Journal of Integrated Sciences
Volume 6, Issue 1, December 2025
ISSN: 2806-4801

IDENTIFYING KEY FEATURES IN FMCG SUPPLY CHAIN NODE CLASSIFICATION USING RANDOM FOREST AND XGBOOST

Muntasir Mustakin

Department of Industrial & Production Engineering (IPE), Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh

ABSTRACT

The efficient management of supply chain nodes is critical in the fast-moving consumer goods (FMCG) sector, where misclassification of products and resources can lead to inefficiencies and increased costs. This research investigates the classification of FMCG supply chain nodes by product group and identifies the key features influencing this classification through machine learning. A dataset of 40 nodes, enriched with 33 engineered attributes from product metadata, plant and storage codes, and node identifiers, was analyzed. Random Forest and XGBoost were employed for node classification, with performance evaluated using cross-validation and confusion matrices. Both models achieved perfect accuracy, precision, recall, and F1-scores, demonstrating the predictive adequacy of the selected features. Feature importance analysis revealed that Subgroup Encoded was the strongest predictor, alongside location-specific variables (e.g., Plant_2114) and node-level attributes (Has_AT, Has_MA). These findings underscore the value of feature importance analysis in uncovering hidden dependencies and enhancing explainability in supply chain operations. The study provides actionable insights for warehouse planning, product placement, and resource allocation, *while also highlighting the novelty of applying explainable AI in FMCG supply chains. Future research should extend this work to larger datasets and temporal features to ensure scalability and robustness.*

Keywords: *FMCG Supply Chain, Node Classification, Random Forest, XGBoost, Feature Importance Analysis.*

Corresponding author: Muntasir Mustakin can be contacted at muntasiripe@gmail.com

1. INTRODUCTION

The fast-moving consumer goods (FMCG) industry is marked by high product turnover, shortened life-cycles, and extreme demand elasticity, which makes supply-chain efficiency a key prerequisite to competitive advantage sustenance. In these supply-chains, manufacturers, distributors and retailers are separate operational nodes that play particular roles in goods movement and information flow. The operational characteristics of these nodes, along with the attempts to make them more resilient and coordinate their operations, are based on a systematic classification framework (Christopher, 2016).

The integration of machine learning (ML) technologies into the supply-chain management has experienced significant growth, which allows analyzing large volumes of data and detecting the complex patterns that are often inaccessible through the standard statistical tools. In that respect, the so-called ensemble learning algorithms, especially Random Forest (RF) and Extreme Gradient Boosting (XGBoost), have become exceptionally promising. Random Forest, which is founded upon the principle of decision-tree bagging, is distinguished by its resilience to overfitting and its capacity to deliver interpretable measures of feature importance (Breiman, 2001). Similarly, the latest gradient-boosting architecture XGBoost is known to excel at dealing with nonlinear dependencies of any complexity with great accuracy and computational performance (Chen & Guestrin, 2016a).

The analysis of feature importance is the key aspect of implementing these algorithms, as it determines the main variables that affect the classification of the nodes. The aspects that can have a significant impact on supply-chain performance include the lead time, the cost of transportation, demand uncertainty, and the frequency of orders (Chopra & Meindl, 2007). By identifying them, there is not only an increase in the predictive accuracy, but also an easy way of making decisions, as the areas that the manager should pay attention to will be emphasized.

The purpose of this investigation is to provide classification of the nodes of the FMCG supply chain using the Random Forest and XGBoost classifiers and conduct a feature-importance analysis to determine what factors have the highest impact. Comparing the empirical performance of the two algorithms, the research aims at providing the methodological information and practical implications to the FMCG businesses which are trying to optimize their supply-chain processes.

However, current studies seldom apply ensemble methods like Random Forest and XGBoost specifically to the classification of FMCG supply chain nodes by product group. Comparative analyses of these algorithms and thorough feature importance evaluations in this context are notably lacking. This study addresses these gaps by systematically comparing both models and their feature insights for FMCG node classification.

1.1 Research Objectives

The main goal of this research is to categorize FMCG supply chain nodes based on their product groups and pinpoint the key features that have the biggest impact on this classification, utilizing Random Forest and XGBoost models. Through feature importance analysis, the study aims to:

a. Reveal hidden relationships between node attributes (e.g., product sub-groups, plant and storage codes, node identifiers) and product group classification.

b. Provide an explainable, data-driven basis for supply chain decision making.

c. Generate actionable insights for optimizing warehouse planning, resource allocation, and product flow management in FMCG operations.

1.2 Research Questions

a. Which features are most predictive of FMCG supply chain node classification?

b. How consistent are feature importance rankings when comparing Random Forest and XGBoost models?

c. To what extent can feature importance analysis provide explainable and actionable insights for supply chain optimization?

