# National College of Ireland

# Project Submission Sheet – 2021/2022

| | | | |
|---|---|---|---|
| **Student Name:** | Nitish Sharma, Komal, Rajbharath Jothimani,Akshay Menon, Babita Chaini | | |
| **Student ID:** | x21154147,x21148082, x21133000,x21173036, x21139211 | | |
| **Programme:** | MSc. Data Analytics | **Year:** | 2022 |
| **Module:** | Domain Application of Predictive Analytics (MSCDAD_A) | | |
| **Lecturer:** | Vikas Sahni | | |
| **Submission Due Date:** | 12 August 2022 | | |
| **Project Title:** | Predicting the late delivery shipments in-order to improve the supply chain and gain customer satisfaction | | |
| **Word Count:** | 3236 | | |

**I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.**
**ALL internet material must be referenced in the references section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.**

**Signature:** Nitish Sharma, Komal, Rajbharath Jothimani,Akshay Menon, Babita Chaini

**Date:** 12 August 2022

**PLEASE READ THE FOLLOWING INSTRUCTIONS:**

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. Projects should be submitted to your Programme Coordinator.
3. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
4. You must ensure that all projects are submitted to your Programme Coordinator on or before the required submission date. **Late submissions will incur penalties.**
5. All projects must be submitted and passed in order to successfully complete the year. **Any project/assignment not submitted will be marked as a fail.**

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Predicting the late delivery shipments in order to improve the supply chain and gain customer satisfaction

Babita Chaini
*School of Computing*
*National College of Ireland*
Dublin, Ireland
x21139211@student.ncirl.ie

Komal
*School of Computing*
*National College of Ireland*
Dublin, Ireland
x21148082@student.ncirl.ie

Nithish Sharma
*School of Computing*
*National College of Ireland*
Dublin, Ireland
x21154147@student.ncirl.ie

Rajbharath Jothimani
*School of Computing*
*National College of Ireland*
Dublin, Ireland
x21133000@student.ncirl.ie

Akshay Menon
*School of Computing*
*National College of Ireland*
Dublin, Ireland
x21173036@student.ncirl.ie

***Abstract*—The primary goal of this project is to create a machine learning model that predicts the risk of delayed delivery anticipated in supply chain deliveries in order to increase customer satisfaction. In this paper various predictive methods are being built using different machine learning and data modelling techniques in order to predict late delivery. Python is used as programming language and Tableau is used as data visualization tool. XGBoost machine learning algorithm is the model opted for this research based on the analysis of previous researches conducted on similar studies and is evaluated based on different criteria such as confusion matrix, accuracy, F1-score and recall. Hence, it is used to predict the delivery risk involved in the supply chain management system. This paper also provides tips for using the machine learning model obtained to help business in making successful late delivery predictions, along with business insights and conclusions from the data being observed.**

***Keywords*–recall, insights, model, accuracy, algorithm**

## I. INTRODUCTION

Predictive analytics is an useful tool for supply chain optimization and late delivry risk mitigation as it helps in lowering logistics costs and raising customer service standards. According to Survey done by MHI annual industry report 2020 Predictive analytics is used by 30% of supply chain professionals, and 57% of businesses that aren't yet utilizing it aims to implement it within the next five years [1]. The main goal of this paper is to identify the source of supply chain risks. This paper focuses on one of the risk area of supply chain which is late delivery. In order to minimize losses incurred as a result of this risks, the organization can try to optimize the factors that are responsible for the risk. Hence, a risk detection model has been trained, tested using XGBoost machine learning algorithm to identify on how to address the risk in this area.

In this DataCo supply chain dataset, issue with late delivery has been identified, and a model is required to determine if a specific product will reach the customer on time or will be delayed, which clearly falls under classification problem.

The objective of this paper is :

- To develop a machine learning model to predict late delivery risks, thus finding the reasons causing it.

To achieve this, XGBoost model is being built to predict the late delivery and it is evaluated based on accuracy, recall and confusion matrix to determine the performance of prediction algorithm being used in predicting late delivery risk.

## II. HYPOTHESIS

*1) Null Hypothesis:* The late delivery risk of supply chain management can be predicted accurately by the XGBoost machine learning model.

*2) Alternate Hypothesis:* XGBoost machine learning model cannot accurately predict the late delivery risk involved in supply chain management.

