Mishra, and H. A. Sanjay, "Machine Learning Approach for Student Academic Performance Prediction," *Evolution in Computational Intelligence*, vol. 1176, pp. 611–618, 2021, https://doi.org/10.1007/978-981-15-5788-0_58.

[10] A. Maulana *et al*., "Leveraging Artificial Intelligence to Predict Student Performance: A Comparative Machine Learning Approach," *Journal of Educational Management and Learning*, vol. 1, no. 2, pp. 64-70, 2023, https://doi.org/10.60084/jeml.v1i2.132.

[11] H. Chen, Y. Wang, C. Xu, C. Xu and D. Tao, "Learning Student Networks via Feature Embedding," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 25-35, 2021, https://doi.org/10.1109/TNNLS.2020.2970494.

[12] A. Vultureanu-Albişi and C. Bădică, "Improving Students' Performance by Interpretable Explanations using Ensemble Tree-Based Approaches," *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pp. 215-220, 2021, https://doi.org/10.1109/SACI51354.2021.9465558.

[13] Z. U. Abideen *et al*., "Analysis of Enrollment Criteria in Secondary Schools Using Machine Learning and Data Mining Approach," *Electronics*, vol. 12, no. 3, p. 694, 2023, https://doi.org/10.3390/electronics12030694.

[14] L. H. Alamri, R. S. Almuslim, M. S. Alotibi, D. K. Alkadi, I. Ullah Khan, and N. Aslam, "Predicting Student Academic Performance using Support Vector Machine and Random Forest," *Proceedings of the 2020 3rd International Conference on Education Technology Management*, pp. 100–107, 2020, https://doi.org/10.1145/3446590.3446607.

