WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

Masters Programmes: Group Assignment Cover Sheet

Student Numbers: Please list numbers of all group members	5661438, 5600375, 5676282, 5662483, 5594410, 5667293
Module Code:	IB9BW0
Module Title:	Analytics in Practice
Submission Deadline:	2 Dec 2024, 12:00
Date Submitted:	2 Dec 2024
Word Count:	1949 words
Number of Pages:	16
Question Attempted: *(question number/title, or description of assignment)*	Attempted all questions
Have you used Artificial Intelligence (AI) in any part of this assignment?	Yes

Academic Integrity Declaration

We're part of an academic community at Warwick. Whether studying, teaching, or researching, we're all taking part in an expert conversation which must meet standards of academic integrity. When we all meet these standards, we can take pride in our own academic achievements, as individuals and as an academic community.

Academic integrity means committing to honesty in academic work, giving credit where we've used others' ideas and being proud of our own achievements.

In submitting my work, I confirm that:

- I have read the guidance on academic integrity provided in the Student Handbook and understand the University regulations in relation to Academic Integrity. I am aware of the potential consequences of Academic Misconduct.
- I declare that this work is being submitted on behalf of my group and is all our own, , except where I have stated otherwise.
- No substantial part(s) of the work submitted here has also been submitted by me in other credit bearing assessments courses of study (other than in certain cases of a resubmission of a piece of work), and I acknowledge that if this has been done this may lead to an appropriate sanction.
- Where a generative Artificial Intelligence such as ChatGPT has been used I confirm I have abided by both the University guidance and specific requirements as set out in the Student Handbook and the Assessment brief. I have clearly acknowledged the use of any generative Artificial Intelligence in my submission, my reasoning for using it and which generative AI (or AIs) I have used. Except where indicated the work is otherwise entirely my own.
- I understand that should this piece of work raise concerns requiring investigation in relation to any of points above, it is possible that other work I have submitted for assessment will be checked, even if marks (provisional or confirmed) have been published.
- Where a proof-reader, paid or unpaid was used, I confirm that the proof-reader was made aware of and has complied with the University's proofreading policy.

Upon electronic submission of your assessment you will be required to agree to the statements above

Table Of Contents

5676282, 5662483, 5594410, 5667293, 5600375, 5661438

Abstract

The report details a predictive model using Gradient Boosted Decision Trees (GBDT) to identify customers likely to leave positive reviews. By leveraging the CRISP-DM framework, it analyses customer data to enhance engagement, support marketing strategies, and strengthen platform credibility, aligning with Nile's goal of fostering favorable feedback and long-term growth.

1. Business Understanding

For e-commerce platforms like Nile, engaging customers to generate positive reviews is crucial for brand reputation and sales. Nile faces the challenge of being unable to target the right customers for review requests efficiently. To address this, Nile plans to implement a predictive model to identify customers likely to leave favourable reviews, optimising resource allocation. This report outlines how Nile can enhance review generation using data mining techniques guided by the CRISP-DM methodology. It will cover data processing, predictive modelling methods such as GBDT and Decision Trees, and model evaluation using metrics like precision and recall. The report concludes with actionable strategies to improve customer engagement and review targeting.

2. Data Understanding

The Entity-Relationship Diagram (ERD) in Figure 1 demonstrates the eight tables provided by Nile, detailing their features and interconnections. The ‘order_review’ dataset provided valuable customer feedback crucial for quality and satisfaction monitoring. The ‘review_score’ field could be categorised as positive or negative. This approach aligns with identifying customers likely to leave positive reviews. The ‘customer_unique_id’ further enhances analysis by tracking customer activity across multiple orders. The orders table is the central hub, connecting customers, products, reviews, and payment data.

Figure 1: ERD

Figure 2 illustrates the review score distribution. It revealed that most customers provide positive feedback, with a significant proportion assigning scores 4 and 5. In contrast, lower review scores (1 to 3) are less frequent, indicating generally satisfactory customer experiences.

Figure 2: Review score distribution

Service quality is increasingly recognised as playing an important role in influencing the purchase intentions of online customers[1]. Accordingly, ACG developed three time-phase features (Table 1), derived from time-based metrics in the orders table, to evaluate the impact of different process stages on customer review scores.

