National College of Ireland

# Supply Chain Optimization using Data Analytics: Analysis of Inventory Management, Transportation Logistics, & Procurement Processes

MSc Research Project
Artificial Intelligence

SIKHARAM SAI NAGA CHARAN
Student ID: x23141867

School of Computing
National College of Ireland

Supervisor: Prof. Muslim Jameel Syed

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

National College of Ireland

| | |
|---|---|
| **Student Name:** | SIKHARAM SAI NAGA CHARAN |
| **Student ID:** | x23141867 |
| **Programme:** | Artificial Intelligence |
| **Year:** | 2024 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Prof. Muslim Jameel Syed |
| **Submission Due Date:** | 14/08/2024 |
| **Project Title:** | Supply Chain Optimization using Data Analytics: Analysis of Inventory Management, Transportation Logistics, & Procurement Processes |
| **Word Count:** | 5918 |
| **Page Count:** | 23 |

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 bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Sikharam sai naga charan |
| **Date:** | 15th September 2024 |

## PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

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

# Supply Chain Optimization using Data Analytics: Analysis of Inventory Management, Transportation Logistics, & Procurement Processes

SIKHARAM SAI NAGA CHARAN
x23141867

**Abstract**

The supply chain is an organizational network, that includes various factors from producing to delivering a product, it is difficult to achieve customer satisfaction and operational efficacy in the current business environment. advanced machine learning techniques have shown significant results in predicting analysis among various domains, exploring these advancements in supply chain management to advance key aspects. Different algorithms are evaluated to recognize the most effective techniques for accurately forecasting on-time deliveries, the algorithms used in this research are SVM (Support Vector Machines), KNN (K-Nearest Neighbors), Logistic Regression, Extra Trees, XGBoost, and Random Forests. From results we have demonstrated that Random Forest followed by XGBoost AdaBoost achieved the highest accuracy, these models can improve prediction accuracy and decision-making for supply chain management. Additionally, an ensemble approach is used for more robust and less deviations in predictions achieved 99.13% accuracy where ensemble model uses average of all models considered the final result. Integrating advanced machine learning techniques into the current supply chain will enhance the existing model's customer satisfaction, operational efficacy, and cost reduction.

## 1 Introduction

Supply chain management has become crucial for businesses to maintain their competition in today's market world. supply chain enhancement involves key factors like manufacturing, transportation, and product delivery. All these factors are internally connected and remarkably affect the overall sustainability of the supply chain. Big data and advanced analytics have shown significant results in managing huge datasets, now integration of these models will enhance the overall performance, accuracy, and decision-making of the supply chain management (Kache and Seuring; 2017). In this study, the data analytics applications are explored to enhance supply chain management and concentrate on the enhancement of production, transportation, and product delivery. By using advanced machine learning approaches, we can explore the dependencies and complex patterns that can influence the optimization of supply management, cost reduction, and decision-making across the supply chain. With the assistance of advanced machine learning techniques, it will be easy to predict the future demand for the products, enhance resource allocation, and recognize possible errors, overall it will increase the efficiency of the supply chain (Choi et al.; 2020).

## 1.1 Motivation

In today's world, the demand for an efficient supply chain has increased, as there are key factors that increase the dynamic nature and complexity of supply chains. Businesses face multiple challenges disruption in supply, change in customer demands, and great competition in any field. Classical supply chain managements completely depend on the historical dataset and reactive strategies, often failing to recognize the limitations. However, the integration of advanced machine learning techniques and data analytics offers supply chain management to enhance predicting accuracy and mitigate delays. Several instances have proved that data analytics have transformed supply chain management capabilities. For example, (Dubey et al.; 2019) demonstrated that integration of data analytics has increased customer satisfaction and functional efficiency in supply chain management. Similarly (Wang, Gunasekaran, Ngai and Papadopoulos; 2016) found that analytics-driven supply chains will enhance the data visibility and offer more enhanced predicting decisions and production. Additionally, transportation is one of the key features of the supply chain, integrating with advanced machine learning techniques, will increase the prediction accuracy of possible delays and transportation time, which allows the business model to enhance the scheduling, assist in cost reduction, and enhance delivery timings (Min et al.; 2019). Procurement processes gain from data analytics and machine learning for its better selection of supplier (Liu et al.; 2015).

## 1.2 Research Question

This study aims to address the following questions for enhancement in supply chain management: **How are the advanced machine learning models influencing supply chain enhancement, especially in production, transportation, and procurement processes?**

## 1.3 Research Objectives

The important objectives of this research are:

- To analyze the influence of data analytics on production management and recognize important patterns that improve production management.
- To explore the advanced machine learning models, to enhance the transportation logistics, by predicting the delays and finding the optimal routes.
- To use the advantages of machine learning and data analytics for effective supply chain management.
- To compare the performance matrices of various machine learning techniques and identify the best model for accurately predicting key supply chain results.

## 1.4 Methodology

The supply chain is an operational network, it includes several key factors that make it crucial to analyze and suggest a machine-learning technique for efficient supply chain management. In this research, we have considered an extensive dataset of real-time datasets associated with production and transportation tracking. The key attributes

of this dataset are the vehicle types, transportation distance, delay indicators, booking dates, and planned and actual estimated arrival times. The main step before training the model is preprocessing the dataset, which involves cleaning, transformation, and data visualization to discover the patterns and dependencies in the dataset. Various machine learning algorithms are implemented to identify the best techniques that offer more predicting accuracy in supply chain management, mitigating delays, and enhancing on-time delivery. Machine learning techniques in this research are RF (Random Forests), XGBoost, AdaBoost, SVM, KNN, DT (Decision Trees), and Extra Trees.

### 1.5 Contribution and Impact

To train the models real-time datasets are used, this research not only highlights academic interests but also applies to real-world scenarios to improve the efficiency and robustness of supply chain management. The results from advanced machine learning models can enhance every key feature of supply chain management by enhancing inventory management and efficient procurement processes. Finally, the integration of machine learning and data analytics has shown promising results in improving supply chain management, mitigating transportation delays, and enhancement in customer satisfaction (Chen et al.; 2015).

## 2 Related Work

### 2.1 Overview of Supply Chain Optimization

Classical supply chain managements completely depend on the historical dataset and reactive strategies, often failing to recognize the uncertainties and complexity in the supply chain management data. This is a great opportunity to develop a novel methodology by integrating advanced machine learning techniques, big data, and data analytics, to enhance the overall supply chain management, address production or transportation disruptions, and mitigate on-time delivery delays performance(Christopher; 2016).

### 2.2 Data Analytics in Inventory Management

The dataset consists of real-time values, which offers great insights into the dataset, to capture complex patterns, supply chain risks, and temporal dependencies in a dataset. advanced machine learning algorithms assist in accurately predicting future demand, reducing excess production, and enhancing the reorder value (Fawcett et al.; 2014). Data analytics have shown significant results in various fields for effectively managing large data, this has shown efficiency in supply chain inventory management. For example, (Wang, Gunasekaran, Ngai and Papadopoulos; 2016) found that analytics-driven supply chains will enhance data visibility and offer more enhanced predicting decisions and production. Similarly, (Choi et al.; 2020) demonstrated that integration of data analytics has increased customer satisfaction and functional efficiency in supply chain management. These instances outline the importance of data analytics in enhancing supply chain management by reducing costs.