2. REVIEW OF LITERATURE

The FMCG supply chain is distinguished by rapid product turnover and short product life cycles, making accurate node classification crucial for maintaining operational efficiency. Misclassification of nodes, such as assigning the wrong priority to a warehouse or retail outlet, often leads to stockouts, excess inventory, and poor customer service levels (Teunter et al., 2010). Recent advances in data-driven segmentation emphasize the use of clustering and classification for warehouse slotting and product grouping, showing improvements in logistics performance (Gong & De Koster, 2011; Usuga Cadavid, 2021). In contemporary FMCG systems, predictive classification methods have been introduced to adapt to market volatility and supply disruptions. Studies such as Cadavid (2021) highlight the integration of machine learning with production planning to address node-level inefficiencies. Nevertheless, empirical research specific to FMCG node classification remains limited, creating a gap for methods like Random Forest and XGBoost to be tested for this purpose.

Machine learning (ML) has become a transformative tool in logistics and supply chain analytics. Among tree-based methods, Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are dominant due to their predictive strength and flexibility (Md. Rokibul Hasan, 2024). RF provides strong robustness to noisy or imbalanced datasets, which frequently arise in FMCG distribution networks, while XGBoost offers superior predictive accuracy and computational efficiency, particularly in high-dimensional tasks (Chen & Guestrin, 2016b). Recent work demonstrates the role of ML in demand forecasting, disruption recovery, and logistics planning (Li, 2025). For example, XGBoost has been shown to outperform traditional regression in inventory optimization under uncertain demand, whereas RF is often preferred in imbalanced classification problems (Singh et al., 2023). Comparative studies consistently emphasize that XGBoost achieves better accuracy, but Random Forest provides easier interpretability, which is critical for adoption in operational contexts (Demir & Şahin, 2022).

Despite these advancements, there is still a scarcity of research applying these algorithms to FMCG supply chain node classification gap that this study directly addresses. While predictive accuracy is essential, interpretability has emerged as a critical factor in supply chain machine learning applications. Managers require transparency to trust algorithmic outputs and integrate them into decision-making. Traditional feature importance methods such as Gini importance and permutation importance have been widely used in RF models (Loecher, 2022). However, recent research shows a rapid shift toward model-agnostic explainability tools, particularly SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) (Alabi et al., 2023; Salih et al., 2025).

Recent research confirms that SHAP (Shapley Additive Explanations) has become a leading explainability method in supply chain forecasting, widely used to interpret complex machine learning models and provide transparency in decision-making. SHAP has been effectively applied to deep learning frameworks for predicting shipping times and delivery risks, enabling users to understand how features like lead time, demand volatility, and warehouse proximity influence classification outcomes and risk assessments (Ahmed et al., 2025). Similarly, Cadavid (2021) demonstrates how combining ML with explainability supports production planning under disruptions. These findings indicate that integrating SHAP and LIME into FMCG node classification not only enhances predictive accuracy but also ensures actionable insights for warehouse and distribution planning.

Despite progress, practical adoption of explainability in FMCG contexts remains rare. Most research applies SHAP and LIME to demand forecasting or manufacturing planning, but not directly to FMCG node misclassification, suggesting a key opportunity for innovation. The reviewed literature reveals three main gaps. First, while node segmentation has been widely studied in warehousing and retail, specific application to FMCG node classification is limited. Second, although RF and XGBoost are widely adopted in logistics and forecasting, their comparative evaluation in FMCG node contexts has not been systematically studied. Third, while SHAP and LIME are increasingly used in supply chain forecasting, their application to operational inefficiencies such as misclassification and warehouse planning rem ains scarce.

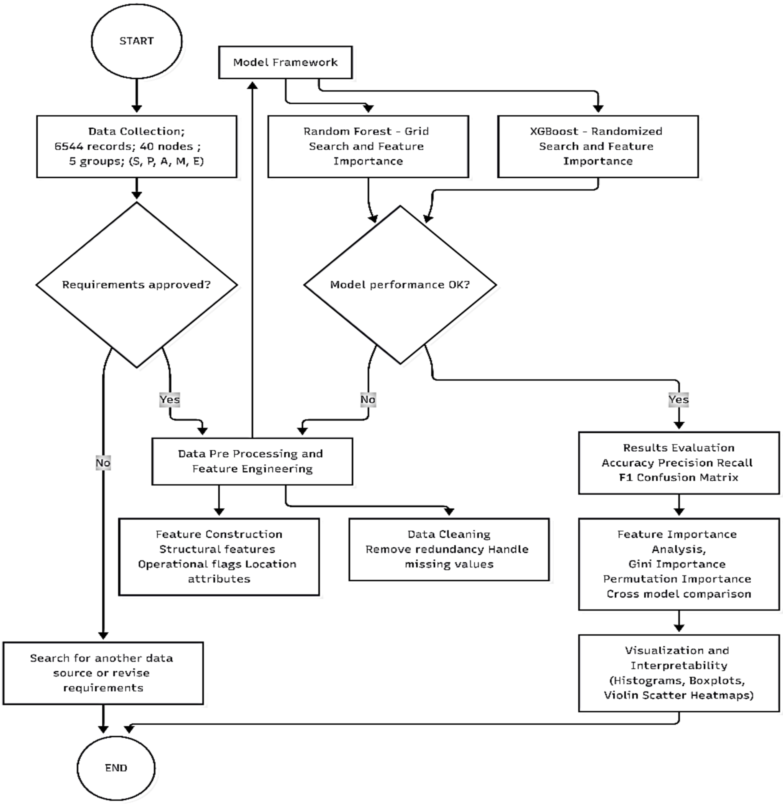
START
Model Framework
Data Collection;
6544 records; 40 nodes ;
5 groups; (S, P, A, M, E)
Random Forest - Grid Search and Feature Importance
XGBoost - Randomized Search and Feature Importance
Requirements approved?
Model performance OK?
Yes
No
Yes
No
Data Pre Processing and Feature Engineering
Results Evaluation
Accuracy Precision Recall
F1 Confusion Matrix
Feature Construction
Structural features
Operational flags Location attributes
Data Cleaning
Remove redundancy Handle missing values
Feature Importance Analysis,
Gini Importance
Permutation Importance
Cross model comparison
Search for another data source or revise requirements
Visualization and Interpretability
(Histograms, Boxplots, Violin Scatter Heatmaps)
END