Features	Implementation	Calculation
order_approving_time	Understand the approval latency for customer orders	the time difference between 'order_approved_at' and 'order_purchase_timestamp'
delivery_time	Evaluate the shipping performance	the difference between 'order_delivered_customer_date' and 'order_purchase_timestamp'

Features	Implementation	Calculation
delivery_date_diff	Highlight whether orders are delivered earlier, on time, or later than expected.	the difference between 'order_delivered_customer_date' and 'order_estimated_delivery_date'

Table 1: Time-phased features

Figure 3 demonstrates the impact of various time-phase features on customer review scores. The diagram suggests a strong correlation between shorter approval times, faster delivery, and consistent, accurate delivery estimates, leading to more positive reviews.

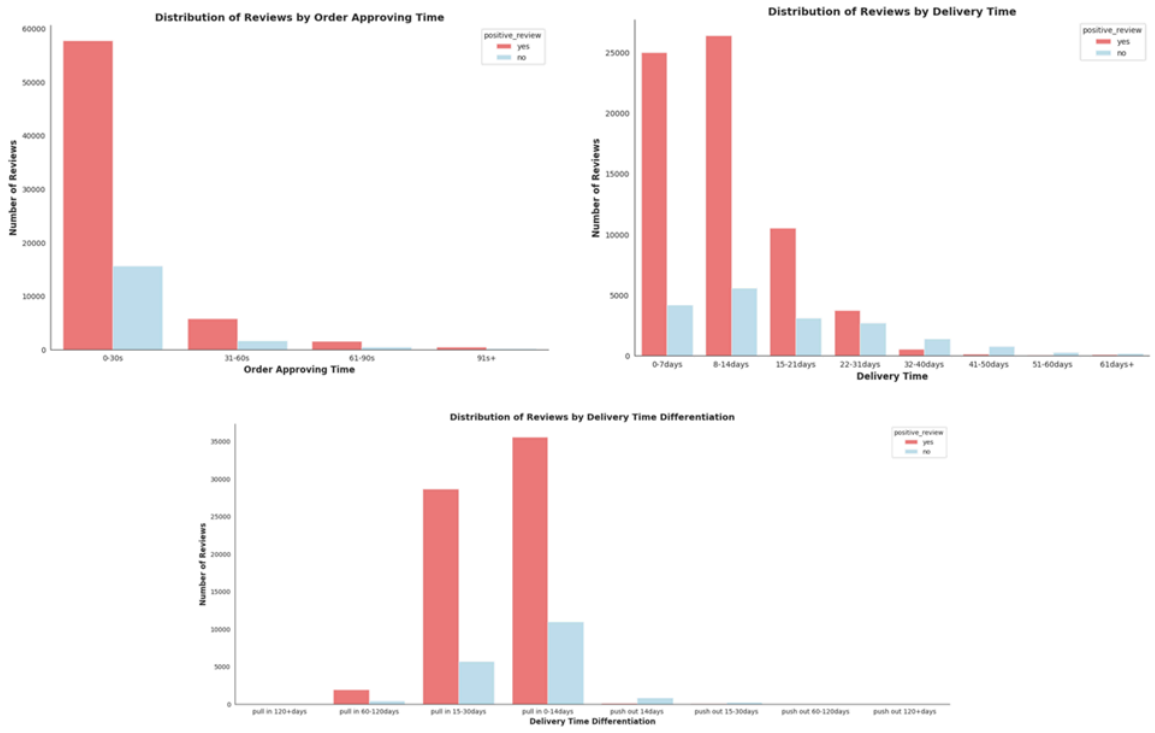

Figure 3: Review score within time-phase features

3. Data selection

The selected data columns provided a strong foundation for analysing customer review behaviour. Order details, such as status, approval, and delivery time,

emphasise the impact of service reliability on satisfaction. Regional demographics highlight expectation differences and review tendencies. Product attributes, including price and categories, payment types and transaction values, shape perceptions of convenience and value.

4. Data Preparation

4.1. Handling Duplicates in 'order_id'

'order_id' field is the reliable foreign key for merging tables. To ensure every order is uniquely represented in the dataset, ACG removed duplicates by retaining only the first occurrence of each 'order_id'.

4.2. Handling Outliers

By calculating the time differences between purchase, approval, and delivery timestamps, several significant delays, such as those exceeding one day, are identified. Therefore, ACG sorted and removed the top four entries with the most important values (Figure 4).

