ERP-Integrated Supply Chain Analysis and Risk Management: A Machine Learning Approach

Pratiksha Agarwal[1]

[1] Senior Product Marketing Manager, SAP, USA
pratikshaag86@gmail.com

Abstract. Integration and optimization of corporate activities inside the always shifting framework of supply chain management depend on Enterprise Resource Planning (ERP) technologies. Although there have been notable progress, the complexity of supply chain data makes precisely predicting and risk reduction challenging even with great advances. Late delivery is one such risk. Accurate prediction of delayed delivery would help a company's output to be much enhanced as well as the customer delight. Still, modern techniques sometimes find it difficult to understand the many linkages and patterns in the data, which reduces performance to less than ideal. This study proposes to forecast delayed delivery using the Random Forest classification model. Our approach calls for thorough data preparation, which entails activities including date conversion, date resolution of missing values, one-hot encoding for categorical variables, and MinMaxScaler application to standardize numerical features. To do complete feature selection, the study also uses feature importance from the original models and association analysis. The hyperparameters are optimized and the performance of the random forest model is improved by the grid search approach. In order to find the most appropriate tactics, the study assesses the performance of logistic regression, support vector machines, linear discriminant, and Gaussian naive Bayes among other models. With an accuracy of 99.7%, a f1-score of 99.79%, and a recall of 99.59%, the random forest model shows to be better than preceding models. With an accuracy rate of 84.98%, a recall rate of 88.06%, and an F1-score of 86.01%, the GNB model shown below-average performance.

Keywords: Enterprise Resource Planning, Supply Chain Analysis, Risk Management, Machine Learning.

1 Introduction

Customer satisfaction should be given top priority in the fast-paced corporate environment of today in order to maximize operations by means of efficient supply chain management [1]. ERP systems improve data visibility and lower operations, therefore helping supply chain management. Still, the complicated and erratic character

N. Pathak et al. (eds.), *Proceedings of the 2nd International Conference on Emerging Technologies and Sustainable Business Practices-2024 (ICETSBP 2024)*, Advances in Economics,
Business and Management Research 296,
https://doi.org/10.2991/978-94-6463-544-7_36

of supply chain data continues to be a major barrier to precisely forecasting and risk reduction, especially with regard to late delivery [2].

Various machine learning techniques have been investigated recently to handle these challenges. Because of its dependability in high-dimensional settings [3] and capacity to manage complex decision constraints, SVMs have become somewhat popular among scholars. Logistic regression (LR) has become rather common in binary classification issues thanks to its simplicity and efficiency. In many different applications, the Random Forest approach surpasses conventional ensemble techniques. This is so since the usage of numerous decision trees [4] efficiently solves overfitting and variability problems. Moreover, new neural network designs including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), together with advanced methods including gradient boosting machines (GBMs), have shown the ability to improve prediction accuracy and spot complex patterns in data. These developments notwithstanding have certain drawbacks. Research on the relationships between several modeling techniques and data inside supply networks is under progress in great volume. Moreover, even if ensemble methods such as Random Forest show promise, their whole potential has not yet been reached when combined with hyperparameter tuning and extensive feature engineering. These disparities draw attention to the requirement of a more complete supply chain risk management strategy including sophisticated machine learning algorithms and thorough data preparation.

This work presents a thorough approach for Random Forest classification model-based supply chain delivery delay forecasting. Our approach comprises of a complete data preprocessing process including converting date columns, encoding categorical variables with one-hot encoding, addressing missing values, scaling numerical features with MinMaxScaler, and carefully selecting features based on correlation analysis and feature importance from preliminary models. By means of hyperparameter modification, grid search helps to maximize the performance of the model. The results underline the need of carefully choosing and improving pertinent models created especially for the characteristics of the dataset to improve supply chain risk management's efficiency. Our individual research efforts follow:

1. Examine the DataCo Supply Chain Dataset holistically, stressing important trends and how they affect delivery performance.
2. We have transformed and ready the data for efficient modeling using cutting-edge feature engineering methods.
3. Fine-tune the Random Forest model using Grid Search, enhancing its predictive accuracy and robustness.
4. To identify the most effective approach, compare the performance of multiple models, including SVM, LR, LDA, and GNB.
5. Assessed utilizing comprehensive evaluation criteria and confusion matrices to gain a clear understanding of the strengths and limitations of each model.