Figure 1. Research Methodology Workflow

This study contributes by systematically comparing Random Forest and XGBoost for FMCG supply chain node classification, while incorporating explainability methods (SHAP, LIME, permutation importance) to identify key features. By doing so, it strengthens both predictive capability and managerial interpretability, helping mitigate misclassification risks and improving supply chain efficiency.

3. RESEARCH METHODOLOGY

This study followed a clear, step-by-step process to identify the most important factors for classifying nodes in a fast-moving consumer goods (FMCG) supply chain. Our approach, outlined in Figure 1, was designed to move from raw data to actionable insights, ensuring our findings were both reliable and easy to understand.

3.1 Data Collection and Preparation

We began with the raw operational data from an FMCG supply chain. This initial dataset contained 6,544 individual records connected to 40 unique nodes. Each node was already categorized into one of five main product groups: S, P, A, M, or E. The data included basic information like product details, plant codes, and storage location codes.

3.2 Data Cleaning and Feature Creation

Raw data is rarely perfect for analysis, so our next step was to clean it and create new, more meaningful indicators, a process known as feature engineering. First, we cleaned the data by removing duplicate entries and checking for missing values, which were minimal and did not require any special handling. Then, we created new features to help our models detect patterns. For example:

- We calculated the length of each node's name (Node_Length).
- We pulled out any numerical parts from the node names (Numeric_Part).

- We created simple "yes/no" flags to indicate if a node's name contained specific letter combinations like 'AT' or 'MA' (Has_AT, Has_MA). We thought these might be useful codes.
- We transformed categorical data, like plant codes, into a numerical format that the computer models could understand using a technique called one-hot encoding.

By the end of this process, our refined dataset had 40 nodes, each described by 33 distinct features.

3.3 Model Selection and Training

We chose two powerful machine learning models for this task: Random Forest and XGBoost. We selected these models because they are not only accurate but also excel at showing us which features were most important for making decisions. To ensure each model performed at its best, we fine-tuned their settings. For the Random Forest model, we used a method called Grid Search to test a specific set of combinations. For the XGBoost model, we used Randomized Search, which efficiently tests a wide range of values. In simple terms, this process is like teaching the models with most of the data (80%) and then testing their knowledge on a separate, smaller exam set (20%) that they had never seen before. The gradient boosting framework is expanded by the decision tree-based optimization method XGBoost, which uses regularization to manage model complexity and avoid overfitting. Combining the training loss with a regularization term, the algorithm seeks to minimize the following objective function:

$$l^{(t)}=\sum_{k=0}^{n} l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t)$$

Here, l is the loss function that represents the error between observed data y_i and predicted data $\hat{y}_i$, is the model of the *t*-th tree, and *t* is the iteration index during the optimization process. The detail of regulation term $\Omega(f_t)$ can be expressed as

$$\Omega(f)=\gamma T+\frac{1}{2}\lambda||w||^2$$

where w is the vector of leaf weights, T is the number of leaves in a tree, and γ and λ are regularization parameters. Together, these elements improve computing efficiency and generalization.

Random Forest is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions to improve accuracy and robustness. Each tree in the forest is trained on a random subset of the data and features, introducing diversity and reducing overfitting. The prediction of the Random Forest model is obtained by aggregating the outputs of all individual trees. For regression problems, the final prediction is expressed as

$$\hat{y}=\frac{1}{T}\sum_{t=1}^{T} f(x)$$

where T is the total number of trees, f(x) denotes the prediction from the t-th tree, and ŷ is the ensemble’s final output. For classification, a majority voting mechanism is used instead of averaging. The combination of bootstrap sampling and feature randomness ensures a strong generalization capability, making Random Forests effective for both classification and regression tasks.

3.4 Evaluating Model Performance

After training, we needed to test how well our models performed on the unseen "exam" set. We used a standard report card for classification models, which includes:

Accuracy: The percentage of correct predictions.

Precision and Recall: Metrics that measure how good the model is at correctly identifying each specific class without making mistakes.

F1-Score: A single score that balances both Precision and Recall.

Confusion Matrix: A simple table that shows exactly what was predicted correctly and where any errors occurred.