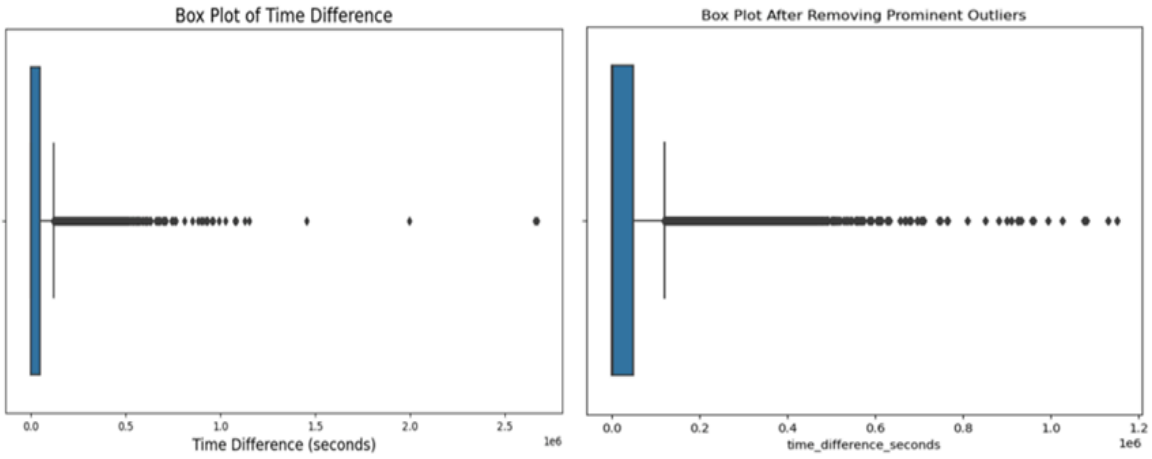

Figure 4: Boxplot for timestamp

4.3. Data Joining Strategy

To create a comprehensive dataset, using the 'order_reviews' table as the primary (left) table for merged purposes. This approach retained all reviews, ensuring that no critical insights were overlooked. In addition, the following three joins were performed to consolidate into a single table, guaranteeing the aggregation of all relevant features.

- **Products, Order Items, and Order Reviews:** This integration enriched the dataset with product details, pricing, and freight costs for each order item.
- **Customers, Orders, and Order Reviews:** This process provides customer demographic information, timestamps, and statuses to facilitate behavioural analysis.
- **Order Payments and Order Reviews:** This method adds payment type and values, enhancing the dataset's depth.

4.4. Addressing Missing Values

Figure 5 illustrates the number of missing values in the dataset. The removed rows represented only a tiny fraction (approximately 1.5%) of the total dataset, which exceeded 80,000 records.

```
print(final_df.isnull().sum())
```

```
customer_unique_id                0
customer_zip_code_prefix          0
customer_state                    0
order_status                      0
order_purchase_timestamp          0
order_approved_at                 0
order_delivered_customer_date     0
order_estimated_delivery_date     0
review_score                      0
product_category_name          1214
product_description_lenght     1214
product_photos_qty             1214
seller_id                         2
price                             2
freight_value                     2
payment_type                      0
payment_value                     0
```

Figure 5: Missing value

4.5. Feature Engineering

Time-based features, such as ‘order_approving_time’, ‘delivery_time’, and ‘delivery_date_diff’, were derived to capture operational efficiency. Besides, binary classification is classifying given information based on predefined classes[2]. Therefore, review scores of 4 and 5 were categorised as positive reviews and class 1, while scores from 1 to 3 are classified as negative reviews and class 0. Furthermore, figure 6 presented each column’s data type for analysis preparation.