## III. LITERATURE REVIEW

Literature review is essential for this research since a proper study has to be done on the previous researches to done in the selection of the machine learning which is best suited for the prediction of late delivery risk. Also, this involves examination of previous researches conducted on the supply chain to have a better understanding on how the previous researches can contribute to this research.

### A. Methods optimizing delivery time in Supply chain management

By building a model of supply chain coordination, this research [2] aims to address the issue of the low rate of order on-time delivery. In the current economic climate, demand uncertainty is a highly common phenomenon that has a significant negative impact on supply chain members and lowers the effectiveness of resource allocation. By using this methodology, we can increase the order delivery on-time rate and fully utilize the information resources from downstream IT SMEs. As per [3], Manufacturers who use the Engineering-To-Order production method create, construct, and assemble a unique product in accordance with the demands of a certain client. In this kind of operation, producers are confronted with several supply chain and design process risks, which frequently result in inconsistent delivery times and high total costs. By including a buffer in the project plan, project managers can deal with lead time risks. By employing a project management method, this study's goal is to provide an optimization model for choosing the right delivery deadline for products in an ETO setting.

In this study [4], multidisciplinary research is done on highly specific business and information technology domains. Our proposed assessment methodology for supply chain management's real-time predictive delivery performance indicator was improved, and a blockchain framework called DelivChain was created to enable it. This paper [5] describes how the healthcare sector attempted to speed up delivery by using drones. Due to restricted zones, rugged terrains, war-prone places, bad road conditions, crowded traffic, and remote locations, the medical supply chain systems and delivery operations among healthcare stakeholders suffer. As a result, Healthcare 5.0 supply chains use the Internet of Drones (IoD) to streamline and accelerate the process of medical delivery through open channels. A promising technology called blockchain can manage the security and dependability of drone deliveries across dubious open channels.

The research conducted in [6]In order to investigate the ideal scenario where the cooperative delivery approach is effective in both centralized and decentralized models, the Stackelberg game model lead by the store is used. The efficiency of the cooperative delivery approach is determined by the service cost structure, according to analytical findings. In [7] it is identified that businesses are forced to compete with one another on delivery reliability and timeliness as a result of the increasingly tough rivalry. Production scheduling and logistics scheduling must be coordinated and taken into account simultaneously in order to achieve the required delivery performance at the lowest total cost.

### B. Rationale behind adapting XGBoost algorithm

According to experimental findings of [8], the XGBoost-based fraud prediction model outperforms existing machine learning models like the Logistic regression method and Gausian Naive Bayes algorithm. The XGBoost technique employs the tree's complexity function as the constant term of the objective loss function to prevent overfitting. The research performed in [9] describes that XGBoost algorithm may be quickly converted to C and installed on embedded terminal equipment. It also has some use in real-world situations. The research's findings revealed that the XGBoost classifier's accuracy is higher than that of Random Forest and SVM. As per the study conducted in [10] it is found that for the two-tier classification architecture, the XGBoost model produced the best results. The Sequential Feature Selection library was used in this study to test Forward and Backward feature removal on the XGBoost model. As a result, this research has demonstrated that XGBoost is a very effective model for predicting this significant decline, and adding more data to the training set will enable accuracy to be further increased.

As per the study conducted in [11], it is identified that the F1-score and accuracy were used to gauge algorithm effectiveness, and the median time required for each fit was used to gauge efficiency. The outcomes demonstrate that for evaluating efficacy, LGBM and XGBoost are suitable replacements for the benchmarking methods.Research performed in [12] states that in order to solve classification issues including missing values in data on the gene expression of hepatocellular carcinomas, the performance evaluation of the XGBoost approach should be examined. The XGBoost Machine Learning model can handle missing values in a dataset without imputation, yet performance evaluation on classification of Hepatocellular Carcinoma Gene Expression Data can be improved with the imputation method.

According to the research [13], In order to reduce production costs, businesses have shifted toward obtaining goods from far-off markets worldwide. The effectiveness of machine learning techniques in resolving demand forecasting issues in many contexts demonstrates its extraordinary impact on enhancing supply chain effectiveness. This may motivate stakeholders to prepare corrective actions based on supply chain and demand forecasting uses of machine learning.

## IV. METHODOLOGY

In order to analyze the data and predict late delivery risk associated with supply chain data, the methodology chosen is cross Industry Standard methodology, abbreviated as CRISP-DM as seen in Fig 1 which is the variant of CRISP-DM being built based on the objective of the project.