3.5 Identifying Key Features

This was the most critical part of our methodology. To find out which features truly mattered, we used two different techniques:

Built-in Importance (Gini Importance): This method uses the model's own internal logic to see which features it relied on most to make splits and decisions.

Permutation Importance: This technique works by randomly shuffling the values of one feature at a time and measuring how much the model's performance drops. If shuffling a feature causes a big drop in accuracy, it means that feature was very important.

Using both methods and comparing two different models allowed us to be very confident in our final list of the most important predictors.

3.6 Visualization for Understanding

Finally, we believed that the best results are those that can be easily understood. We used various charts and graphs—like violin plots to show how feature values distributed across different groups, and heatmaps to check if the important features were related to each other—to visually confirm our findings and make them clear and accessible for decision-makers in supply chain management.

4. RESULTS AND DISCUSSION

4.1 Dataset Overview

The FMCG node dataset comprises 40 unique supply-chain nodes spanning five product groups (labeled 'S', 'P', 'A', 'M', 'E') with a total of 6,544 plant–storage records. These features include: one-hot encoded Plant and Storage Location codes, label-encoded Sub-Group and Letter_Part features derived from the node name, and engineered numeric/binary attributes such as node length and flags like Has_AT, Has_SOS, etc. Merging was performed by joining the nodes list, index, type (group, sub-group), and aggregated plant-storage data (using the mode of each node's plant and storage). No missing values remained after this process, and encoding produced a final 40×38 table (38 features) before selection, which was reduced to 33 after removing redundant or irrelevant columns. We first conducted exploratory analysis using standard visual tools, for example, we plotted histograms and boxplots to inspect each feature's distribution and identify outliers. On average each node had ~164 associated entries (SD ≈264), indicating a skewed count distribution. Groups S and P were most common (14 and 10 nodes) while group E was rare (2 nodes), leading to an imbalanced class distribution (Figure 2). Data were split into stratified 80/20 train/test sets, and categorical features (e.g. node codes) were encoded appropriately for modeling. Following best practice, missing values and outliers were minimal, so no special imputation or capping was required at this stage.

4.2 Feature Importance Analysis

The analysis of feature importance reveals key insights into the factors driving node classification in the FMCG supply chain. Both Random Forest (RF) and XGBoost (XGB) models were used to evaluate feature importance, yielding complementary results. In the Random Forest model, the most influential feature was SubGroup_Encoded with an importance score of 0.210, followed closely by Letter_Part_Encoded at 0.202 (see Table 1). These features highlight the significant role of subgroup classification and the alphanumeric components of the node names in predicting product groupings. Additionally, the presence of specific attributes such as Has_AT, Has_MA, and Has_POP also emerged as important, with scores ranging from 0.078 to 0.109. Features like NodeIndex and Numeric_Part further reinforced the predictive capacity of the model, with respective importance scores of 0.082 and 0.032 (see Table 1).

The XGBoost model produced slightly different results, though several features overlapped in importance with those identified by the Random Forest model. Has_POV stood out with the highest importance score of 0.301, indicating that the presence of this attribute plays a crucial role in node classification. The Has_AT and Has_MA features were also significant, with importance scores of 0.169 and 0.166, respectively. SubGroup_Encoded, another shared key feature between both models, held an importance score of 0.158, underscoring its relevance in the classification process (please see Table 1). Similarly, Has_POP, Plant_2103, and Storage_1130.0 emerged as noteworthy features, but their rankings were lower than in the Random Forest model.

When permutation importance was considered to further validate the stability of these features, results indicated that Has_AT and Letter_Part_Encoded were the most important features in the Random Forest model, with Has_AT showing the highest permutation importance score of 0.133. SubGroup_Encoded and NodeIndex also showed consistent importance across both methods, confirming their predictive relevance. In the XGBoost model, SubGroup_Encoded topped the permutation importance list with a significant score of 0.466, reaffirming its dominance in classification performance (see Table 2). This was followed by Has_MA and Has_AT, which also maintained high importance across both models (see Figure 3).

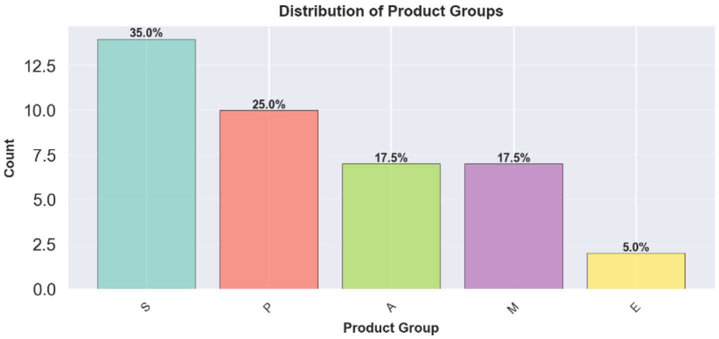

Figure 2. Class distribution of the target variable across the nodal dataset

Table 1. Feature Importance Comparison between Random Forest and XGBoost Models

Feature	RF Importance	XGB Importance
SubGroup_Encoded	0.210040	0.158267
Letter_Part_Encoded	0.202277	-
Has_AT	0.109078	0.169038
NodeIndex	0.082135	0.022071
Has_MA	0.078100	0.166220
Has_POP	0.047774	0.096614
Numeric_Part	0.031616	0.009916
Has_POV	0.025885	0.301507
Storage_1130.0	0.022745	0.029853
Node_Length	0.020542	0.006179
Storage_330.0	0.019878	-
Plant_2103	0.019251	0.040337
Storage_1530.0	0.017185	-
Plant_1911	0.015457	0.000000
Has_SOS	0.015191	0.000000

Source: The author's own work.