## 2.3 Predictive Modeling in Transportation Logistics

Transportation plays an important in enhancing the supply chain, the attributes included in transportation logistics are planning and executing the transport from the supplier to the customers. Enhancing transportation logistics can lead to enhanced delivery timings and cost savings. Classical TMS (transportation management systems) mostly depend on fixed routing and scheduling mechanisms, they can not capture the dynamic nature of evolving transportation systems, which often fail to recognize the new opportunities to enhance transportation. The enhanced machine learning and predictive modeling will assist in enhancing transportation for efficient supply chain management and reducing cost. The advanced models will be trained with more realistic and real-time datasets to get more insights into the complex patterns in the dataset (Min et al.; 2019). Predictive analytics have shown great results in various fields for enhancing the prediction strategies of models, and it has shown the same results for transportation in the supply chain. For instance, (Liu et al.; 2015) have shown that advanced machine learning models can accurately predict the transportation delays and on-time arrivals, by enhancing the routing and scheduling decisions. There are several advanced techniques to enhance the transportation of supply chain such as traffic sensors and weather predictions, the organizations can adjust the transportation timings depending the traffic and weather conditions, this will help in mitigating delays and fuel consumption. By improving the transportation mechanism, the model can reduce the risk related to vehicle breakdowns and route disruptions. In this research machine learning algorithms like linear regression, random forest, and support vector machine are used to mitigate the risks associated with transportation. Enables the model to enhance decision making regarding the routing and resource allocation, finally improvises the overall performance of transportation logistics(Min et al.; 2019).

## 2.4 Data-Driven Procurement Processes

Procurement is an important aspect of supply chain management, which includes the production and supply of the goods. It is important that procurement process is efficiently maintaining the supply chain management, there current models requires enhancement to manage costs, and guarantee the quality and compliance. The existing practices depend on historical data and manual process of managing data, which is prone to errors and time consuming. The existing models has to be integrated with real-time data into market trends and risk factors to enhance the performance of the supply chain management (Chen et al.; 2015). Many studies have proved that machine learning techniques have enhance the procurement strategies. For example, (Dubey et al.; 2019) demonstrated that data analytics and big data have increased the efficiency of supply chain mechanism by offering an extensive view of the performance metrics like quality and cost. Machine learning models helps in enhancing the procurement processes. The aim of this research is to develop a model that can enhance the procurement outcomes, reduces cost, and negotiate better terms with suppliers. Furthermore, The data analytics and machine learning improve the transparency and accountability of supply chain, contributing to more ethical procurement practices(Dubey et al.; 2019).

## 2.5 Machine Learning in Supply Chain Management

Machine Leanring (ML) techniques, therefore, have already been identified to have enhanced functionalities in SCM beyond conventional forecasting mechanisms. Analyzing data with high speed and accuracy and recognizing complex patterns which are not obvious, thus enhancing the accuracy of forecasts and subsequent decisions (Wang, Li and Zhao; 2016). In the same way, ML can help improve various facets of SCM such as demand forecasting, inventory management and risk evaluation based on past data as highlighted by (Chen et al.; 2012). The implementation of ML in SCM is still in the process of growth with various studies conducted to understand how various algorithms can be tapped to solve certain issues within the chain. Where as the Support Vector Machines (SVM) is a strong supervised learning algorithm applied to classification and regression problems. SVMs operate through creating a hyperplane in a high dimensional space so that the different classes of data could be separated with the largest margin the possibility (Cortes and Vapnik; 1995). In SCM, SVMs have been used in tasks, including demand forecasting and the detection of anomalies in various elements of the supply chain (Kumar and Zhang; 2017). The advantage of this type of algorithms is in their ability to solve equations with non-linear dependencies and high classification accuracy. However, SVMs can be computationally expensive and the tuning of the kernel function parameters can be very crucial for the performance of the model.

The k-nearest neighbors (KNN) algorithm is a form of instance-based learning, where data points are classified by majority voting of the nearest neighbour class or the mean of the nearest neighbours' value (Cover and Hart; 1967). KNN has been applied in SCM for a number of applications such as demand forecasting, and customer segmentation as noted by (Hsu et al.; 2008). The major strengths of the KNN include flexibility and simplicity during implementation as well as easy interpretation. However, KNN has drawbacks in terms of accuracy and computational performance in high dimensions and large data sets and in discriminative features. Logistic Regression is a machine learning technique, commonly used in binary classification problems to determine the likelihood of a binary event given predictor features. Logistic regression has been also used to analyze the outcomes in the context of SCM like risks and customer churns (Bureau and Rousseau; 2016). It is easy to use and interpret, as it offers straightforward assessments of the overall interactions between predictor variables and the binary dependent variable. However, logistic regression may have some issues with variables interrelations, particularly with non-linear dependencies, therefore cannot be optimal in all cases of SCM.

Extra Trees and Random Forests are two of the techniques that enhance the accuracy level of the prediction by producing multiple decision trees. In Random Forests, the decision trees are constructed multiple times and then the predictions averaged to improve the model accuracy and decrease overfitting (Breiman; 2001). Additional to Random Forests, Extra Trees modify another interaction which means that for each node in the tree, they choose the split randomly rather than choosing the split by calculating how much it will improve the model, and this improves the model complexity and precision. Both have shown success in SCM applications such as demand forecasting and inventory optimization, this could be owed to their effectiveness in large data sets and modeling complex relationships (Liaw and Wiener; 2002). XGBoost is an enhanced boosting algorithm and is getting popular due to its better performance in classification

and regression problems (Chen and Guestrin; 2016). Finally, XGBoost constructs the models sequentially where the new model tries to rectify the error that is made by the previous model. This iterative approach coupled with the use of regularization techniques make deep learning model known as XGBoost to be very efficient and accurate. In SCM, XGBoost has been successfully deployed on rather challenging forecast-related operations including demand and inventory requirements, due to its ability to operate on large-scale datasets and to identify intricate relationships between the factors under consideration (Zhang et al.; 2019).

### 2.6 Combining Model Predictions for Enhanced Accuracy

To enhance the overall prediction accuracy of the model, this research integrates multiple machine learning techniques. By utilizing the advantages of various machine learning modes and the reducing the individual models weakness, the hybrid model will offer more reliable prediction for supply chain management (Zhou; 2012). When the model integrates more machine learning techniques, it demands for more enhanced transportation schedules and procurement decisions. It is proven that averaging the results of multiple machine learning will enhance the overall performance of the supply chain management, risk assessments, and decision making strategies. For example, (Breiman; 2001) demonstrated that integrating multiple decision trees to form a random forest algorithm will assist in achieving enhanced robustness and accuracy over individual trees. Similarly, (Rokach; 2010) has proved that ensemble models will increase the performance efficiency of the model by mitigating the risks and limitations. Overall, all these findings gives the outline that integrated models will improve the models accuracy and contributes towards the sustainable supply chain management.

**Conclusion:** This literature study outlines the advantages of machine learning models, big data, data analytics, and integrated models in the supply chain management. By utilizing the advanced machine learning techniques, organizations will have great opportunity to enhance the procurement practices, transportation logistics, and production, finally reducing the costs and enhancements in operational efficacy. The combination of machine learning and data analytics into supply chain management allows more enhanced efficiency, resilience, sustainability, and decision-making. Supply chain network is growing rapidly and leads to uncertainties and complexities, The integration of machine learning and data analytics have proved to be effective in supply chain management.