This paper's organization is as follows: Following section II provides an overview of previous supply chain risk management research and highlights areas where further research is needed. Section III outlines our suggested methodology, encompassing data preparation, feature engineering, model selection, and fine-tuning. Section IV showcases the outcomes of our investigation, encompassing performance measurements and a juxtaposition of various approaches. Section V explores the consequences of our discoveries and possible avenues for future investigation. Section VI concludes the article by providing a final summary of our contributions and the significance of our work in supply chain management.

2 Related Works

Supply chain risk management has developed remarkably in recent years. Many academic publications have examined the application of statistical methods and machine learning to forecast and reduce likely risks. Machine learning has been applied by several academics to improve supply chain prediction reliability and accuracy. Using an SVM model, a research was conducted to explore supply chain disruptions and found its applicability in complex environments and unambiguous categorization of several categories [5]. In another study, delayed deliveries were predicted using logistic regression, therefore demonstrating its simplicity and relevance in binary classification environments where the connection between components and the aim variable is usually linear [6].

Over recent years, ensemble learning methods have grown in favor especially. Many companies now use a hybrid classification system [7.8]. An initial research project used a Random Forest model to project supply chain risks. Using several decision trees to lower volatility and overfitting [9] this approach exceeded individual models. Our results show that Random Forest was more able than other approaches to forecast late delivery. Furthermore, researchers looked at how GBM may improve prediction accuracy by means of repeated education of inexperienced students [10].

Along with conventional machine learning methods, deep learning techniques draw growing interest. Time-series data in supply chain environments has been examined using CNNs, well-known for their capacity to understand spatial hierarchies and patterns [11]. Especially LSTM networks, RNNs have proven capacity to capture temporal correlations and represent sequential data. In supply chain management [12], these attributes are absolutely crucial.

Moreover, we underlined the need of using anomaly detection methods to find unique data points in supply chain data and possible hazards. The work on Autoencoders for Anomaly Detection [13] proved the ability of the model to provide complete representations and detect deviations from predicted patterns. Additionally utilizing Gaussian Mixture Models (GMM), the researchers assessed anomaly detectability and classification. The authors underlined the need of GMM's probabilistic approach in faithfully reflecting complicated distributions [14].

Using several machine learning approaches is helping hybrid models to be more common. A good substitute that can maximize the benefits of both linear and ensemble approaches is using Random Forest models combined with Logistic Regression. As so, the forecast's accuracy improved considerably as well as one's capacity to understand and examine the data [15]. To increase the accuracy and flexibility of their forecasts, the researchers used Neural Networks and Support Vector Machines (SVMs [16]).

While supply chain risk management has come a long way, some areas need for more study even if it has achieved great strides. Most studies have concentrated on techniques fit for a certain model. Still, these approaches might not cover all the complex and linked elements of supply chain data. Moreover, using ensemble techniques like Random Forest shows promise. Still, adding hyperparameter optimization and careful feature engineering will help them to be even more effective and produce a lot of pertinent information. By means of intense preprocessing, feature selection, and fine-tuning techniques, our suggested solution effectively addresses the constraints. This method makes use of RF to improve prediction reliability and accuracy. This method increases efficiency and helps to better understand the causes of delayed delivery, therefore strengthening the management of supply chain vulnerabilities. It so helps to improve supply management by nature.

3 Method

3.1 Dataset

We used the DataCo Supply Chain Dataset, derived from the Mendeley Data repository [17]. There are 180,509 records in all and 53 different traits in the collection. It provides a complete picture of all aspects of supply chain operations—including orders, consumers, products, and delivery methods. When assessing several elements, it is important to include elements including the type of transaction (e.g., DEBIT, TRANSFER, or PAYMENT), the duration of shipping (both actual and scheduled), the profit per order, the sales per customer, and the delivery status, so indicating whether a shipment was on time, early, or late.A binary variable in the dataset denotes the likelihood of a delivery running late as well. It also addresses specifics on product categories, consumer locations and segments, order dates, and pricing. To look at the data, we completed many preparation chores. DateTime format was adopted from the date fields to maximize the computation of shipment durations. We also followed the required actions to correct any missing values therefore maintaining the data's accuracy. Using one-hot encoding, we transformed categorical data including "Customer Segment" and " Shipping Mode" into a numerical representation thereby streamlining the modeling process. Furthermore, numerical features were standardized with MinMaxScaler or StandardScaler to guarantee their equal influence throughout model development. The large dataset provided a strong foundation for research and prediction of logistics network disturbances. We have effectively found main elements influencing risk and delivery performance.

