A Comparative study of classification Algorithms for shipment Delay Forecasting

[1]V Sumalatha, [2] Tempalli Sriapoorva, [3] Madhugani Harshika, [4] Sunkari Sidharatha
[1]AssistantProfessor, [234]Students
Department of Computer Science & Engineering
Siddhartha Institute of Technology & Sciences, Narapally
cse.hod@siddhartha.co.in, 23TQ1A0516@siddhartha.co.in, 23TQ1A0557@siddhartha.co.in, 23TQ1A0545@siddhartha.co.in,

Abstract

Shipment delay prediction plays a crucial role in logistics and supply chain management, as delays can lead to increased operational costs, poor customer satisfaction, and inefficient resource utilization. This study focuses on developing a predictive system to forecast shipment delays using machine learning classification algorithms. The main objective is to identify the most effective algorithm for accurately predicting whether a shipment will be delivered on time or delayed.

The dataset used in this study includes various shipment-related attributes such as shipping mode, distance, delivery time, warehouse location, product importance, and customer details. To improve model performance, data preprocessing techniques such as handling missing values, feature scaling, and encoding of categorical variables were applied. Multiple machine learning algorithms, including Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, and Gradient Boosting, were implemented and analyzed.

The models were evaluated using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The experimental results indicate that ensemble learning techniques, particularly Random Forest and Gradient Boosting, achieved higher prediction accuracy compared to other models. These algorithms effectively captured complex patterns in the data and minimized prediction errors.

I. Introduction

The rapid growth of global supply chains and e-commerce logistics has significantly increased the complexity of transportation and delivery systems. Ensuring timely shipment delivery has become one of the most critical challenges for logistics companies and supply chain managers. Delivery delays can result in increased operational costs, reduced customer satisfaction, and disruptions across the entire supply chain network. Therefore, predicting shipment delays in advance is an important task in logistics management and transportation analytics .

Traditional forecasting methods mainly rely on statistical models, which may not effectively capture complex relationships among various operational factors such as weather conditions, warehouse workload, carrier efficiency, and transportation schedules. These limitations make it difficult to accurately predict delays in dynamic logistics environments. With the advancement of Artificial Intelligence (AI) and Machine Learning (ML), data-driven approaches have emerged as powerful tools for predicting logistics performance and operational risks .

Machine learning models are capable of analyzing large volumes of historical data and identifying hidden patterns that influence shipment outcomes. Algorithms such as Decision Trees, Random Forest, Support Vector Machines, Logistic Regression, and Gradient Boosting have shown promising results in predictive analytics. Among these, ensemble learning techniques like Random Forest and Gradient Boosting provide higher accuracy and robustness by combining multiple models.

II. Literature Survey

Several research studies have explored the application of machine learning and artificial intelligence in shipment delay prediction and logistics forecasting. Fernandes et al. (2026) analyzed vessel arrival data over a 15-year period and found that only 1.77% of shipments arrived on time, highlighting the complexity of delay prediction and the need for advanced AI-based models .

Kahl et al. (2026) evaluated different AutoML platforms for forecasting supply chain operations and demonstrated that automated machine learning systems can effectively support prediction tasks in logistics environments . Similarly, El-Sayegh et al. (2026) developed a hybrid deep learning model (CNN-BiLSTM) for delay prediction in construction projects, achieving high accuracy of 95.56%, which proves the effectiveness of AI-based predictive models .

Zhang et al. (2026) proposed a stacked ensemble model for ETA prediction in shipping, combining algorithms such as LightGBM, XGBoost, and Random Forest. Their approach significantly reduced prediction error, demonstrating the strength of ensemble learning in transportation forecasting . Likewise, Ben-Daya et al. (2024) applied deep learning techniques such as CNN-LSTM for late delivery risk prediction and achieved improved performance compared to traditional models .

Park et al. (2025) developed a real-time risk prediction system for drone-based logistics using machine learning, improving safety and prediction accuracy in dynamic environments . Nugroho et al. (2024) demonstrated the effectiveness of deep neural networks in prediction tasks, achieving high accuracy compared to traditional algorithms

III. System Analysis

System analysis focuses on understanding the need for an efficient shipment delay prediction system in modern logistics. With the rapid growth of e-commerce and global supply chains, timely delivery has become critical for maintaining customer satisfaction and operational efficiency. Delays in shipments can lead to increased costs, resource mismanagement, and supply chain disruptions. Traditional forecasting methods are not capable of handling complex and dynamic logistics data. There is a need for intelligent systems that can analyze multiple operational factors such as shipping mode, distance, warehouse conditions, and delivery schedules. Machine learning provides a suitable approach by identifying hidden patterns in large datasets. The system must ensure high accuracy, scalability, and reliability. Data preprocessing and feature engineering play an important role in improving model performance. The system should also support real-time prediction and decision-making. Overall, the aim is to develop a smart predictive system for delay forecasting.