The delivery time of orders in supply chain management is

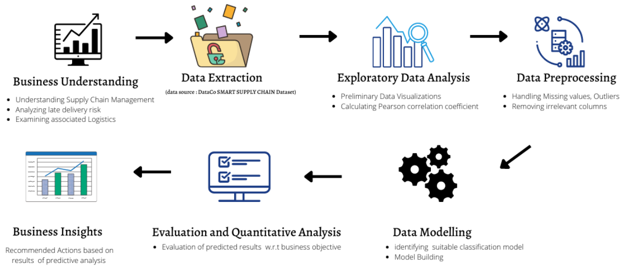

Business Understanding
Understanding Supply Chain Management
Analyzing late delivery risk
Examining associated Logistics
Data Extraction
(data source : DataCo SMART SUPPLY CHAIN Dataset)
Exploratory Data Analysis
Preliminary Data Visualizations
Calculating Pearson correlation coefficient
Data Preprocessing
Handling Missing values, Outliers
Removing irrelevant columns
Business Insights
Recommended Actions based on results of predictive analysis
Evaluation and Quantitative Analysis
Evaluation of predicted results w.r.t business objective
Data Modelling
identifying suitable classification model
Model Building

Fig. 1. Project Methology

predicted and classified as late delivery or not based on the impact of available features in the dataset, which are found to impact late delivery. XGBoost is chosen as the classification technique based on the researches discussed in the literature review section and is applied in order to predict the late delivery risk.The data being used consists of outliers, missing values which may generate biased results. Hence they are handled to build model that can generate unbiased results.

### A. Data Understanding

The dataset being used for this project consist of 180519 rows and 53 columns among which both continuous and categorical columns exist. It contains attributes that are related to supply chain management and 3 years of data beginning from January 2015 till February 2018.

### B. Data Pre-processing and Feature Selection

Data pre-processing is done to handle the missing values, outliers and irrelevant columns identified through data understanding. Columns such as Customer Zipcode, Order Zipcode and Product Description are found to have missing values. Columns Product Description and Order Zipcode are removed since they are found to have more than 80 percentage of missing values. There are 30,037 outliers identified in the dataset using boxplots which are found to be present in 9 columns which are continuous. Potential outliers which are found to have impact on the results are removed. Pearson correlation coefficient is also determined for the columns in the dataset. No duplicate record are being identified in the dataset. Product status only has values as zero and hence, this column is also removed. A total of 10 irrelevant columns such as Customer Email, Customer Fname, Customer Id, Customer Lname, Customer Password and Customer Street are identified and removed since they are found to have data which does not helps in prediction of the target column late delivery risk. Columns such as Days for shipping (real) and Days for shipment (scheduled) which are directly impacting the target column late delivery risk is also removed to avoid model overfitting scenario. Multi-collinearity between the independent columns are identified using correlation coefficient values and impacted columns are removed according to avoid multicollinearity as seen in Fig 2. As a result of pre-processing and feature selection, 25 variables were left in the dataset for model building purpose.

### C. Modelling

Extreme Gradient Boosting classification algorithm abbreviated as XGBoost is used in the model building process. This machine learning algorithm is chosen based on the observations from previous researches as mentioned in the literature review section. The pre-processed data taken is being randomly split into training and test data with 80 percent of the data moving under training dataset and the remaining 20 percent under test dataset. Standardization is carried out on the values present in the dataset. Model is being built using XGBoost classifier and the importance of the features influencing the target variable are identified. The model being built is used for the prediction purpose. Confusion matrix and ROC Curve are obtained as seen in Figures Fig. 3 and Fig. 4 respectively.