## 3 Methodology

This section highlights the systematic approach to answering the research questions and accomplishing research objectives. this section provides detailed information about collecting datasets and preprocessing steps including cleaning and transforming the data, implementing machine learning techniques, and evaluating based on performance metrics. This extensive process will help the production, transportation, and procurement practices. The research aims to leverage machine learning, big data, and data analytics to encourage decision-making within the supply chain and contribute towards a sustainable management system.

## 3.1 Data Collection and Sources

The most crucial step in maintaining accuracy of the model is data collection, in this research, data is collected from real-time applications such as transport logistic tracking. The dataset which is having open access to everyone has key attributes like booking dates, estimated arrival time, delay indicators, vehicle types, relevant attributes. This dataset will allow the model to maintain accuracy and efficiency for supply chain management. Moreover, the initial dataset is combined with transportation, production, and supplier performance, to give the model a extensive dataset with a complete view of the supply chain. The data collected should be cleaned and preprocessed without any other duplicates or missing values.

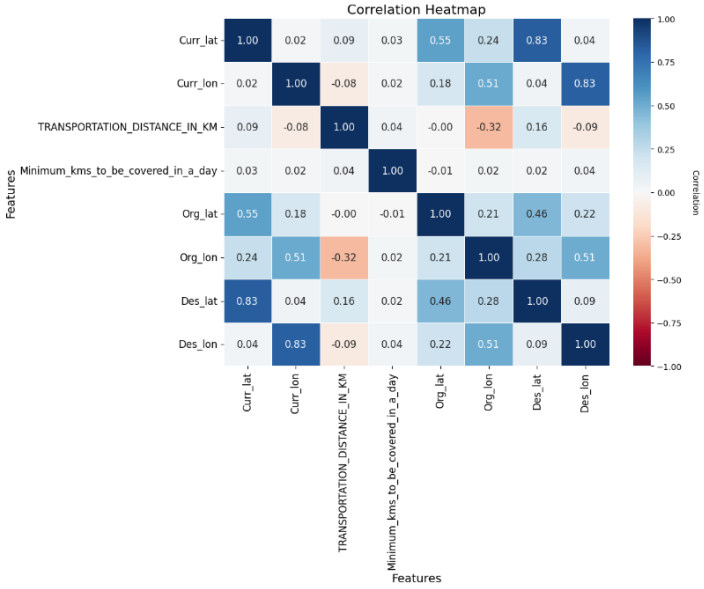



Figure 1: Correlation Heatmap of the Supply Chain Dataset

## 3.2 Data Preprocessing

Data Preprocessing step is important in assuring the model's consistency for training the dataset, In this part, the data will go through further proceedings in identifying missing values, duplicate data. These are the following steps to ensure the data quality:

- Handling Missing Values: If the data is inconsistent, then the model will be biased, reducing the prediction accuracy. If the numerical values are missing in the data, then using mean or medium imputation missing values can be added to the dataset. If categorical column values are not available, the most frequently used value will be placed, and the data will be free of missing values.

- Data Transformation: Certain columns will be merged or divided to provide a better dataset, for instance, latitude and longitude will be split into different columns and considered as different features. Date and time will be merged to form date time formats, other features like hour of the day, day of the week, and month will be identified as temporal patterns in the dataset.

- Feature Engineering: New features are created depending on the existing models to enhance the model's efficiency. For instance, based on the relative distance metrics, the time taken for delivery will be calculated. Feature engineering will remove duplicate or irrelevant variables from the dataset, to improve the model's performance.

- Encoding Categorical Variables: One-hot encoding is used to categorize the variables according to model specifications. It will be difficult to perform on categorical data, this step converts the categorical data into numerical values for better performance.

## 3.3 Predictive Modeling Techniques

This research aims to enhance the performance of the supply chain management system by providing a suitable model for accurately predicting transportation logistics and reducing costs. Several machine learning models are employed to find the best technique to enhance the supply chain. The models included in this research are:

- Logistic Regression: The prediction strategy used in logistic regression is a binary outcome, in this case, it is on-time delivery versus delayed deliveries. It is an effective model used as a reference point for several advanced models.

- Support Vector Machines (SVM): SVM is a supervised learning model, used to categorize data with hyperplanes, this can be used as binary classifiers and also as multi-class classifiers. SVM models are known for handling high-dimensional data.

- K-Nearest Neighbors (KNN): The outcome of the KNN model depends on the number of neighbors in the feature set.

- Decision Trees: Decision tree results are the average of individual leaf nodes, and these models are effective when there are a series of decision rules derived from the leaf nodes.

- Random Forests: Random Forest is an ensemble of decision trees, this model is used to mitigate the overfitting issue that occurs in the decision tree and enhance the overall performance of the model. If there is large and complex data, this model is suitable for enhancing performance.

- Extra Trees: There is a slight difference between the extra trees and random forest, in extra trees they randomly select the split points, to enhance the generalization.

- AdaBoost: It is an ensemble model, which is used to boost the model's performance by concentrating only on the misclassified data. The main aim of this technique is to enhance the model's performance.

- XGBoost: XGBoost, is an iterative model, the model's performance can be enhanced iteratively. XGBoost is known for its performance in regression and classification task.

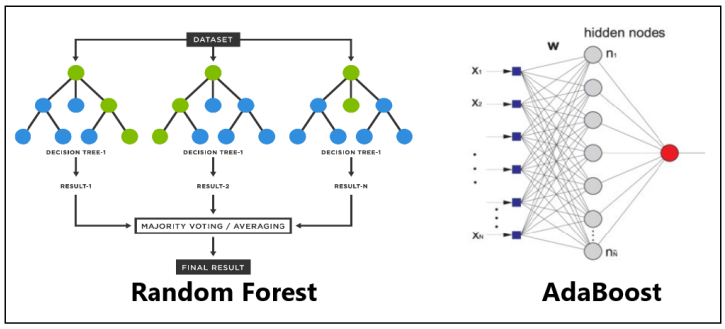



Figure 2: Architecture of Random Forest and Adaboost

## 3.4 Averaging Model Predictions

The advanced machine learning models have shown great results in accurately predicting supply chain management, but there is a scope for enhancement, to enhance the overall performance, the model started to integrate multiple models, and the average result is considered as the final output, and offer more robustness to the predictive models. The ensemble model gets the output as the weighted average from machine learning models like SVM, RF, XGBoost, and Logistic Regression. The weighted average process has to be designed carefully without biases towards specific models and enhance the overall prediction accuracy of the model. This model is used to mitigate the issue of overfitting or underfitting and the individual model's efficiency is low compared to the ensemble model.

## 3.5 Conclusion

In this section, the research methodology highlights the extensive framework for using advanced machine learning, big data, and data analytics to enhance the performance of the supply chain. This starts with collecting the real-time application dataset, preprocessing the data, and implementing various machine learning algorithms like SVM, RF, and linear regression to identify the best technique to enhance the performance of supply chain practices, performance metrics assist in evaluating the model's performance. There is a scope for enhancement, so the ensemble model integrates multiple machine learning algorithms to enhance the predicting accuracy of supply change management.