Existing System

The existing system for shipment delay forecasting mainly relies on traditional statistical methods and manual analysis. Logistics companies use historical data and basic forecasting techniques to estimate delivery times. These methods include rule-based systems and simple regression models. However, such approaches are limited in handling large-scale and complex datasets. They do not consider multiple influencing factors simultaneously, such as weather, traffic, warehouse load, and carrier performance. Manual monitoring and decision-making are commonly used, which can lead to errors and inefficiencies. Existing systems lack automation and real-time prediction capabilities. They also provide limited accuracy in dynamic logistics environments. Most systems do not use advanced machine learning techniques for prediction. As a result, they fail to provide reliable delay forecasts and proactive solutions.

Disadvantages of Existing System

- Low prediction accuracy
- Inability to handle complex data patterns
- Lack of real-time prediction
- Heavy reliance on manual analysis
- Time-consuming forecasting process
- Limited use of multiple influencing factors

Proposed System

The proposed system is a machine learning-based shipment delay prediction model that performs a comparative analysis of multiple classification algorithms. The system uses a dataset containing shipment-related attributes such as shipping mode, distance, delivery time, warehouse location, and customer information. Data preprocessing techniques such as cleaning, normalization, and encoding are applied to improve data quality. Various classification algorithms including Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, and Gradient Boosting are implemented. Each model is trained and tested using the dataset. Performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. The system compares all models to identify the most effective algorithm for delay prediction. Ensemble methods like Random Forest and Gradient Boosting are expected to provide better performance. The system can be integrated into logistics platforms for real-time predictions. Overall, it provides an efficient and intelligent solution for shipment delay forecasting.

Advantages of Proposed System

- High prediction accuracy
- Ability to handle large and complex datasets
- Supports real-time prediction
- Reduces manual effort
- Uses multiple influencing factors for better analysis
- Provides comparative analysis of algorithms
- Improves decision-making in logistics

IV. Methodology

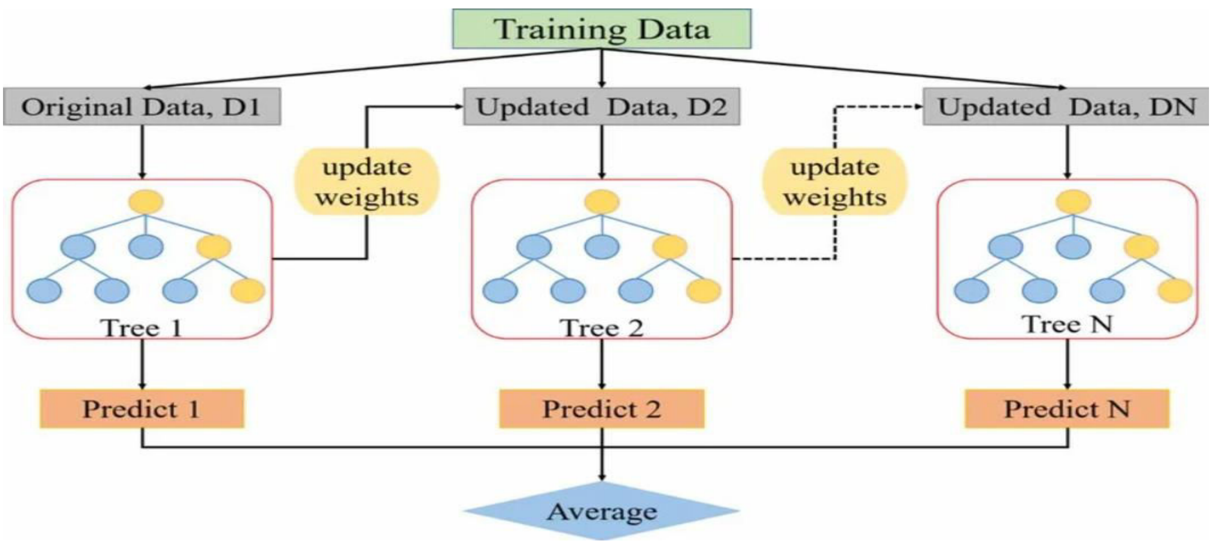
The proposed system follows a structured methodology to predict shipment delays using multiple machine learning classification algorithms. Initially, a logistics dataset is collected containing features such as shipping mode, distance, delivery time, warehouse location, product importance, and customer details. The data then undergoes preprocessing, which includes handling missing values, removing inconsistencies, scaling numerical features, and encoding categorical variables. After preprocessing, feature selection techniques are applied to identify the most relevant attributes that influence shipment delays.