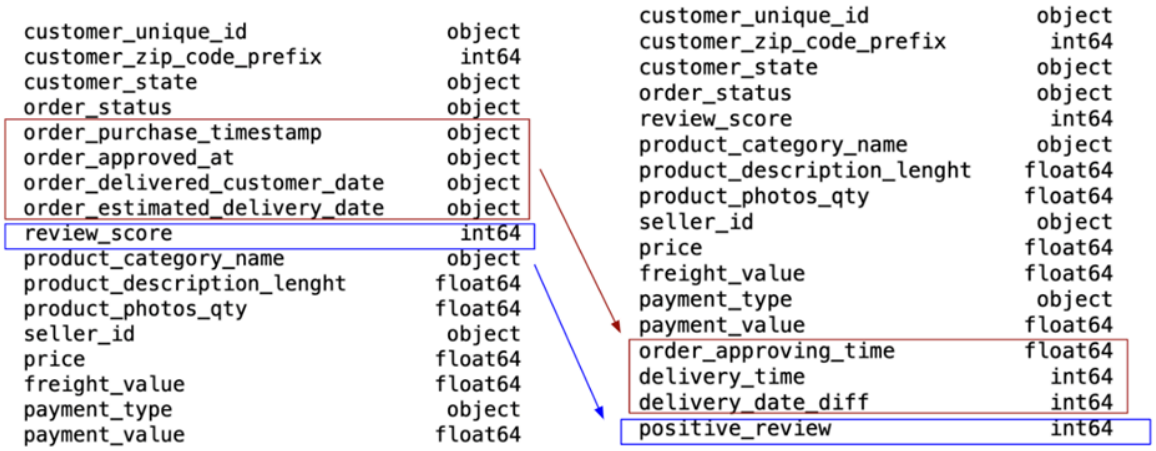

Figure 6: Data type

4.6. Encoding categorical features

Most machine learning models work better with numerical data[3]. The categorical variables must be encoded into numerical values for algorithms to work. Moreover, encoding does not change data dimensionality; it is significantly more memory-efficient[4]. Therefore, two encoding techniques were employed.

- **One-Hot Encoding:** Categorical features with low cardinality, such as 'order_status', 'customer_state', 'product_category_name', and 'payment_type'.
- **Frequency Encoding:** categorical features with high cardinality, including 'customer_unique_id' and 'seller_id'.

5. Modelling

5.1. Data modelling

Gradient boosting ensemble machine learning process focused on the two-class prediction[5] to identify the key factors driving customers to leave positive reviews. Moreover, GBDT is particularly effective for structured datasets, as it captures complex relationships and delivers high predictive performance, especially in scenarios involving imbalanced data[6]. Thus, to improve prediction performance and avoid overfitting, ACG used a gradient-boosting classifier and random search cross-validation instead of a decision tree model. Additionally, ACG performed gradient-boosted decision trees (GBDT), eXtreme gradient boosting (xGBoost), and random forest models for pre-modelling. Alternatively, Random Forest provides a more

resource-efficient solution while maintaining robust accuracy. It is computationally less demanding and more accessible to implement, making it a practical choice for clients prioritising cost-effectiveness[7]. Separately, XGBoost can deal with the bias-variance tradeoff more carefully, which decreases the training time of the model[8]. Finally, according to the precision consequences, ACG applied gradient-boosted decision trees as the modelling deployment.

5.2. Split data

The dataset comprises 114 features, including 14 variables and 100 dummy variables. Since dataset splitting is considered indispensable and highly necessary to eliminate or reduce bias in training data in machine learning models[9], ACG divided the dataset into an 80% training set and a 20% testing set, containing 67,554 and 16,889 records, respectively.

5.3. Binary Classification: Gradient Boosting Classifier

The aim is to quantitatively evaluate the classification performance using key metrics: accuracy, precision, recall, and F1-score. The parameter is set to macro to calculate the arithmetic mean of these metrics for binary classes, enabling an assessment of the model's overall performance across all classes without depending on class distribution. Table 2 presents the macro value of the training and testing dataset.

Type	Training Value	Testing Value	Remark
Accuracy	0.822	0.821	Train and test samples were correctly classified
Macro Precision	0.815	0.807	Strong performance in minimising false positives
Macro Recall	0.607	0.595	Low capacity to identify true positives
Macro F1-score	0.627	0.611	Balanced trade-off between precision and recall

Table 2: Macro value of training and testing dataset

5.4. Fine-tuning hyperparameter: RandomizedSearchCV for class 1 and macro

The objective is to determine the optimal set of hyperparameters to improve the model's classification performance while ensuring its effectiveness across all classes without bias toward class distribution. The former approach focuses on optimising the model for precision performance on class 1 (positive review), while the latter evaluates the overall classification performance using macro-averaged metrics. Table 3 presents the performance of these two methods.

Type	Class 0 (negative review)	Class 1 (positive review)	Macro-averaged
Precision	0.79	0.82	0.81
Recall	0.20	0.99	0.61
F1-score	0.32	0.90	0.62

Table 3: Hyperparameter results

6. Evaluation

This analysis compares the performance of three machine learning models, GBDT, XGBoost and random forest, in a binary classification task.

6.1. Confusion matrix and classification report

The three models' confusion matrices and classification reports were analysed to assess their ability to correctly classify each class (Class 0 and Class 1). Figure 7 compared each model's confusion matrix, and Table 4 presented the classification report of the three models.