3.2 Proposed Work

Referred to as Algorithm 1, the proposed method consists in numerous steps including data preprocessing, feature engineering, feature selection, model initialization, fine-tuning, threshold modification, and assessment. Every phase is painstakingly created to ensure exact forecasts and complete data analysis.

Algorithm 1: Supply Chain Analysis and Risk Management
Inputs:
Supply Chain Dataset including several samples. Each sample includes transaction features and indicates late delivery status.
Outputs:
Predicted labels for late delivery risk.
Steps:

1. Import required libraries.
2. Load and preprocess the datasets.
3. **Feature Engineering:**
    - Convert the shipping date to a date format.
    - Calculate the delivery time by subtracting the order date from the shipping date.
    - Encode categorical variables.
    - Scale numerical features.
4. **Feature Selection:**
    - Analyze how features are distributed according to the class label.
    - Select features based on their correlation with the output and overall importance.
5. Split the dataset into training, testing, and validation sets.
6. **Model Initialization and Fine-Tuning:**
    - Initialize various models: RF, SVM, LR, LDA, and GNB.
    - Perform cross-validation on the training data.
    - Fine-tune the model hyperparameters using techniques like Grid Search or Random Search.
7. **Threshold Tuning on Validation Set:**
    - Set various threshold levels.
    - Predict on the validation data and compute the F2-score to select the optimal threshold.
8. **Prediction and Evaluation on Test Set:**
    - Apply the optimal threshold to the test data.
    - Evaluate the performance using the F2-score.
9. Document findings and model performance.
10. Return the predicted labels for the test set.

To start, we import the essential libraries for data processing and machine learning, such as pandas, numpy, matplotlib, seaborn, and sklearn. These libraries are indispensable at different phases of our analysis.

```
Import libraries: pandas, numpy, matplotlib, seaborn, sklearn
```

Subsequently, import and preprocess dataset D to guarantee data integrity and address any missing values. Maintaining the quality and trustworthiness of our analysis depends on this phase.

$$\text{Load and preprocess dataset } D$$

Raw data is being turned into relevant features by feature engineering that improve our algorithms' forecasting accuracy. First step is converting the "shipping date" into a DateTime format so that the "delivery time" may be computed more easily.

$$\Delta t_i = \text{shipping date}_i - \text{order date}_i \tag{1}$$

One-hot encoding is a process that converts categorical data into a numerical format that may be used with machine learning models. One-hot encoding creates binary columns for each category, ensuring that the models can reliably interpret categorical data.

$$\text{One-hot encoding: } X\_\text{encoded} = \text{OneHotEncoder} X_{\text{cate}} \tag{2}$$

Ultimately, the numerical characteristics were rescaled using the MinMaxScaler technique. Scaling ensures that all features contribute equally during model training and helps improve the convergence of gradient-based optimization algorithms.

$$X_{\text{scaled}} = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{3}$$

Feature selection is conducted to determine the most pertinent features for the classification task. Examine and depict the distribution of each characteristic according to the class label in order to comprehend their influence on the target variable. Correlation analysis and feature importance scores from preliminary models help us select the most significant features.

Analyze feature distributions by class label and relevant features.

The dataset D is split into training $((X_{\text{train}}, y_{\text{train}}))$, validation $((X_{\text{val}}, y_{\text{val}}))$, and testing $((X_{\text{test}}, y_{\text{test}}))$ sets. This split allows us to train the models, tune hyperparameters, and evaluate performance on unseen data.

$$\text{Split dataset into training, validation, and testing sets}$$

Initialized several ML models with default hyperparameters, including RF, SVM, LR, LDA, and GNB. Cross-validation on the training set helps evaluate initial performance and identify potential areas for improvement.