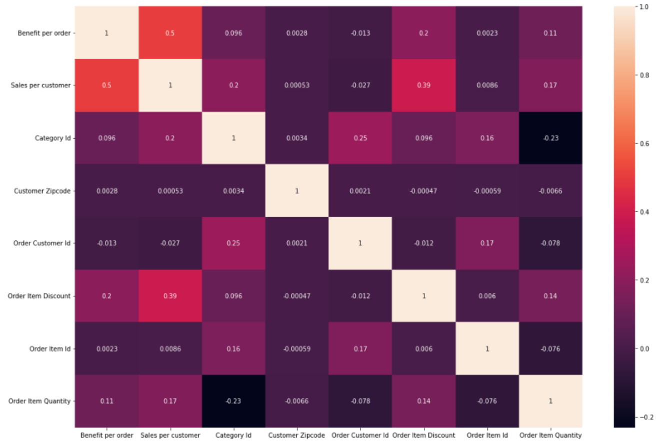


Fig. 2. Heat Map after handling multi-collinearity between independent columns

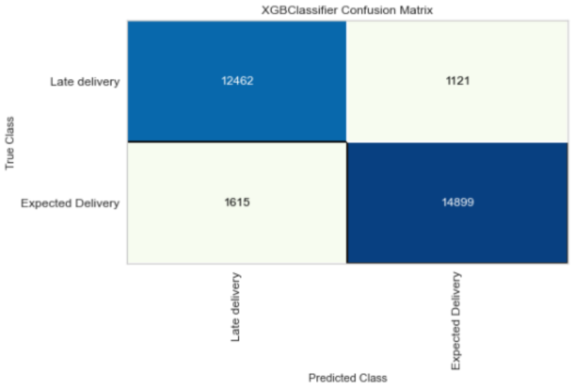


Fig. 3. Confusion Matrix

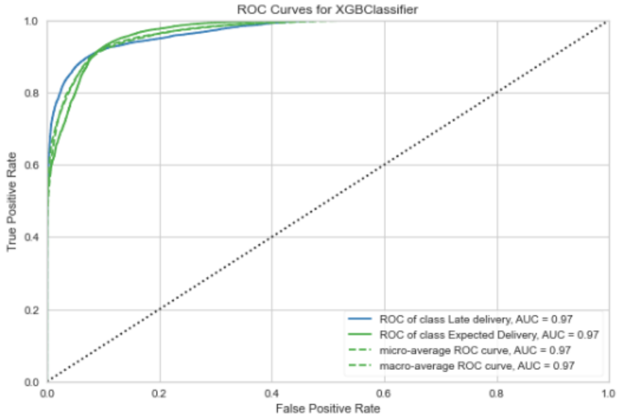
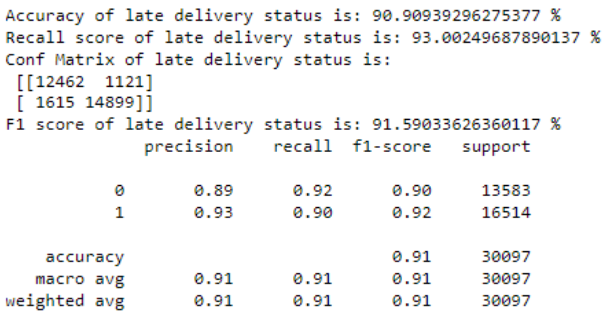


Fig. 4. ROC Curve

### D. Evaluation and Quantitative Analysis

Evaluation of the results are carried out using various evaluation metrics such as accuracy, recall score, f1 score and precision. Thus the results obtained are verified to confirm if they are unbiased. It is observed that accuracy obtained is 90.91 percentage. Six fold Cross validation is executed and it is observed that the accuracy got dropped to 89 percentage, since the data is validated with different subsets to avoid over-fitting scenarios. The recall score and F1 score are found to be 93 percentage and 91.59 percentages respectively as seen in Fig. 5

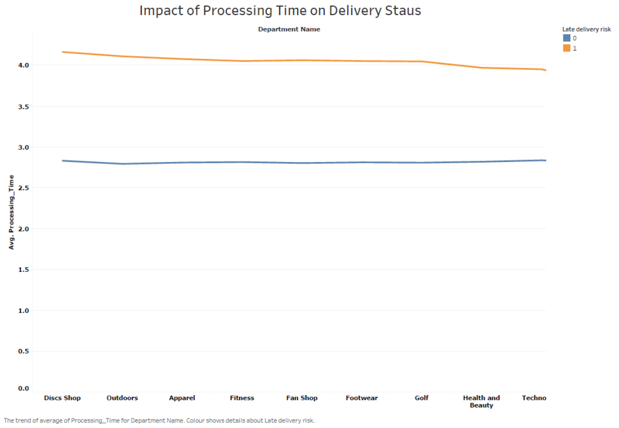


```
Accuracy of late delivery status is: 90.90939296275377 %
Recall score of late delivery status is: 93.00249687890137 %
Conf Matrix of late delivery status is:
 [[12462  1121]
 [ 1615 14899]]
F1 score of late delivery status is: 91.59033626360117 %
              precision    recall  f1-score   support

           0       0.89      0.92      0.90     13583
           1       0.93      0.90      0.92     16514

    accuracy                           0.91     30097
   macro avg       0.91      0.91      0.91     30097
weighted avg       0.91      0.91      0.91     30097
```