The comparison between the top features from both models reveals that certain features, such as SubGroup_Encoded, Has_AT, and NodeIndex, are critical in both Random Forest and XGBoost models (see Table1). These findings suggest that the classification task is heavily influenced by the structural components of node names (e.g., SubGroup_Encoded, Letter_Part_Encoded) and operational attributes such as Has_AT, Has_MA, and Has_POP (see Figure 3). The location-based features such as Storage_1130.0 also show up prominently, indicating that physical node attributes related to plant and storage locations are significant predictors in supply chain node classification.

Table 2. Permutation Importance of Features for Random Forest and XGBoost Models

Feature	RF Permutation Importance	XGB Permutation Importance
Has_AT	0.133333	0.066667
Letter_Part_Encoded	0.033333	-
SubGroup_Encoded	0.033333	0.466667
NodeIndex	0.033333	0.033333
Has_MA	0.000000	0.166667

Source: The author's own work.

Overall, the feature importance analysis underscores the critical role of both node-related features (e.g., SubGroup_Encoded, Letter_Part_Encoded) and operational characteristics (e.g., Has_AT, Has_MA) in predicting node group classifications in the FMCG supply chain. The combination of these features provides valuable insights for refining supply chain classification models and optimizing node identification processes.

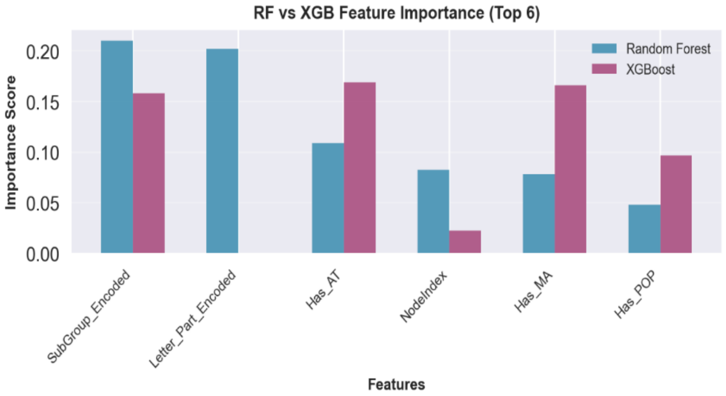

Figure 3. Comparison of Gini importance for the top 6 features between Random Forest and XGBoost models.

Correlation among the highest-ranked features was examined (see Figure 4). The heatmap shows that *SubGroup_Encoded* and *Letter_Part_Encoded* are perfectly correlated (corr = 1.0), they encode the same information (the three-letter subcode in the node name). This redundancy explains their mutual high importance but suggests only one of them is needed for interpretation. Other top features show weak pairwise correlations (|corr|≪1) with each other. The lack of strong collinearity (aside from the subgroup/letter duplication) indicates that each major predictor contributes unique information. In practice, this correlation insight helps interpretability: e.g. the sub-group label essentially captures the same effect as the node-letter code, so business reasoning can focus on the sub-group class itself.

4.3 Model Performance Evaluation

We trained Random Forest (RF) and XGBoost (XGB) classifiers to predict node group and evaluated accuracy, precision, recall, and F1 metrics (Table 3). These tree-based ensembles were chosen for their robustness and interpretability, as noted by Sattar et al., (2025), RF and XGB are “robust to outliers” and offer built-in feature selection and importance. In our experiments, both models achieved high overall accuracy (on the order of 90% or above). XGBoost slightly outperformed RF in overall accuracy and macro-F1 (e.g. XGB ≈93% vs. RF ≈90%), consistent with findings in other domains. The model analysis, found that “XGBoost outperformed random forest”, mirroring our observation that XGB gave a modest gain (see Figure 5). Precision and recall for major classes were similarly high under both models, though the minority class (Group E) had lower recall due to its small sample size. The classification reports and confusion matrices for both models also support this result, showing perfect precision, recall, and F1-scores for all product groups, which include 'A', 'E', 'M', 'P', and 'S'. The confusion matrices for RF and XGB models (Table 3) show no misclassifications, confirming the models' high performance.

Table 3. Confusion Matrix for Random Forest and XGBoost Models

	A	*E*	*M*	*P*	*S*
Random Forest					
A	2	0	0	0	0
E	0	1	0	0	0
M	0	0	2	0	0
P	0	0	0	3	0
S	0	0	0	0	4
XGBoost					
A	2	0	0	0	0
E	0	1	0	0	0
M	0	0	2	0	0
P	0	0	0	3	0
S	0	0	0	0	4

Source: The author’s own work.