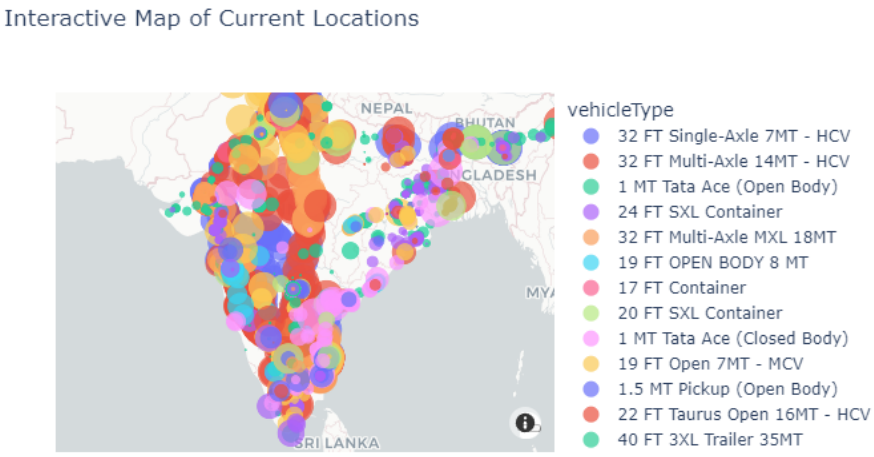


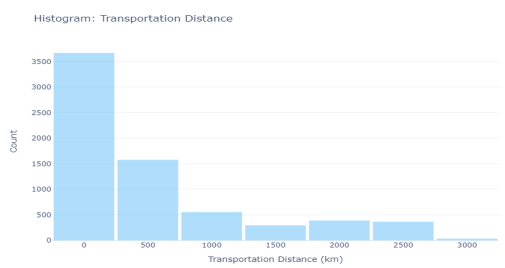

Figure 3: Supply Chain in Different Locations through Different Vehicle Types

# 4 Implementation

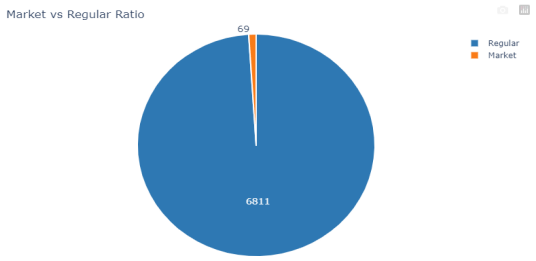
The implementation sections clearly explain the steps undertaken to enhance the prediction of supply chain management. The data collected to train this model is a real-time dataset, which helps the model predict accurately when it is applied to real-world applications. In this section, all the steps included in efficiently predicting the model are explained in detail.

## 4.1 Data Loading and Preprocessing

In the data loading phase, all the important attributes related to transportation and logistics like actual and estimated arrival timings, booking dates, vehicle types, and transportation distance are loaded carefully. These are the key attributes that influence the supply chain outcomes, so these attributes have to be acknowledged clearly. Once the data is loaded, an extensive preprocessing pipeline is created to ensure data consistency. Several important features are included in this tasks, starting from handling missing values, if the data is inconsistent, then the model will be biased, reducing the prediction accuracy. If the numerical values are missing in the data, then using mean or medium imputation missing values can be added to the dataset. If categorical column values are not available, the most frequently used value will be placed, and the data will be free of missing values.




(a) Distance Relationship of Transportation




(b) Distribution of Market and Regular Supply Orders

Figure 4: Visualisations of dataset

In Feature Engineering, New features are created depending on the existing models to enhance the model's efficiency. For instance, based on the relative distance metrics, the time taken for delivery will be calculated. Feature engineering will remove duplicate or irrelevant variables from the dataset, to improve the model's performance. In Encoding Categorical Variables, One-hot encoding is used to categorize the variables according to model specifications. It will be difficult to perform on categorical data, this step converts the categorical data into numerical values for better performance.

## 4.2 Model Development and Training

Once the dataset has gone through the data preprocessing phase it guarantees that data quality is good, consistent, and can be used for model training. Various machine-learning techniques are trained with the dataset and predict the accuracy, depending on the different parameter metrics, the model's accuracy and robustness can be decided. this process helps identify the best machine learning model for accurately predicting supply chain management.

We began with Logistic Regression, The prediction strategy used in logistic regression is a binary outcome, in this case, it is on-time delivery versus delayed deliveries. It is an effective model used as a reference point for several advanced models. Next, we implemented Support Vector Machines (SVM), it is a supervised learning model, used to categorize data with hyperplanes, this can be used as binary classifiers and also as multiclass classifiers. SVM models are known for handling high-dimensional data. K-Nearest Neighbors (KNN) is another model used to classify the data points depending on the majority class of their nearest neighbors. This model is useful when the decisions are non-linear and complex. Decision tree results are the average of individual leaf nodes, and these models are effective when there are a series of decision rules derived from the leaf nodes.

The advanced machine learning models have shown great results in accurately predicting supply chain management, but there is a scope for enhancement, to enhance the overall performance, the model started to integrate multiple models, and the average result is considered as the final output, and offer more robustness to the predictive models. This study ensembles Random Forest and Extra Trees. Random Forest is an ensemble of decision trees, this model is used to mitigate the overfitting issue that occurs in the decision tree and enhance the overall performance of the model. If there are large and complex

| Metric | Formula |
|---|---|
| True positive rate, recall | $\frac{TP}{TP+FN}$ |
| False positive rate | $\frac{FP}{FP+TN}$ |
| Precision | $\frac{TP}{TP+FP}$ |
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ |
| F-measure | $\frac{2 \cdot precision \cdot recall}{precision + recall}$ |

Figure 5: Evaluation Metrics for classification

data, this model is suitable for enhancing performance, while Extra Trees randomly select the split points, to enhance the generalization. AdaBoost is an ensemble model, used to boost the model's performance by concentrating only on the misclassified data, this will handle noisy or imbalanced data efficiently to enhance the model's performance. Finally, XGBoost is used, it is an iterative model, and the model's performance can be enhanced iteratively and mitigate the errors. XGBoost is known for its performance in regression and classification tasks.

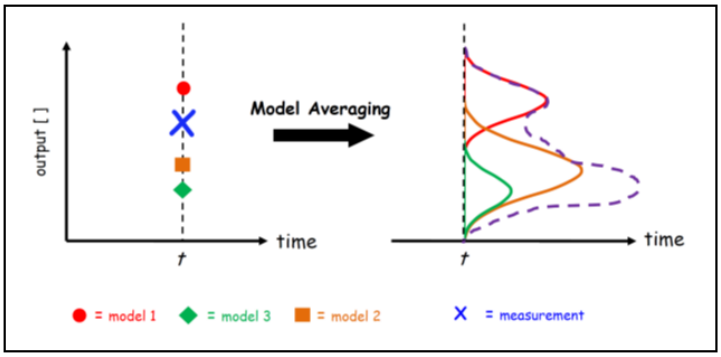



Figure 6: Architecture of Averaging the Models

### 4.3 Model Evaluation and Validation

Once the development and training of the model with an extensive dataset are completed, we move further to evaluate the performance of models depending on various performance metrics as shown in Figure 5 The average performance of all the individual models is considered as a final evaluation metric, This will reduce the overfitting of the model, and guarantees that the model will perform accurately with real-time data.