Fig. 5. Model Accuracy

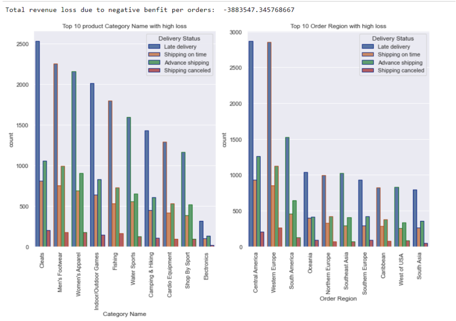
## V. Qualitative Analysis

On time shipment is a key performance indicator in supply chain business. The graph in Fig. 6 shows that all the late delivered product are processed or shipped after 4 days of product order time whereas, the products delivered on time are being processed in 3 days.



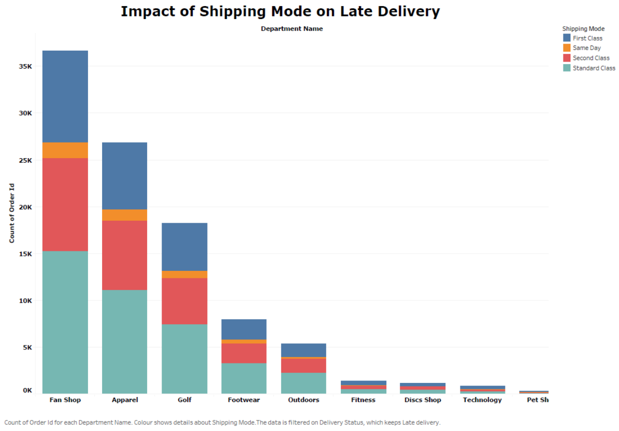
The trend of average of Processing_Time for Department Name. Colour shows details about Late delivery risk.

Fig. 6. Impact of Processing Time on Delivery Status



```
Total revenue loss due to negative benfit per orders:  -3883547.345768667
```

Fig. 7. Delivery Status



Count of Order Id for each Department Name. Colour shows details about Shipping Mode.The data is filtered on Delivery Status, which keeps Late delivery.

Fig. 8. Impact of Shipping Mode on Late Delivery

From the Fig. 7, it is observed that the products delivered late results in huge revenue loss. For the orders delivered late ,Dataco has negative revenue. Region wise and category wise distribution of products delivery time are displayed in Fig. 7. From Fig 8, it is evident the orders shipped in standard class shipping mode accounts for the majority of late deliveries followed by the second class and First class. Same-day deliveries are delivered quick. The same trend can be observed across different departments. Also, it can be seen that the fan shop has most of the late deliveries followed by apparel.

## VI. KEY BUSINESS INSIGHTS

*1) Subscription for Regular Customers:* The supply chain prediction depicts the delivery of shipment which divides the shipment mode into standard class, first class,second class and same day delivery . As per the analysis, more number of customers opts for standard class of shipment mode which eventually had higher risk of late delivery and as an organization to benefit business, implementation of customer based recommender system prompts users to subscribe first and second class mode of shipment with discounted subscription rates, thus resulting in on time delivery of the consignments and gaining revenue generation for the organization.

*2) Product categorization:* The product categorization plays a vital role in enhancing the research regarding customer ability to buy particular product . Our analysis in the project shows higher number of customers bought cleats followed by mens footwear, which illustrates customer product buying pattern. This would be helpful for business to prioritize the most demanded product in the market with high chances of availability thus assisting the supply chain manager in the organization to maintain product logistics flow without facing out of stock situations.

*3) Regional hub locator:* The main motive of the supply chain prediction analysis is to understand geographic location with maximum and minimum number of shipments carried out with in its duration. The data set extracted and analyzed for this project to predict late delivery risk justifies analysis on geographic location with high delivery risk, which would help logistics organizations to build transport hub for overcoming late delivery of shipments.