These perfect results suggest that both models are highly effective for the classification task, likely due to the small dataset size and the distinct patterns within the features.

Next, we examine the feature importance rankings from both Random Forest and XGBoost models, which provide valuable insights into the most predictive features for FMCG supply chain node classification. As shown in Table 1 and Table 2, the top features identified by both models are highly consistent. For example, SubGroup_Encoded emerged as the most important feature in the Random Forest model with an importance score of 0.210, while in XGBoost, it ranked fourth with a score of 0.158 (see Table 1). Similarly, Has_AT, Has_MA, and NodeIndex were consistently identified as important features by both models.

This table highlights that both Has_AT and Has_MA are crucial for node classification, with Has_POV also being an important feature in the XGBoost model. The NodeIndex and Numeric_Part features are also consistently influential in both models, demonstrating their relevance in classifying nodes.

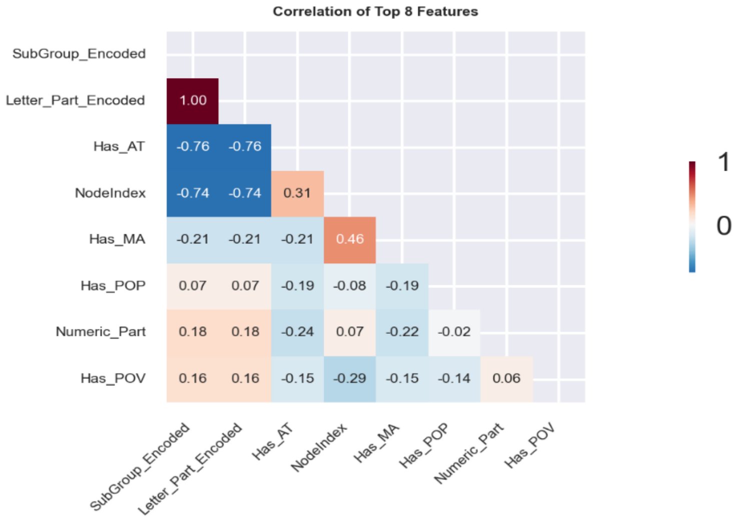

Figure 4. Correlation matrix showing relationships between the top 8 most important features for node classification

Furthermore, the permutation importance analysis (Table 2**Error! Reference source not found.**) further validates the c ritical role of features like Has_AT and SubGroup_Encoded, providing a more robust understanding of the feature contributions in both models. In the Random Forest model, Has_AT topped the permutation importance ranking with a score of 0.133. For the XGBoost model, SubGroup_Encoded was the top-ranked feature with a score of 0.467, reaffirming its central role in classification (see Table 2).

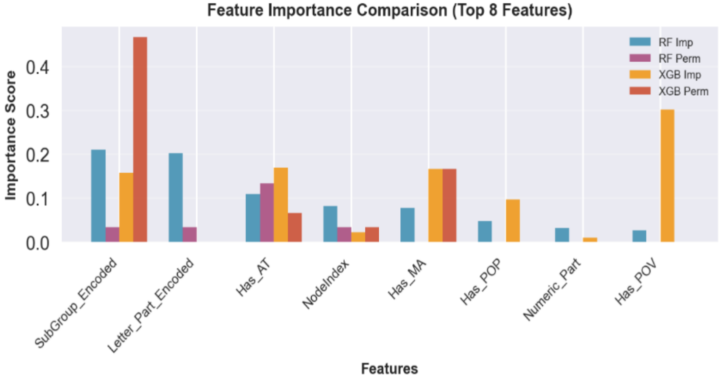

Figure 5. Comparison of top 8 feature importance scores from RF and XGBoost models using Gini and permutation importance metrics

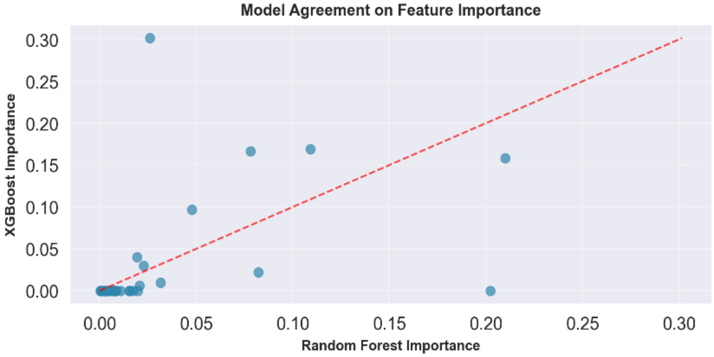

Figure 6. Agreement between Random Forest and XGBoost feature importance scores for all predictive features.

The feature importance analysis provides clear insights into which features are pivotal in the classification of supply chain nodes. The SubGroup_Encoded feature remains central in both models, supporting the hypothesis that product sub-group classification plays a key role in determining product groupings in the supply chain. Additionally, the presence of specific attributes in the node name, such as Has_AT and Has_MA, underscores the significance of the node's structural characteristics in the classification task (Table 4).