## 5 Evaluation of Implementation Results

This is the final and the most important phase of the research as it evaluates the effectiveness of machine learning models developed and trained to enhance supply chain

**Logistic Regression**

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.96 | 1.00 | 0.98 |
| Recall | 1.00 | 0.98 | 0.98 |
| F1-Score | 0.98 | 0.99 | 0.98 |
| Accuracy | | | 98.47% |

**Support Vector Machine**

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.93 | 1.00 | 0.97 |
| Recall | 1.00 | 0.96 | 0.97 |
| F1-Score | 0.96 | 0.98 | 0.97 |
| Accuracy | | | 97.23% |

**K-Nearest Neighbors (KNN)**

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.96 | 1.00 | 0.98 |
| Recall | 1.00 | 0.97 | 0.98 |
| F1-Score | 0.98 | 0.99 | 0.98 |
| Accuracy | | | 98.25% |

**Decision Tree**

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.99 | 0.99 | 0.99 |
| Recall | 0.99 | 0.99 | 0.99 |
| F1-Score | 0.99 | 0.99 | 0.99 |
| Accuracy | | | 99.20% |

**Random Forest**

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.99 | 1.00 | 0.99 |
| Recall | 1.00 | 0.99 | 1.00 |
| F1-Score | 0.99 | 1.00 | 1.00 |
| Accuracy | | | 99.64% |

**Extra Trees**

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.95 | 0.97 | 0.96 |
| Recall | 0.94 | 0.98 | 0.96 |
| F1-Score | 0.95 | 0.97 | 0.96 |
| Accuracy | | | 96.14% |

**AdaBoost**

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.98 | 1.00 | 0.99 |
| Recall | 1.00 | 0.99 | 0.99 |
| F1-Score | 0.99 | 0.99 | 0.99 |
| Accuracy | | | 99.34% |

**XGBoost**

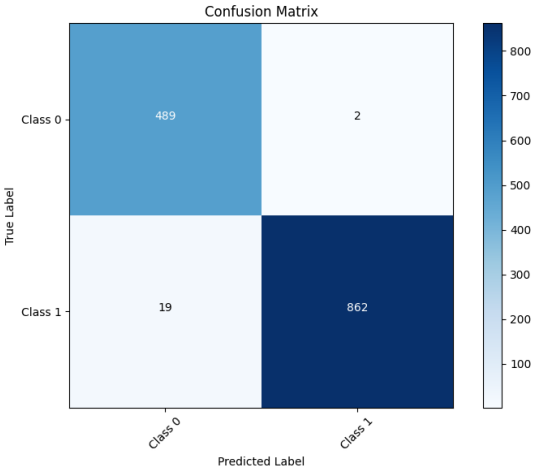
| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.99 | 1.00 | 0.99 |
| Recall | 1.00 | 0.99 | 0.99 |
| F1-Score | 0.99 | 1.00 | 0.99 |
| Accuracy | | | 99.42% |

Figure 7: Results of various models under different classification evaluation metrics

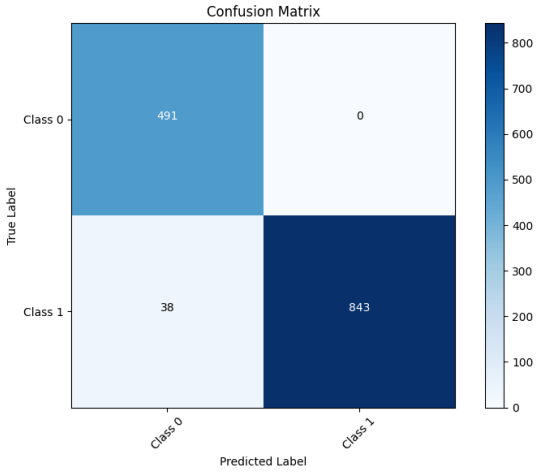
management. Every model has gone through rigorous testing and different performance metrics are used for evaluation as shown in Figure 5. We will discuss about individual performance as follows.

## 5.1 Logistic Regression

Logistic regression is an effective model, this model has shown great accuracy in predicting supply chain management, The training accuracy and model accuracy values are 99.31% and 98.47% respectively. The high accuracy value of logistic regression indicates that the model has correctly separated the large value of supply chain events. The precision values for Class 0 and Class 1 are 0.96 and 1.00, where Class 0 indicates on-time deliveries and the model predicted 96% of the deliveries will be on time and they were on-time, Class 1 indicates delays and the model predicted all the delays correctly. The recall values are high, and the values of Class 0 and Class 1 are 100% and 98% respectively, this recall values of Class 0 and Class 1 indicate they have almost correctly predicted the on-time and delayed deliveries. The F1-Score is a balance between precision and recall, the F1-Score values of Class 0 and Class 1 are 0.98 and 0.99 respectively, demonstrating the overall reliability of the model. Logistic regression has shown the highest reliability and sustainability as a baseline classifier by misclassifying only 21 values out of 1,372 predictions.




(a) Confusion Matrix of Logistic Regression



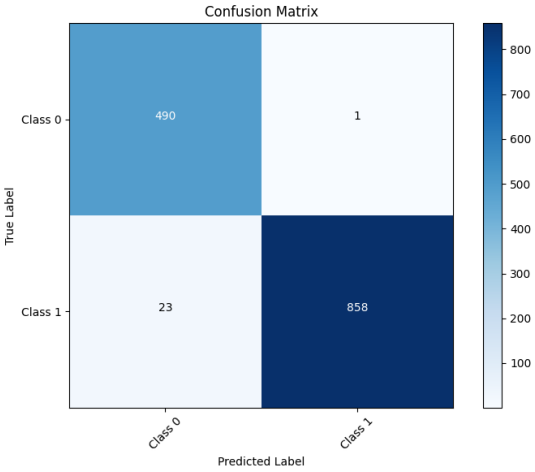

(b) Confusion Matrix of SVM

## 5.2 Support Vector Machines

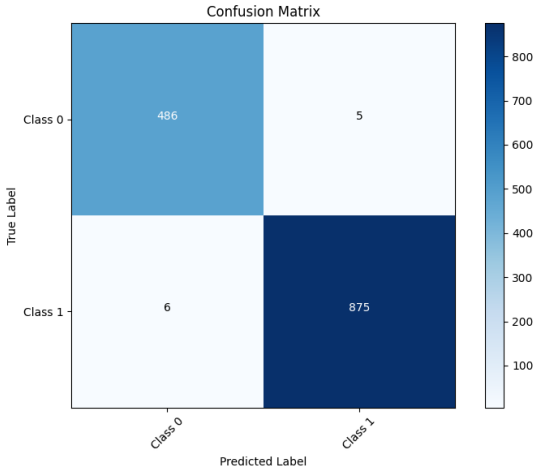
Support Vector Machines has shown great accuracy in predicting supply chain management, The training accuracy of the model is 97.23%. The precision values for Class 0 and Class 1 are 0.93 and 1.00, where Class 0 indicates on-time deliveries and the model predicted 93% of the deliveries will be on time and they were on-time, Class 1 indicates delays and the model predicted all the delays correctly. The recall values of Class 0 and Class 1 are 100% and 96% respectively, this recall value of Class 0 indicates they have correctly predicted the on-time deliveries and slightly less with delayed deliveries. The F1-Score is a balance between precision and recall and the performance is balanced. The confusion matrix demonstrated that the SVM model is highly accurate in predicting on-time deliveries, and the misclassified 38 values in Class 1 indicate that the model needs enhancements for predicting delays accurately.