$$\text{Initialize models: RF, SVM, LR, LDA, GNB}$$

Fine-tuning involves hyperparameter optimization using Grid Search. Defined a range of hyperparameters for each model and searched for the optimal combination that

maximizes performance. For instance, in Random Forest, tuned the estimators. $((n_{\text{estimators}}))$, maximum depth $\left((max_{\text{depth}})\right)$, minimum samples split $\left((min_{\text{samples\_split}})\right)$, minimum samples of leaf $\left((min_{\text{samples\_leaf}})\right)$, and bootstrap.
For SVM, the kernel type (linear, polynomial, RBF), regularization parameter C, and kernel coefficient were optimized. $((\gamma))$.

$$\text{SVM hyperparameters: kernel type, } C$$

Similarly, Adjusted the regularization penalty (l1, l2), inverse of regularization strength C, and solver for Logistic Regression, the solver and shrinkage parameters for LDA, and the $(var_{\text{smoothing}})$ parameter for Gaussian Naive Bayes.

$$\begin{gathered}\text{LR hyperparameters: penalty, } C\text{,solver}\\ \text{LDA hyperparameters: solver, shrinkage}\\ \text{GNB hyperparameters: } var_{\text{smoothing}}\end{gathered}$$

Threshold tuning is performed on the validation set by setting a sequence of threshold values for class probabilities. Predicted the outcomes on the validation set for each threshold, computed the F2-score, and selected the threshold that maximized the F2-score. The F2-score is given by:

$$\text{F2-score} = \frac{5 \times \text{precision} \times \text{recall}}{4 \times \text{precision} + \text{recall}} \tag{4}$$

This step ensures that our model is calibrated to balance precision and recall effectively, minimizing false positives and negatives.

$$\text{Set thresholds and select the optimal one based on the f2 score}$$

Finally, the selected threshold for the test set is used to detect late deliveries. The model's performance is evaluated using the F2 score to ensure its robustness and accuracy.
Evaluate the efficacy of the X_test dataset by employing the most effective threshold and utilizing the F2-score.
Our proposed endeavor aims to enhance supply chain management procedures by providing a dependable and precise instrument for forecasting and reducing the risks associated with late deliveries.

3.3 Evaluation Matrix

To measure the effectiveness of our proposed models, we employ the evaluation matrix shown below. The measures are accuracy, recall, F1-score, true negatives, false positives, false negatives, and true positives. These indicators reflect the model's ability to reliably anticipate late deliveries. The confusion matrix provides a detailed analysis of the model's classification accuracy, including true negatives (TN), false positives

(FP), false negatives (FN), and true positives (TP). The confusion matrix is a visual depiction of the efficacy of a classification model. The tool shows a breakdown of accurate positive and negative classifications, as well as any inaccuracies in incorrect positive and negative classifications. Using these evaluation methods, we can thoroughly investigate the effectiveness of our models in anticipating delayed deliveries and ensure that the chosen model produces consistent and reliable findings. Feature selection is used to determine which features are most important to the classification process. To further understand how each attribute effects the target variable, we evaluate and graph its distribution relative to the class label.

4 Results and Discussion

The performance of each model is assessed using several evaluation metrics: Accuracy, F1-Score, Recall, TN, FP, FN, and TP. The results are summarized in Table 1 and visualized using confusion matrices (Figure 1). Among the models tested, Random Forest Classification achieved the highest performance across all metrics.

The Random Forest Classification model get an accuracy of 99.44%, a recall of 99.59%, and an F1-score of 99.79%. It had the lowest number of FP (123) and no FN, which indicates that it accurately identifies late deliveries while minimizing errors. The RF model's exceptional success can be attributed to its capacity to handle many features and its resilience against overfitting effectively. The ensemble learning approach of Random Forest integrates the predictions of numerous decision trees, resulting in improved accuracy and stability.

The SVM model attained an accuracy rate of 98.25%, a recall-rate of 96.93%, and an f1 score of 98.43%. Although the SVM model's performance is significantly inferior to that of the Random Forest model, it exhibits strong prediction skills. The model exhibited 941 instances of false positives and five instances of false negatives. SVMs are highly efficient in high-dimensional areas and show robust performance when there is a distinct separation between classes.