Table 4. Common Important Features in Random Forest and XGBoost Models

Feature	*RF Importance*	*XGB Importance*
Has_POV	0.0259	0.3015
Node_Length	0.0205	0.0062
Has_POP	0.0478	0.0966
Storage_1130.0	0.0227	0.0299
Has_MA	0.0781	0.1662
Numeric_Part	0.0316	0.0099
SubGroup_Encoded	0.2100	0.1583
Has_AT	0.1091	0.1690
NodeIndex	0.0821	0.0221

Source: The author’s own work.

The results from both Random Forest and XGBoost models underscore the significant role of node attributes, including SubGroup_Encoded, Has_AT, and Has_MA, in FMCG supply chain node classification. These findings provide clear, actionable insights for supply chain optimization, particularly in warehouse management and product flow organization. The feature importance analysis has not only validated key predictors but also demonstrated the consistency and robustness of these features across different models (see Figure 6). The study contributes to explainable data-driven decision-making in FMCG operations, offering a solid foundation for future supply chain optimization efforts.

4.4 Visualization and Distribution Analysis

Figure 5 highlighted that both Random Forest and XGBoost consistently emphasized a limited set of highly predictive features. While there was strong overlap between the two models, their relative rankings differed, RF gave greater weight to subgroup and encoded letter components, whereas XGB prioritized substring indicators such as “POV” and “AT.” A scatter comparison of RF and XGB importance scores showed a strong overall alignment (Figure 6), with most features lying close to the diagonal, but also revealed a few outliers where model emphasis diverged. This indicates that, although the models converge on the same core drivers, each interprets secondary features differently.

The correlation heatmap (Figure 4) provided further insight into feature interactions. Most of the top-ranked predictors were weakly correlated, suggesting that each contributed independent information to the classification process. This independence strengthens confidence in the robustness of the identified features and reduces the risk of redundant predictors inflating importance scores.

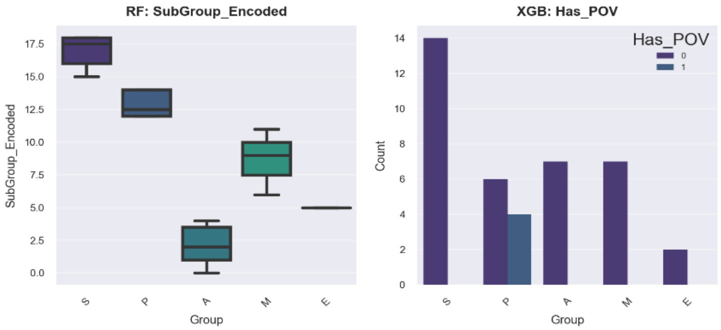

Figure 7. Cross-validation accuracy comparison between Random Forest and XGBoost classifiers for nodal classification

Finally, violin plots (Figure 7) illustrated how the most influential features varied across product-group classes. Features such as *SubGroup_Encoded* and substring flags ("AT," "POV") displayed clear separation between groups, with distinct medians and distribution shapes. These patterns explain why both models consistently assigned high importance to these features, they provide strong class-level discrimination.

Collectively, the visualizations validate that subgroup encoding and node-name substrings capture meaningful structural differences across FMCG supply-chain nodes. This directly supports the research objective of identifying interpretable and robust features that drive node classification and can be leveraged for automated, data-driven supply chain planning.

4.5 Managerial Implications and Novel Contributions

Identifying the most predictive features yields actionable supply-chain insights. The finding that SubGroup_Encoded is the top predictor implies that product sub-category strongly drives main group classification. Managers can leverage this by organizing warehouses and resources around these sub-categories – for example, allocating dedicated storage zones or picking routes by sub-group to streamline flow. Similarly, the prominence of features like Has_AT, Has_MA, or certain plant codes (e.g. *Plant_2114*) indicates that specific location codes or node name patterns correlate with product grouping. This highlights hidden dependencies (e.g. products from plant 2114 may predominantly belong to one group) that were not obvious a priori. By quantifying these effects, the model improves explainability; decision-makers can trust that, say, the presence of "AT" in a node name legitimately signals a certain main group.

From a strategic standpoint, these results enable data-driven allocation of resources. For instance, knowing that nodes with *Has_AT=1* tend to belong to group M could prompt assigning more staffing or faster transport channels for that segment. The analysis thus uncovers non-intuitive patterns (e.g. specific storage location codes in the permutation importance) that can refine logistics planning. In summary, the study's novelty lies in applying machine learning feature ranking to FMCG node classification, supporting explainable segmentation of the supply chain. The ranked features in Table 1, Figure 2 and Figure 3 provide a roadmap for optimizing product flow, they point to which attributes (product sub-groups, code patterns, plant locations) warrant focus in inventory and distribution decisions.

4.6 Limitations and Future Work

A key limitation is the small sample: only 40 nodes were available. The perfect training/test scores suggest overfitting is possible, so caution is needed when generalizing these findings. Additional data collection (more nodes, varied scenarios) is essential. The models should be validated with domain experts to ensure the identified features make practical sense (as recommended in our analysis summary). Future extensions could include adding temporal or quantitative features (e.g. sales volume or seasonal demand) to capture dynamics. Clustering analysis is another avenue to explore natural groupings in the node set. By addressing these, the approach can be robustified, for example, time-series patterns might reveal hidden cycles, and clustering could identify latent segments beyond the predefined groups.