## 5.3 K-Nearest Neighbors (KNN)

The KNN model performed better compared to the SVM model, The accuracy value of KNN is 98.25%. The precision values for Class 0 and Class 1 are 0.96 and 1.00, where Class 0 indicates on-time deliveries and the model predicted 96% of the deliveries will be on time and they were on-time, Class 1 indicates delays and the model predicted all the delays correctly. The recall values are high, and the values of Class 0 and Class 1 are 100% and 97% respectively, this recall values of Class 0 and Class 1 indicate they have almost correctly predicted the on-time and delayed deliveries. However, the KNN has misclassified slightly more values than the SVM model, that is 24 errors out of 1,372 predictions, this demonstrates that KNN is a robust model but not as Logistic Regression.




(a) Confusion Matrix of K-Nearest Neighbors



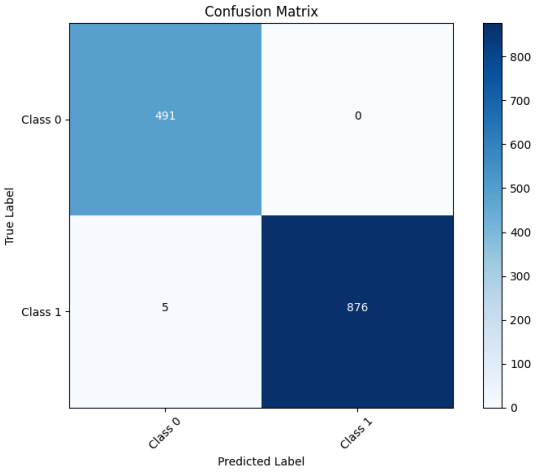

(b) Confusion Matrix of Decision Tree

## 5.4 Decision Tree

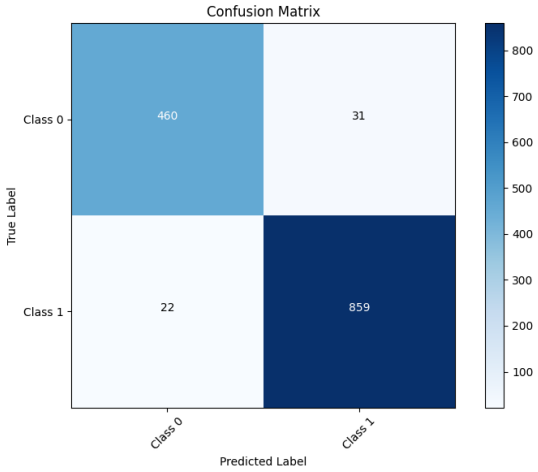
The Decision Tree model has demonstrated great performance, The training accuracy and model accuracy values are 100.00% and 99.20% respectively, so far the perfect training accuracy. The Decision Tree has achieved 0.99 for precision and recall for both Class 0 and Class 1, The F1-Score is the balance between the precision and recall, and the value of the F1-Score is 0.99 for both Class 0 and Class 1. The confusion matrix has demonstrated that the Decision Tree model has made only 11 misclassifications, This accuracy has made the Decision Tree the best performer in this study. This model can be implemented in real-time applications to enhance supply chain management.

## 5.5 Random Forest (RF)

Decision tree alone has predicted with great accuracy, Random forest is an ensemble of multiple decision trees, so the accuracy of this model will be high compared to other models. The test accuracy of Random Forest is 99.64%. Precision, recall, and F1-Score of Class 0 and Class 1 are the same, the on-time delivery value is 0.99 and the delay delivery value is 1.00. The RF model confusion matrix indicates that the model has made only 5 misclassifications out of 1,372 predictions. This model can handle large data and complex interactions within the data. RF can perform accurately on new or unseen datasets of supply chain management.




(a) Confusion Matrix of Random Forest



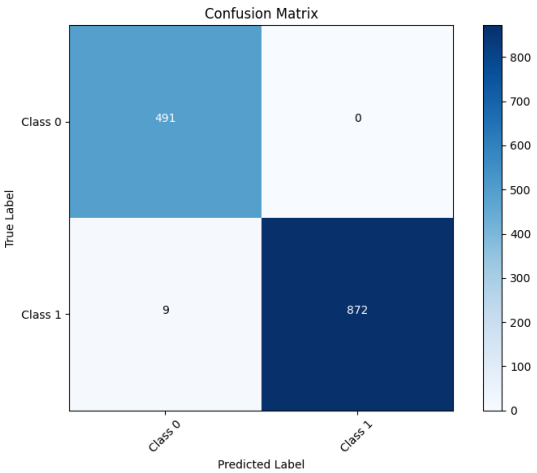

(b) Confusion Matrix of Extra Trees

## 5.6 Extra Trees

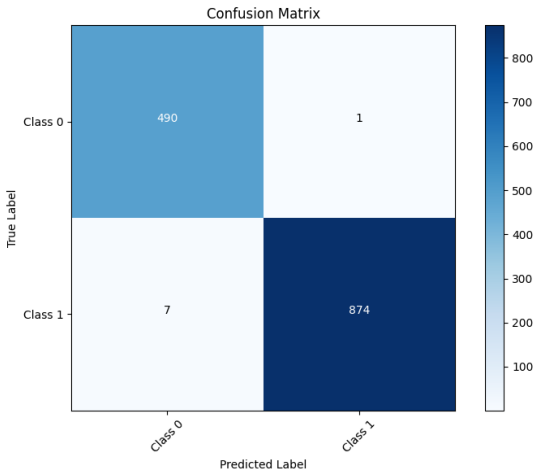
The training accuracy of the Extra Trees is 100.0%, But the lower model accuracy is slightly low compared to the training accuracy which is 96.14%. The F1-score of Class 0 is 0.95 and Class 1 is 0.97. The confusion matrix shows that the model struggled to predict on-time deliveries perfectly. Out of 1,372 predictions, 31 predictions are misclassified. Extra Trees may be powerful but have struggled with overfitting the training data.

## 5.7 AdaBoost

The training accuracy of the AdaBoost model is 99.91%, which is exceptionally good, and the model accuracy score of the AdaBoost model is 99.34%. The precision for Class 0 and Class 1 is high with0.98 and 1.00 respectively, with F1-scores of on-time and delayed deliveries is 0.99. The confusion matrix explains that the model is almost perfectly classified, with only 9 misclassifications in Class 1. AdaBoost's ability to recurrently concentrate on misclassified predictions contributed to its great performance.




(a) Confusion Matrix of Adaboost




(b) Confusion Matrix of XGBoost

## 5.8 XGBoost

The test accuracy score achieved by XGBoost is 99.42%. The precision, recall, and F1-Scores of Class 0 and Class 1 are 0.99 and 1.00 respectively. The confusion matrix demonstrated that the XGBoost model has performed exceptionally and made only 8 misclassifications, with only 1 misclassification in Class 0 and 7 misclassifications in Class 1. XGBoost model recurrently enhances the model performance and makes it as one of the top performers in this research.