LR achieved a training accuracy of 98.25%, a recall of 96.93%, and an f1-score of 98.43%. There were 941 instances when the test incorrectly identified something as positive when it was negative and seven instances where it incorrectly identified something as negative when it was positive. LR is a straightforward and efficient model used for binary classification problems. It performs excellently when a roughly linear connection between the data and the target variable exists. The LDA model produced an training accuracy rate of 96.21%, a recall rate of 96.26%, and an F1-score of 96.56%. The model exhibited 1120 instances of false positives and 935 instances of false negatives. LDA assumes that the characteristics adhere to a Gaussian distribution and seeks to identify a linear combination of features that optimally distinguishes the classes. This model's performance is marginally inferior to the previous models, possibly because it assumes linearity and a Gaussian distribution. The GNB model achieved an accuracy rate of 84.98%, a recall rate of 88.06%, and an F1-score of 86.01%. Out of all the models assessed, this particular model had the greatest number of incorrect positive predictions (3390) and incorrect negative predictions (4743). The NB algorithm implies that the features in the dataset are independent of each other, but this assumption may not be valid in this particular dataset. As a result, the algorithm's performance may be negatively affected.

Table 1. Comparison of Classification Models.

Model	Acc	Recall	F1-score	TN	FP	FN	TP
RF	99.77	99.59	99.79	24288	123	0	29745
SVM	98.25	96.93	98.43	23470	941	5	29740
LR	98.25	96.93	98.43	23470	941	7	29738
LDA	96.21	96.26	96.56	23291	1120	935	28810
GNB	84.98	88.06	86.01	21021	3390	4743	25002

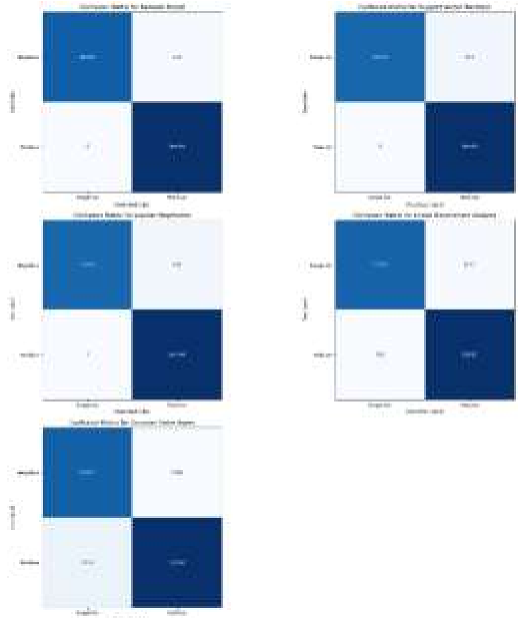

Fig. 1. Confusion Matrices for Different Models.

The confusion matrices for each model, depicted in Figure 1, present a comprehensive classification. The confusion matrix of the RF model demonstrates its exceptional performance, as it exhibits the largest count of TP and TN while minimizing false positives and avoiding false negatives.

The RF Classification model exhibited superior performance to the other models due to its capacity to effectively handle many features and its resilience against overfitting. This model employs an ensemble learning technique that amalgamates the predictions of numerous decision trees, augmenting its accuracy and stability by mitigating the variance exhibited by individual trees. RF is highly efficient in handling intricate datasets with interplay among features, a common occurrence in supply chain data. Among all the models studied, the Gaussian Naive Bayes model performed the worst. This is the outcome of a strong presumption of feature independence, often erroneous in real-world datasets where features could be related. Moreover, Gaussian Naive Bayes requires that features follow a normal distribution, which would not fairly depict the true distribution of data. Reduced accuracy, memory, and F1-score all point to more incorrect classifications brought on by these limitations.

5 Conclusion

This paper investigates and projects delayed delivery events using the Kaggle DataCo Supply Chain Dataset and additional classification techniques. Among the many subjects covered in the 180,509 records overall are goods, orders, consumers, and delivery status. of the application of the method, data preprocessing, feature engineering, feature selection, model initializing, fine-tuning, threshold adjustment, and assessment consisted of sequential phases. Date column conversion, missing value resolution, one-shot encoding of category data, and MinMaxScaler normalizing of numerical features started the paper. We investigated among different models Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), and Gaussian Naive Bayes (GNB). Every model was polished and hyperparameter optimization was achieved by means of grid search, therefore enhancing performance.The Random Forest model presented fairly amazing performance with an accuracy rate of 99.77%, a recall rate of 99.59%, and an F1-score of 99.79%. The results reveal the model's general lifetime as well as its capacity for feature interaction control. Underperformance instead came from the GNB model with an F1-score of 86.01%, 88.06% recall, and 84.98% accuracy. One can simplify this situation by considering feature independence and normal distribution. However, this assumption manifestly contradicted the characteristics of the data. The Linear Discriminant Analysis (LDA) model achieved 96.21% accuracy, while the Support Vector Machine (SVM) and Logistic Regression models achieved 98.25% accuracy. The findings underscore the need of carefully selecting and developing models that are appropriate for the dataset's specific characteristics.