*4) Sales Prediction:* Market trend analysis have been performed through visualization and machine learning parameters. In our project the loss of revenue have been analysed with respect to the risk of late delivery which seems to be disadvantage for an organization. Such analysis could be used to predict financial depression and mitigate malfunctioning of supply chain risk management regarding late delivery of the most demanded product in the present market scenario.

*5) Customer satisfaction:* The supply chain delivery risk prediction model is a data driven risk mitigation framework focused on attaining customer satisfaction by performing data understanding , risk analysis and decision making module implementation to provide business with real time tracking of the shipment along with customer retention process in case of late delivery by providing alternate solutions such conditional offers, to maintain customer trust in the business.

## VII. CONCLUSION

Any company that engages in logistics must consider supply chain management as a key component. Orders that are delivered on time helps the company receive favorable customer feedback, which leads to satisfied customers. In this research, the XGBoost machine learning algorithm is used on supply chain management data to identify several factors that aid in the prediction of late deliveries. The anticipated outcome is discovered to have higher prediction accuracy. By addressing the issues that contribute to late delivery risks, this model can help supply chain management to perform better and gain better customer satisfaction.The accuracy of the prediction obtained through XGBoost machine learning model in predicting late delivery risk is found to be 90.90 percentage which clearly explains that the null hypothesis is satisfied.

## REFERENCES

[1] G. Baryannis, S. Dani, and G. Antoniou, "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability," *Future Generation Computer Systems*, vol. 101, pp. 993–1004, 2019.

[2] W. Cheng and X. LV, "Study over on-time delivery rate of orders executed by it manufacturing sme based on supply chain collaboration," in *2015 International Conference on Logistics, Informatics and Service Sciences (LISS)*, 2015, pp. 1–5.

[3] B. s. Wibowo, "Managing on-time delivery in engineering-to-order supply chain with buffer time optimization," in *2018 4th International Conference on Science and Technology (ICST)*, 2018, pp. 1–5.

[4] M. H. Meng and Y. Qian, "A blockchain aided metric for predictive delivery performance in supply chain management," in *2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, 2018, pp. 285–290.

[5] R. Gupta, P. Bhattacharya, S. Tanwar, N. Kumar, and S. Zeadally, "Garuda: A blockchain-based delivery scheme using drones for healthcare 5.0 applications," *IEEE Internet of Things Magazine*, vol. 4, no. 4, pp. 60–66, 2021.

[6] S. Chen, X. Wang, Y. Wu, Y. Lin, L. Li, and Q. Guo, "Equilibrium decisions of a dual channel supply chain with the cooperative delivery strategy," in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 2017, pp. 373–378.

[7] D. Liu, W. Wang, L. Huang, and D. Proverbs, "Coordinated production and delivery operations with parallel machines and multiple vehicles," *IEEE Access*, vol. 8, pp. 32 947–32 956, 2020.

[8] Y. Zhou, X. Song, and M. Zhou, "Supply chain fraud prediction based on xgboost method," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 2021, pp. 539–542.

[9] J. Zhang, Y. Li, Z. Zhang, Z. Zeng, Y. Duan, and Y. Cao, "Xgboost classifier for fault identification in low voltage neutral point ungrounded system," in *2019 IEEE Sustainable Power and Energy Conference (iSPEC)*, 2019, pp. 1767–1771.

[10] R. Bain, C. Lynch, D. McDonnell, and K. Witheephanich, "An xgboost approach for industrial component degradation classification," in *2022 33rd Irish Signals and Systems Conference (ISSC)*, 2022, pp. 1–7.

[11] H. Łoś, G. S. Mendes, D. Cordeiro, N. Grosso, H. Costa, P. Benevides, and M. Caetano, "Evaluation of xgboost and lgbm performance in tree species classification with sentinel-2 data," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 5803–5806.

[12] M. A. Latief, A. Bustamam, and T. Siswantining, "Performance evaluation xgboost in handling missing value on classification of hepatocellular carcinoma gene expression data," in *2020 4th International Conference on Informatics and Computational Sciences (ICICoS)*, 2020, pp. 1–6.

[13] A. E. Filali, E. H. Ben Lahmer, and S. E. Filali, "Exploring applications of machine learning for supply chain management," in *2021 Third International Conference on Transportation and Smart Technologies (TST)*, 2021, pp. 46–52.