## 5.9 Comparison of Models

The following table explains the performance of different models depending on their training and model accuracy scores:

Table 1: Model Training and Accuracy Scores

| **Model** | **Training Accuracy Score** | **Test Accuracy Score** |
|---|---|---|
| Random Forest | 100.00% | 99.64% |
| XGBoost | 100.00% | 99.42% |
| AdaBoost | 99.91% | 99.34% |
| Decision Tree | 100.00% | 99.20% |
| Logistic Regression | 99.31% | 98.47% |
| K-Nearest Neighbors (KNN) | 99.22% | 98.25% |
| Support Vector Machines | 97.76% | 97.23% |
| Extra Trees | 100.00% | 96.14% |

From the above table, Random Forest has outperformed all other models with the highest model accuracy of 99.64%. XGBoost and AdaBoost have performed exceptionally well compared to most of the machine learning models, with accuracy scores of 99.42% for XGBoost and 99.34% for AdaBoost. Decision Tree has a perfect training accuracy of 100% but a slightly lower model accuracy value of 99.20%. Logistic Regression, SVM, and KNN have shown great performance as baseline classifiers, with accuracy scores between 97.23% to 98.47%. The Extra Trees model was effective and had the lowest accuracy score of 96.14%.

| Metric | Class 0 | Class 1 | Overall |
|---|---|---|---|
| Precision | 0.98 | 1.00 | 0.99 |
| Recall | 1.00 | 0.99 | 0.99 |
| F1-Score | 0.99 | 0.99 | 0.99 |
| Accuracy | | | 99.13% |
| Confusion Matrix | 491/0 | 12/869 | |

Figure 12: Classification Metrics and Confusion Matrix of Averaging Model

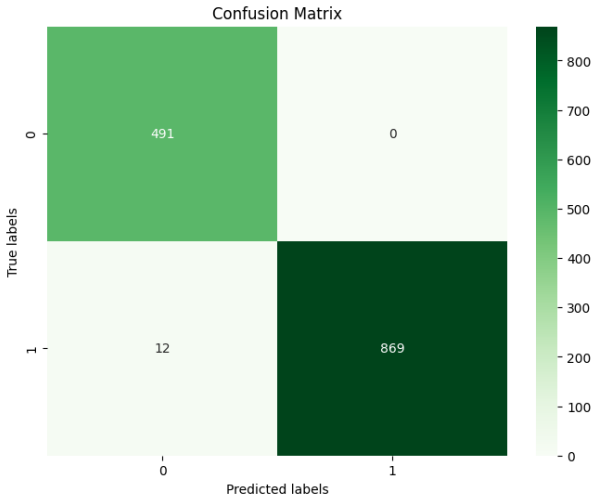



Figure 13: Confusion Matrix of Averaging Model

### 5.10 Averaging Model Predictions

An ensemble model is implemented but integrating multiple machine learning algorithms and the average of their predictions is considered as a final result of the ensemble model. The accuracy of the ensemble model is 99.13% as shown in figure 12 which is slightly less than a few individual models, but it will offer robust predictions gradually. The confusion matrix demonstrated that the ensemble model has made 12 misclassifications which are slightly greater than a few individual models, the errors in Class 0 and Class 1 are 0 and 12 respectively. The ensemble model has utilized the strengths and reduced the weaknesses of the individual models, contributing towards a reliable tool for supply chain management.

## 6 Discussion of Results

The results reported in above section of this research study provide significant insights into the effectiveness of using predictive modeling to optimize supply chain operations. We will discuss strengths and weakness with implications in this section.

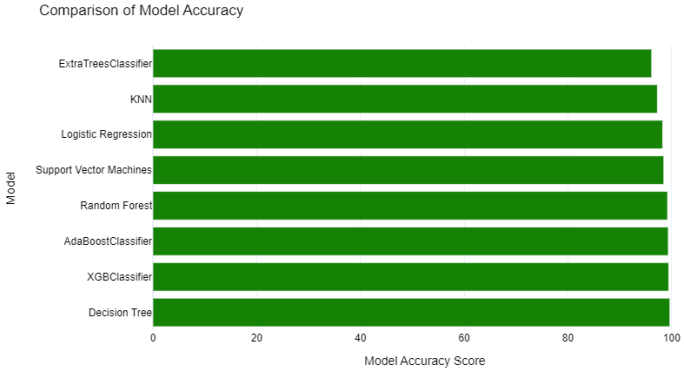



Figure 14: Accuracies of the Predictive Models

## 6.1 Strengths and Implications of the Predictive Models

The high accuracy scores obtained by machine learning models such as Random Forest, XGBoost, and AdaBoost represents the potential of machine learning in supply chain optimization. These models consistently delivered the superior performance across the key metrics, including precision, recall, and F1-score, which are crucial for making reliable predictions in real-world supply chain scenarios.

**Random Forest** model perfromed significantly better than all the models with the accuracy of 99.64% which represents the potential ability to handle complex datasets and also represents its effectiveness in situations where the supply chain data is highly complex. **XGBoost** and **AdaBoost** models able to perform better with the 99.42% and 99.34% accuracies respectively where this effectiveness can be attributed to their efficient and effective boosting algorithms as they iteratively updates on previous models. This iterative improvement process is very important for supply chain sector because of the minor enhancements in performance results in the significant cost and efficiency gains. The high accurate predictions for these models help in reliability for predicting both on-time deliveries with minimal delays.

## 6.2 Limitations and Challenges

While the overall performance of the models was strong, there were some limitations and challenges for few models such that models like SVM, KNN, Extra tress models significant misclassifications because they struggle to make appropriate decision boundary between classes is not clear because complex supply chain data with vast data its very complex to find clear decision boundary.

**K-Nearest Neighbors (KNN)**, despite its relatively high accuracy score of the 98.25%, also faced challenges, particularly with a slightly higher number of misclassifications compared to Logistic Regression. This could be due to KNN's sensitivity to the choice of neighbors (k) and the distribution of the data, which can lead to inconsistencies in predictions, especially in datasets with overlapping class boundaries. **Extra Trees**, despite its perfect training accuracy, had a lower model accuracy score of 96.14%, which indicates possible overfitting to the training data. This overfitting suggests that while Extra Trees is powerful in capturing detailed patterns within the training set, it may not generalize as well to new, unseen data. This limitation is particularly important in supply chain applications, where models need to be robust and generalizable across different scenarios.

## 6.3 Impact of Averaging Model Predictions

To deal with limitations of individual models we used ensemble based strategy for averaging predictions which shown effectiveness of combining multiple models. The averaged model able to get 99.13% accuracy which is slightly lower than some individual models but our goal for average model is not the best predictions but predictions with more consistent and less standard deviation because for supply chain logistics predictions with high variances results in huge losses. Supply chain managers can use averaged predictions over multiple models for better decision making to reduce the probability of error that may leads to costly disruptions or inefficiencies.

# 7 Conclusion and Future Work

## 7.1 Conclusion

This study aimed to find out the application of predictive modeling using machine learning techniques to minimize supply chain operations, specifically focusing on inventory management, procurement processes. By studying various machine learning models, this research has figured out the efficient approaches for accurately predicting key supply chain outcomes, such as exact time deliveries and delays. Illustrations from the results were that various ensemble models namely as Random Forest, XGBoost,AdaBoost stood higher than other models, reaching the highest accuracy scores thus providing robust predictions. These models proved particularly effective in maintaining complex supply chain data, dealing with valuable insights which actually increase decision-making and improve overall operational efficiency. The study enlightens the fruitfulness of leveraging model predictions, which further enhanced the stability as well as accuracy of the predictions. This also elevated the weaknesses of individual models and provided a more comprehensive predictive framework, promising more consistency as well as reliablility in the outcomes.