Further study into the use of cutting-edge ensemble methods, complex neural network architectures, or hybrid approaches that combine many models could help to improve the accuracy and durability of forecasts in supply chain risk control.

References

1. Li, Q., Wu, G.: Erp system in the logistics information management system of supply chain enterprises. Mobile information systems 2021(1), 7423717 (2021)
2. Bialas, C., Bechtsis, D., Aivazidou, E., Achillas, C., Aidonis, D.: Digitalization of the healthcare supply chain through the adoption of enterprise resource planning (erp) systems in hospitals: an empirical study on influencing factors and cost performance. Sustainability 15(4), 3163 (2023)

3. Muthuswamy, M., Ali, A.M.: Sustainable supply chain management in the age of machine intelligence: addressing challenges, capitalizing on opportunities, and shaping the future landscape. Sustainable Machine Intelligence Journal 3, 3–1 (2023)
4. Kim, S., Seo, J., Kim, S.: Machine learning technologies in the supply chain management research of biodiesel: A review. Energies 17(6), 1316 (2024)
5. Cannas, V.G., Ciano, M.P., Saltalamacchia, M., Secchi, R.: Artificial intelligence in supply chain and operations management: a multiple case study research. International Journal of Production Research 62(9), 3333–3360 (2024)
6. Pietukhov, R., Ahtamad, M., Faraji-Niri, M., El-Said, T.: A hybrid forecasting model with logistic regression and neural networks for improving key performance indicators in supply chains. Supply Chain Analytics 4, 100041 (2023)
7. Sharma, S., Vardhan, M.: Mtjnet: Multi-task joint learning network for advancing medicinal plant and leaf classification. Knowledge-Based Systems, 112147 (2024)
8. Vardhan, M., Sharma, S.: Enhancing plant pathology with cnns: A hierarchical approach for accurate disease identification. In: Proceedings of the 2024 13th International Conference on Software and Computer Applications, pp. 159–164 (2024)
9. Ali, M.R., Nipu, S.M.A., Khan, S.A.: A decision support system for classifying supplier selection criteria using machine learning and random forest approach. Decision Analytics Journal 7, 100238 (2023)
10. Camur, M.C., Ravi, S.K., Saleh, S.: Enhancing supply chain resilience: A machine learning approach for predicting product availability dates under disruption. Expert Systems with Applications 247, 123226 (2024)
11. Hosseinnia Shavaki, F., Ebrahimi Ghahnavieh, A.: Applications of deep learning into supply chain management: a systematic literature review and a framework for future research. Artificial Intelligence Review 56(5), 4447–4489 (2023)
12. Liu, R., Vakharia, V.: Optimizing supply chain management through bo-cnn-lstm for demand forecasting and inventory management. Journal of Organizational and End User Computing (JOEUC) 36(1), 1–25 (2024)
13. Ashraf, M., Eltawil, A., Ali, I.: Disruption detection for a cognitive digital supply chain twin using hybrid deep learning. Operational Research 24(2), 1–31 (2024)
14. Huang, Z., Gou, Z.: Gaussian mixture model based pattern recognition for understanding the long-term impact of covid-19 on energy consumption of public buildings. Journal of Building Engineering 72, 106653 (2023)
15. Clavijo-Buritica, N., Triana-Sanchez, L., Escobar, J.W.: A hybrid modelling approach for resilient agri-supply network design in emerging countries: Colombian coffee supply chain. Socio-Economic Planning Sciences 85, 101431 (2023)
16. Ghalandari, M., Amirkhan, M., Amoozad-Khalili, H.: A hybrid model for robust design of sustainable closed-loop supply chain in lead-acid battery industry. Environmental Science and Pollution Research 30(1), 451–476 (2023)
17. Hasan, R., Kamal, M.M., Daowd, A., Eldabi, T., Koliousis, I., Papadopoulos, T.: Critical analysis of the impact of big data analytics on supply chain operations. Production Planning & Control 35(1), 46–70 (2024)