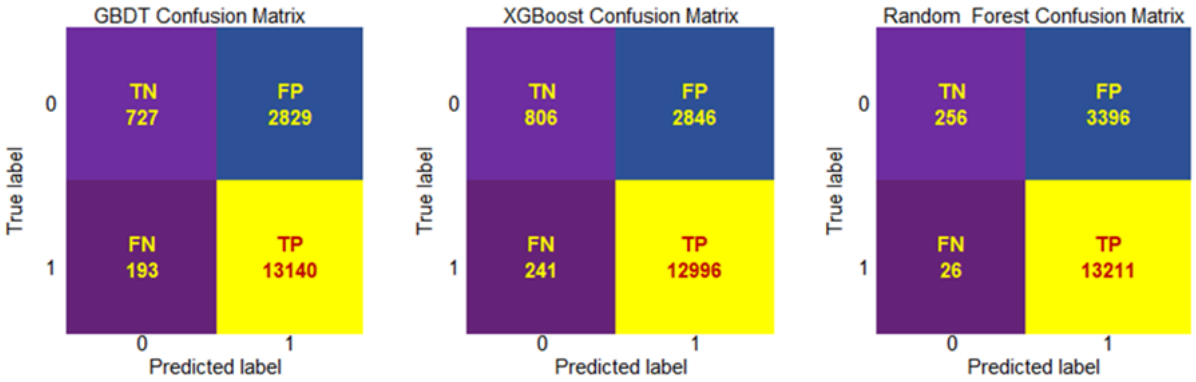

Figure 7: Confusion matrix of GBDT, XGBoost, and random forest

Metric	GBDT			XGBoost			Random Forest		
	Class 0	Class 1	Overall (macro)	Class 0	Class 1	Overall (macro)	Class 0	Class 1	Overall (macro)
Precision	0.79	0.82	0.81	0.77	0.82	0.80	0.91	0.80	0.86
Recall	0.20	0.99	0.59	0.22	0.98	0.60	0.07	1.00	0.53
F1-Score	0.32	0.90	0.61	0.34	0.89	0.62	0.13	0.89	0.51
Accuracy	—	—	0.82	—	—	0.82	—	—	0.80
Weighted Average Precision			0.82			0.81			0.82
Weighted Average Recall			0.82			0.82			0.80
Weighted Average F1-Score			0.78			0.77			0.72

Table 4: Classification reports of GBDT, XGBoost, and random forest

6.2. Model Analysis

GBDT is better suited for applications prioritising accurately identifying positive instances (Class 1) due to its higher recall and F1-Score and ability to minimise false negatives effectively.

6.3. Top 10 most important features

Figure 7 contributes to predictive accuracy. According to Figure 7, delivery date difference is the critical determinant of the target variable.

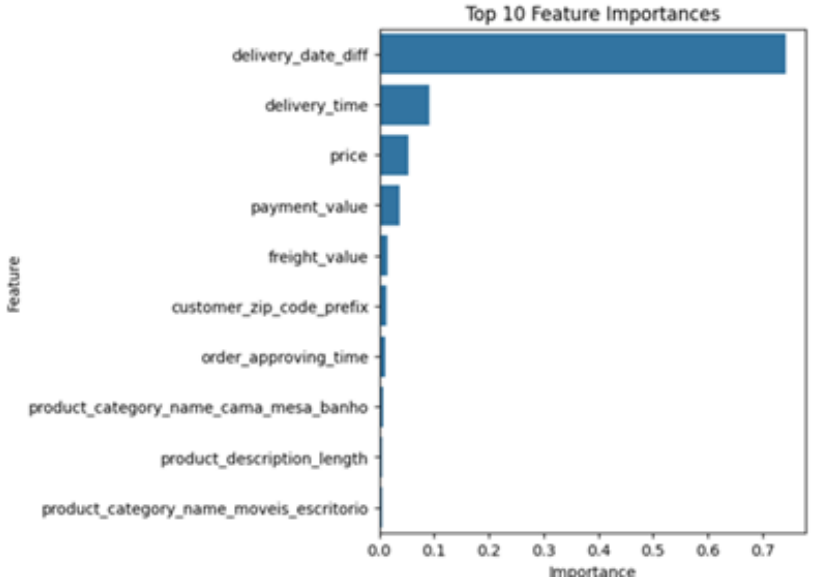

Figure 7: Top 10 most important features

7. Deployment

The three-phase goals are to ensure the model's practical reliability and seamless integration into operational workflows, supporting informed decision-making.