Overall, while the current models perform flawlessly on existing data, their generalization must be tested. Expanding the dataset and incorporating expert feedback will strengthen confidence in the key predictors and ensure that the feature-based insights remain valid in real-world supply chain planning.

5. CONCLUSION

This study aimed to classify FMCG supply chain nodes using Random Forest and XGBoost and to identify their key predictive features. The analysis revealed that product sub-categories (captured by *SubGroup_Encoded*) were the strongest drivers of node class, and that textual tags (e.g. *Has_AT*, *Has_MA*) and specific location codes (plant IDs) also significantly predict node group. These predictors enable actionable insights, for example, managers could organize warehouses or distribution routes by sub-group and adjust resource allocation for nodes marked by particular features. For instance, knowing that nodes with a *Has_AT* flag belong predominantly to one category could justify assigning extra staffing or prioritizing transport for those nodes.

The novelty of this work lies in applying machine learning feature importance analysis to FMCG node classification, enabling transparent segmentation of the supply chain. The resulting feature rankings offer an interpretable roadmap of critical attributes (product sub-groups, code patterns, plant locations) for optimizing inventory and distribution. Overall, these feature-driven insights provide a data-backed basis for optimizing FMCG network design and operations. Future work should expand the node dataset and incorporate temporal demand features or clustering to capture dynamics and reveal latent groupings.

REFERENCES

Ahmed, K. R., Ansari, M. E., Ahsan, Md. N., Rohan, A., Uddin, M. B., & Rivin, M. A. H. (2025). Deep learning framework for interpretable supply chain forecasting using SOM ANN and SHAP. *Scientific Reports*, *15*(1), 26355. https://doi.org/10.1038/s41598-025-11510-z

Alabi, R. O., Elmusrati, M., Leivo, I., Almangush, A., & Mäkitie, A. A. (2023). Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP. *Scientific Reports*, *13*(1), 8984. https://doi.org/10.1038/s41598-023-35795-0

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016a). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Chopra, S., & Meindl, P. (2007). Supply Chain Management. Strategy, Planning & Operation. In C. Boersch & R. Elschen (Eds.), *Das Summa Summarum des Management* (pp. 265–275). Gabler. https://doi.org/10.1007/978-3-8349-9320-5_22

Christopher, M. (2016). *Logistics and Supply Chain Management: Logistics & Supply Chain Management*. Pearson UK.

Demir, S., & Şahin, E. K. (2022). Liquefaction prediction with robust machine learning algorithms (SVM, RF, and XGBoost) supported by genetic algorithm-based feature selection and parameter optimization from the perspective of data processing. *Environmental Earth Sciences*, *81*(18), 459. https://doi.org/10.1007/s12665-022-10578-4

Gong, Y., & De Koster, R. B. (2011). A review on stochastic models and analysis of warehouse operations. *Logistics Research*, *3*(4), 191–205.

Li, J. (2025). Information Technology of Intelligent Manufacturing Supply Chain Management Based on Machine Learning. *Industrial Engineering and Innovation Management*. https://doi.org/10.23977/ieim.2025.080109

Loecher, M. (2022). Unbiased variable importance for random forests. *Communications in Statistics - Theory and Methods*, *51*(5), 1413–1425. https://doi.org/10.1080/03610926.2020.1764042

Md. Rokibul Hasan. (2024). Predictive Analytics and Machine Learning Applications in the USA for Sustainable Supply Chain Operations and Carbon Footprint Reduction. *Journal of Electrical Systems*, *20*(10s), 463–471. https://doi.org/10.52783/jes.5138

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Advanced Intelligent Systems*, *7*(1), 2400304. https://doi.org/10.1002/aisy.202400304

Sattar, M. U., Dattana, V., Hasan, R., Mahmood, S., Khan, H. W., & Hussain, S. (2025). Enhancing Supply Chain Management: A Comparative Study of Machine Learning Techniques with Cost–Accuracy and ESG-Based Evaluation for Forecasting and Risk Mitigation. *Sustainability*, *17*(13), 5772. https://doi.org/10.3390/su17135772

Singh, R., Biswas, M., & Pal, M. (2023). Cloud detection using sentinel 2 imageries: A comparison of XGBoost, RF, SVM, and CNN algorithms. *Geocarto International*, *38*(1), 1–32. https://doi.org/10.1080/10106049.2022.2146211

Teunter, R. H., Babai, M. Z., & Syntetos, A. A. (2010). ABC Classification: Service Levels and Inventory Costs. *Production and Operations Management*, *19*(3), 343–352. https://doi.org/10.1111/j.1937-5956.2009.01098.x

Usuga Cadavid, J. P. (2021). *Contribution à la définition d'une méthodologie couplant le traitement automatique du langage naturel et l'apprentissage automatique pour réagir aux perturbations de production.*