[15] O. Kherif, Y. Benmahamed, M. Teguar, A. Boubakeur and S. S. M. Ghoneim, "Accuracy Improvement of Power Transformer Faults Diagnostic Using KNN Classifier With Decision Tree Principle," *IEEE Access*, vol. 9, pp. 81693-81701, 2021, https://doi.org/10.1109/ACCESS.2021.3086135.
[16] J. Malini, and Y. Kalpana, "Analysis Of Factors Affecting Student Performance Evaluation Using Education DataminingTechnique," *Materialstoday: Proceedings*, vol. 47, pp. 6105-6110, 2021, https://doi.org/10.1016/j.matpr.2021.05.026.
[17] K. D. R, O. K. S, N. P. G, "Student performance classification: A data mining approach," *JIMS8I - International Journal of Information Communication and Computing Technology*, vol. 8, no. 2, pp. 462–466, 2020, http://dx.doi.org/10.5958/2347-7202.2021.00001.3.
[18] E. A. Yekun and A. T. Haile, "Student Performance Prediction with Optimum Multilabel Ensemble Model," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 511–523, 2021, https://doi.org/10.1515/jisys-2021-0016.
[19] N. Chauhan, K. Shah, D. Karn, and J. Dalal, "Prediction of Student's Performance Using Machine Learning," *2nd International Conference on Advances in Science & Technology (ICAST)*, 2019, https://doi.org/10.2139/ssrn.3370802.
[20] Y. Jang, S. Choi, H. Jung, and H. Kim, "Practical early prediction of students' performance using machine learning and eXplainable AI," *Education and Information Technologies*, vol. 27, pp. 12855–12889, 2022, https://doi.org/10.1007/s10639-022-11120-6.
[21] M. Arifin, W. Widowati, F. Farikhin, and G. Gudnanto, "A Regression Model and a Combination of Academic and Non-Academic Features to Predict Student Academic Performance," *TEM Journal*, vol. 2, no. 2, pp. 855-864, 2023, https://doi.org/10.18421/TEM122-31.
[22] N. M. Shahani, X. Zheng, C. Liu, F. U. Hassan, and P. Li, "Developing an XGBoost Regression Model for Predicting Young's Modulus of Intact Sedimentary Rocks for the Stability of Surface and Subsurface Structures," *Frontiers in Earth Science*, vol. 9, p. 761990, 2021, https://doi.org/10.3389/feart.2021.761990.
[23] A. Shehadeh, O. Alshboul, R. E. Al Mamlook, and O. Hamedat, "Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression," *Automation in Construction*, vol. 129, p. 103827, 2021, https://doi.org/10.1016/j.autcon.2021.103827.
[24] A. Tariq, Y. Niaz, A. Amin, "Systematic Approach for Re-Sampling and Prediction of Low Sample Educational Datasets," *International Journal of Computing and Digital Systems*, vol. 12, no. 1, pp. 1203–1214, 2022, https://doi.org/10.12785/ijcds/120196.
[25] Y. Qiu, J. Zhou, M. Khandelwal, H. Yang, P. Yang, and C. Li, "Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration," *Engineering with Computers*, vol. 38, pp. 4145–4162, 2022, https://doi.org/10.1007/s00366-021-01393-9.
[26] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899-67911, 2020, https://doi.org/10.1109/ACCESS.2020.2986809.
[27] X. Zhang, C. Yan, C. Gao, B. A. Malin, and Y. Chen, "Predicting Missing Values in Medical Data Via XGBoost Regression," *Journal of Healthcare Informatics Research*, vol. 4, no. 4, pp. 383–394, 2020, https://doi.org/10.1007/s41666-020-00077-1.
[28] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 243-247, 2019, https://doi.org/10.1109/COMITCon.2019.8862214.
[29] E. Ileberi, Y. Sun and Z. Wang, "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost," *IEEE Access*, vol. 9, pp. 165286-165294, 2021, https://doi.org/10.1109/ACCESS.2021.3134330.
[30] M. I. Khan, Z. A. Khan, A. Imran, A. H. Khan and S. Ahmed, "Student Performance Prediction in Secondary School Education Using Machine Learning," *2022 8th International Conference on Information Technology Trends (ITT)*, pp. 94-101, 2022, https://doi.org/10.1109/ITT56123.2022.9863971.
[31] S. S. Madnaik, "Predicting Students' Performance by Learning Analytics," *San Jose State University*, 2020, https://doi.org/10.31979/etd.6jjb-ua9w.
[32] R. Li *et al*., "Estimation of Blood Alcohol Concentration From Smartphone Gait Data Using Neural Networks," *IEEE Access*, vol. 9, pp. 61237-61255, 2021, https://doi.org/10.1109/ACCESS.2021.3054515.
[33] S. Begum and S. S. Padmannavar, "Genetically Optimized Ensemble Classifiers for Multiclass Student Performance Prediction," *International Journal of Intelligent Engineering and Systems*, vol. 15, no. 2, pp. 316–328, 2022, https://doi.org/10.22266/ijies2022.0430.29.
[34] J. Y. Chan, H. Ng, T. T. V. Yap, and V. T. Goh, "Predictive Modelling of Student Performance in MMU Based on Machine Learning Approach," *Proceedings of the International Conference on Computer, Information Technology and Intelligent Computing (CITIC 2022)*, pp. 258–278, 2022, https://doi.org/10.2991/978-94-6463-094-7_21.
[35] N. Mahat, N. I. Nording, J. Bidin, S. Abu Hasan, and Teoh Yeong Kin, "Artificial Neural Network (ANN) to Predict Mathematics Students' Performance," *Journal of Computing Research and Innovation*, vol. 7, no. 1, pp. 29–38, 2022, https://doi.org/10.24191/jcrinn.v7i1.264.
[36] A. Abdelrahman, T. Hassan, and A. Soliman, "A Predictive Model for Student Performance in Classrooms Using Student Interactions With an eTextbook," *Researh Square*, 2022, https://doi.org/10.21203/rs.3.rs-1353605/v3.

[37] Z. Chen, X. Wang, W. Liao, and Z. Du, "Exploratory Data Analysis AndPrediction on Student Performance," *International Core Journal of Engineering*, vol. 7, pp. 2414–1895, 2021, https://doi.org/10.6919/ICJE.202111_7(11).0038.