7.1. Phase 1 - model deployment

Deployment environments, including cloud platforms and ETL pipelines, are prepared to meet the model's computational requirements. Additionally, during pilot testing and complete model verification, A/B testing is conducted to validate the credibility of the commercial strategy.

7.2. Phase 2 - system integration

Logistics systems and customer relationship management (CRM) systems will be integrated to leverage predictive analytics to identify and address factors contributing to delivery delays. Behavioural and transactional customer data should also be incorporated to design tailored marketing

strategies. APIs will be utilised to integrate these systems, enabling seamless data transfer and efficient processing within a cloud-based infrastructure.

7.3. Phase 3 - continual learning

An automated model retraining system will be designed to continuously monitor data drift and update the model with new datasets to maintain.

8. Conclusions

The GBDT prediction model offers superior precision (0.82) for identifying positive reviews while minimising false negatives. The integration of enriched time-phase features and categorical encodings significantly enhanced predictive accuracy. Moreover, key findings highlight the importance of service quality, including delivery timeliness and customer demographics, in shaping review behaviour.

Moving onto the three phases of the deployment plan: model deployment, system integration, and continual learning. First, the A/B test validates the model's commercial strategy. Second, system integration ensures seamless data transformation and processing through cloud platforms. Finally, continuous model retraining is designed to address data drift and maintain robust performance.

8.1. Imbalanced data

According to Figure 2 (Review score distribution), the numbers of positive and negative reviews are significantly disparate. This imbalance reduces the model's ability to prevent false positive reviews accurately.

8.2. Lack of analysis of review comments

Although numerical review scores are extensively analysed, the lack of exploration of textual review comments represents a missed opportunity. These comments can offer contextual understanding and improve the model's predictive capabilities.

9. Recommendations

9.1. Handle imbalanced data

Apply SMOTE to oversample the minority class or undersampling method to the majority class to improve the model's ability to predict negative reviews.

9.2. Analysis of review comments

Implement natural language processing (NLP) techniques to extract insights from textual review comments through sentiment analysis and topic

modelling. This can complement numerical analysis and enhance the model's understanding of customer feedback.

10. References

1. Kaushik, K., Mishra, R., Rana, N.P. and Dwivedi, Y.K. (2018). Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on Amazon.in. *Journal of Retailing and Consumer Services*, 45(0969-6989), pp.21–32. doi:https://doi.org/10.1016/j.jretconser.2018.08.002.
2. Kumari, R. and Kr., S. (2017). Machine Learning: A Review on Binary Classification. *International Journal of Computer Applications*, 160(7), pp.11–15. doi:https://doi.org/10.5120/ijca2017913083.
3. Nishoak Kosaraju, Sainath Reddy Sankepally and K. Mallikharjuna Rao (2023). Categorical Data: Need, Encoding, Selection of Encoding Method and Its Emergence in Machine Learning Models—A Practical Review Study on Heart Disease Prediction Dataset Using Pearson Correlation. *Lecture notes in networks and systems*, 551(2367-3370), pp.369–382. doi:https://doi.org/10.1007/978-981-19-6631-6_26.
4. Kassymzhomart Kunanbayev, Islambek Temirbek and Amin Zollanvari (2021). Complex Encoding. *International Joint Conference on Neural Networks (IJCNN)*, (2161-4393). doi:https://doi.org/10.1109/ijcnn52387.2021.9534094.
5. Bahad, P. and Saxena, P. (2019). Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics. *International Conference on Intelligent Computing and Smart Communication 2019*, (2524-7565), pp.235–244. doi:https://doi.org/10.1007/978-981-15-0633-8_22.
6. Chen, T. and Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, (978-1-4503-4232-2), pp.785–794. doi:https://doi.org/10.1145/2939672.2939785.
7. Muraina, I. (2022). *IDEAL DATASET SPLITTING RATIOS IN MACHINE LEARNING ALGORITHMS: GENERAL CONCERNS FOR DATA SCIENTISTS AND DATA ANALYSTS*. 7th INTERNATIONAL MARDIN ARTUKLU SCIENTIFIC RESEARCHES CONFERENCEwww.artuklukongresi.org Mardin, Turkey496.
8. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), pp.5–32. doi:https://doi.org/10.1023/a:1010933404324.
9. Uzir, N., Raman, S. and Banerjee, S. (2017). *Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets 1 PUBLICATION 177 CITATIONS SEE PROFILE 1 PUBLICATION 177 CITATIONS SEE PROFILE Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets*.

International Journal of Control Theory and Applications International Journal of Control Theory and Applications.