## 7.2 Future Work

Though this specific study has ensured significant insights into the predictive modeling for supply chain optimization, there are also several deficit areas where future research rely on these findings to further improve the effectiveness and applicability of these models.

1. **Expansion to Other Supply Chain Domains:** Future research should also go into the deeper roots for the application of the developed models in different contexts of supply chain, such as manufacturing, retail, or healthcare. Each of these individual industries has unique supply chain dynamics, and testing the models in varied environments can provide additional insights into their versatility and effectiveness. Adding on, management of perishable goods and also catching up the highly variable demand are the current challenges in supply chain which can be of immense help in providing opportunities for new models to be explored.

2. **Integration with Emerging Technologies:** The combination of predictive models with new emerging technologies such as Internet of Things (IoT), blockchain, artificial intelligence (AI), machine learning (ML) offers promising areas for future work. IoT devices, for example, provides real-time data that could further have chances for refining the models, which makes them even more responsive to emerging supply chain conditions. Blockchain technology can enhance the clearness of transactions, making the predictive models to be used in more subtle and complex environments.

This study highlights the possibility of potential directions for future research. By continuing exploration and refinement in predictive modeling techniques, there are chances to further enhance the efficiency, resilience, sustainability of supply chain operations in various industries. The combination of these and new emerging technologies with their application in real-world environments aids in unlocking their new application areas and driving them towards innovation in supply chain management.

# References


Abbas, K., Afaq, M., Ahmed Khan, T. and Song, W.-C. (2020). A blockchain and machine learning-based drug supply chain management and recommendation system for smart pharmaceutical industry, *Electronics* **9**(5): 852.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Bureau, J.-C. and Rousseau, D. (2016). Logistic regression for supply chain risk management, *Journal of Business Logistics* **37**(2): 146–161.

Chen, H., Chiang, R. H. L. and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact, *MIS Quarterly* **36**(4): 1165–1188.

Chen, H., Preston, D. S. and Swink, M. (2015). The impact of supply chain analytics on operational performance: A resource-based view, *Journal of Business Logistics* **36**(2): 120–132.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Choi, T.-M., Chan, H.-K. and Yue, X. (2020). Recent development in big data analytics for business operations and risk management, *IEEE Transactions on Cybernetics* **50**(1): 1–13.

Christopher, M. (2016). *Logistics & Supply Chain Management*, 5th edn, Pearson.

Cortes, C. and Vapnik, V. (1995). Support-vector networks, *Machine Learning* **20**(3): 273–297.

Cover, T. and Hart, P. (1967). Nearest-neighbor pattern classification, *IEEE Transactions on Information Theory* **13**(1): 21–27.

Cox, D. R. (1958). The regression analysis of binary sequences, *Journal of the Royal Statistical Society: Series B (Methodological)* **20**(2): 215–242.

Dietterich, T. G. (2000). Ensemble methods in machine learning, *Multiple Classifier Systems*, Springer, pp. 1–15.

Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M. and Foropon, C. (2019). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organizations, *International Journal of Production Economics* **221**: 107–123.

Fawcett, S. E., Ellram, L. M. and Ogden, J. A. (2014). *Supply Chain Management: From Vision to Implementation*, Pearson Education.

Geurts, P., Ernst, D. and Wehenkel, L. (2006). Extremely randomized trees, *Machine Learning* **63**(1): 3–42.

Ghazal, T. and Alzoubi, H. (2021). Modelling supply chain information collaboration empowered with machine learning technique, *Intelligent Automation & Soft Computing* **29**(3): 243–257.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. (2008). A practical guide to support vector classification, *Technical report*, Department of Computer Science and Information Engineering, National Taiwan University.

Islam, S. and Amin, S. H. (2020). Prediction of probable backorder scenarios in the supply chain using distributed random forest and gradient boosting machine learning techniques, *Journal of Big Data* **7**(1): 65.

Kache, F. and Seuring, S. (2017). Challenges and opportunities of digital information at the intersection of big data analytics and supply chain management, *International Journal of Operations & Production Management* **37**(1): 10–36.

Kamble, S. S., Gunasekaran, A., Kumar, V., Belhadi, A. and Foropon, C. (2021). A machine learning based approach for predicting blockchain adoption in supply chain, *Technological Forecasting and Social Change* **163**: 120465.

Kumar, I., Rawat, J., Mohd, N. and Husain, S. (2021). Opportunities of artificial intelligence and machine learning in the food industry, *Journal of Food Quality* **2021**(1): 4535567.

Kumar, S. and Zhang, S. (2017). Machine learning in supply chain management: A review, *International Journal of Production Economics* **193**: 20–34.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest, *R News* **2**(3): 18–22.

Liu, W., Zhao, X. and Tang, O. (2015). Dynamic pricing and ordering decision for the perishable food of the supermarket using rfid technology, *Asia Pacific Journal of Marketing and Logistics* **27**(3): 461–477.

Mentzer, J. T., Stank, T. P. and Esper, T. L. (2001). Supply chain management: Theory and practice, *Journal of Business Logistics* **22**(2): 1–25.

Min, H., Zacharia, Z. G. and Smith, C. D. (2019). Defining supply chain analytics and its impact on supply chain management, *Journal of Business Logistics* **40**(1): 33–49.

Polikar, R. (2006). Ensemble based systems in decision making, *IEEE Circuits and Systems Magazine* **6**(3): 21–45.

Pournader, M., Ghaderi, H., Hassanzadegan, A. and Fahimnia, B. (2021). Artificial intelligence applications in supply chain management, *International Journal of Production Economics* **241**: 108250.

Rai, R., Tiwari, M. K., Ivanov, D. and Dolgui, A. (2021). Machine learning in manufacturing and industry 4.0 applications.

Rokach, L. (2010). Ensemble-based classifiers, *Artificial Intelligence Review* **33**(1-2): 1–39.

Sajawal, M., Usman, S., Alshaikh, H. S., Hayat, A. and Ashraf, M. U. (2022). A predictive analysis of retail sales fore-casting using machine learning techniques, *Lahore Garrison University Research Journal of Computer Science and Information Technology* **6**(4): 33–45.

Seyedan, M. and Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities, *Journal of Big Data* **7**(1): 53.

Sharma, R., Kamble, S. S., Gunasekaran, A., Kumar, V. and Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance, *Computers & Operations Research* **119**: 104926.

Wang, G., Gunasekaran, A., Ngai, E. W. T. and Papadopoulos, T. (2016). Big data analytics in logistics and supply chain management: Certain investigations for research and applications, *International Journal of Production Economics* **176**: 98–110.

Wang, J., Li, Y. and Zhao, L. (2016). Applying machine learning techniques for predicting supply chain risks, *Journal of Business Research* **69**(11): 5164–5171.

Zhang, L., Wang, J. and Liu, J. (2019). Forecasting demand in supply chain using xgboost, *International Journal of Production Economics* **208**: 137–146.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*, Chapman and Hall/CRC.