[38] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, A. Dakkak, and Y. El Allioui, "A Multiple Linear Regression-Based Approach to Predict Student Performance," *Advances in Intelligent Systems and Computing, Springer Science and Business Media Deutschland GmbH*, pp. 9–23, 2020, https://doi.org/10.1007/978-3-030-36653-7_2.

[39] J. L. Harvey and S. A. P. Kumar, "A Practical Model for Educators to Predict Student Performance in K-12 Education using Machine Learning," *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 3004-3011, 2019, https://doi.org/10.1109/SSCI44817.2019.9003147.

[40] Y. Sun, M. Peng, Y. Zhou, Y. Huang and S. Mao, "Application of Machine Learning in Wireless Networks: Key Techniques and Open Issues," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3072-3108, 2019, https://doi.org/10.1109/COMST.2019.2924243.

[41] E. Evangelista, "An Optimized Bagging Ensemble Learning Approach Using BESTrees for Predicting Students' Performance," *International Journal of Emerging Technologies in Learning*, vol. 18, no. 10, pp. 150–165, 2023, https://doi.org/10.3991/ijet.v18i10.38115.

[42] A. Hennebelle, L. Ismail, and T. Linden, "Schools Students Performance with Artificial Intelligence Machine Learning: Features Taxonomy, Methods and Evaluation," *Preprints*, 2023, https://doi.org/10.20944/preprints202308.1358.v1.

[43] B. Sekeroglu, R. Abiyev, A. Ilhan, M. Arslan, and J. B. Idoko, "Systematic literature review on machine learning and student performance prediction: Critical gaps and possible remedies," *Applied Sciences*, vol. 11, no. 22, p. 10907, 2021, https://doi.org/10.3390/app112210907.

[44] Karale, A. Narlawar, B. Bhujba, and S. Bharit, "Student Performance Prediction using AI and ML," *Int J Res Appl Sci Eng Technol*, vol. 10, no. 6, pp. 1644–1650, 2022, https://doi.org/10.22214/ijraset.2022.44032.

[45] H. Lee and J.-W. Lee, "Why East Asian Students Perform Better in Mathematics than Their Peers: An Investigation Using a Machine Learning Approach," *Centre for Applied Macroeconomic Analysis*, 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3896033.

[46] Z. Fan, J. Gou, and C. Wang, "Predicting secondary school student performance using a double particle swarm optimization-based categorical boosting model," *Eng Appl Artif Intell*, vol. 124, 2023, https://doi.org/10.1016/j.engappai.2023.106649.

[47] N. Aslam, I. U. Khan, L. H. Alamri, and R. S. Almuslim, "An Improved Early Student's Performance Prediction Using Deep Learning," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 12, pp. 108–122, 2021, https://doi.org/10.3991/ijet.v16i12.20699.

[48] C. Dervenis, V. Kyriatzis, S. Stoufis, and P. Fitsilis, "Predicting Students' Performance Using Machine Learning Algorithms," *ACM International Conference Proceeding Series, Association for Computing Machinery*, 2022, https://doi.org/10.1145/3564982.3564990.

[49] R. Mehdi and M. Nachouki, "A neuro-fuzzy model for predicting and analyzing student graduation performance in computing programs," *Educ Inf Technol (Dordr)*, vol. 28, no. 3, pp. 2455–2484, 2023, https://doi.org/10.1007/s10639-022-11205-2.

[50] E. De Leon Evangelista and B. D. Sy, "An approach for improved students' performance prediction using homogeneous and heterogeneous ensemble methods," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 5, pp. 5226–5235, 2022, https://doi.org/10.11591/ijece.v12i5.pp5226-5235.

[51] B. Sekeroglu, Y. K. Ever, K. Dimililer, and F. Al-Turjman, "Comparative Evaluation and Comprehensive Analysis of Machine Learning Models for Regression Problems," *Data Intell*, vol. 4, no. 3, pp. 620–652, 2022, https://doi.org/10.1162/dint_a_00155.

CatBoost_Data_Mining_Light_Gradient_Boosting_Machine_...

ORIGINALITY REPORT

20%	15%	14%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

3%

★ Submitted to University of Warwick

Student Paper

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		