The dataset is then split into training and testing sets to evaluate model performance. Various classification algorithms such as Decision Tree, Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting are implemented. Each model is trained using the training data and tested on unseen data. Performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix are used for evaluation. A comparative analysis is conducted to determine the best-performing algorithm. Finally, the selected model is used to predict whether a shipment will be delayed or delivered on time, supporting efficient logistics decision-making.

System Architecture

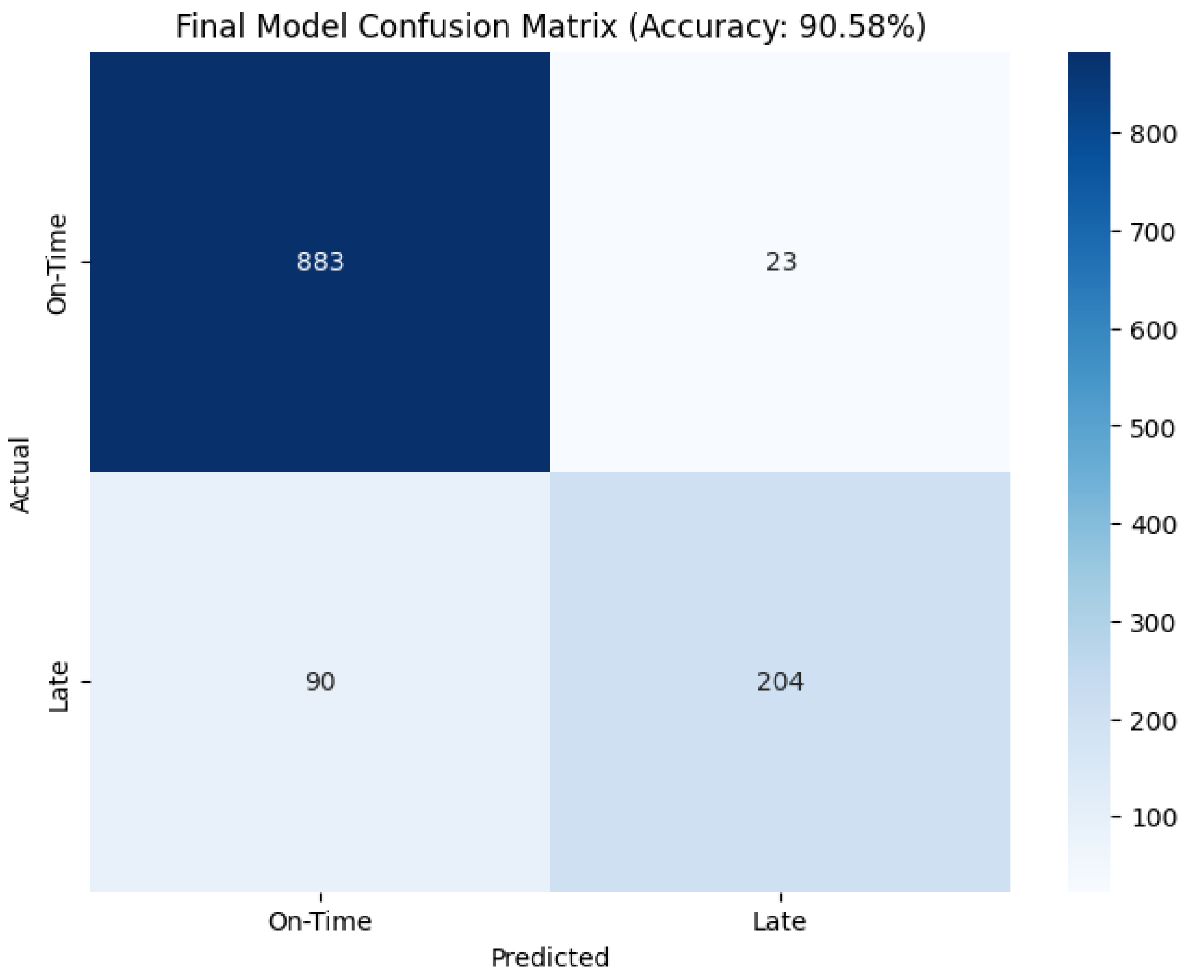
The system architecture for the shipment delay prediction model is designed to process logistics data efficiently and generate accurate predictions. Initially, shipment data is collected either from user input or historical datasets, which include attributes such as shipping mode, distance, warehouse location, delivery time, and customer details.

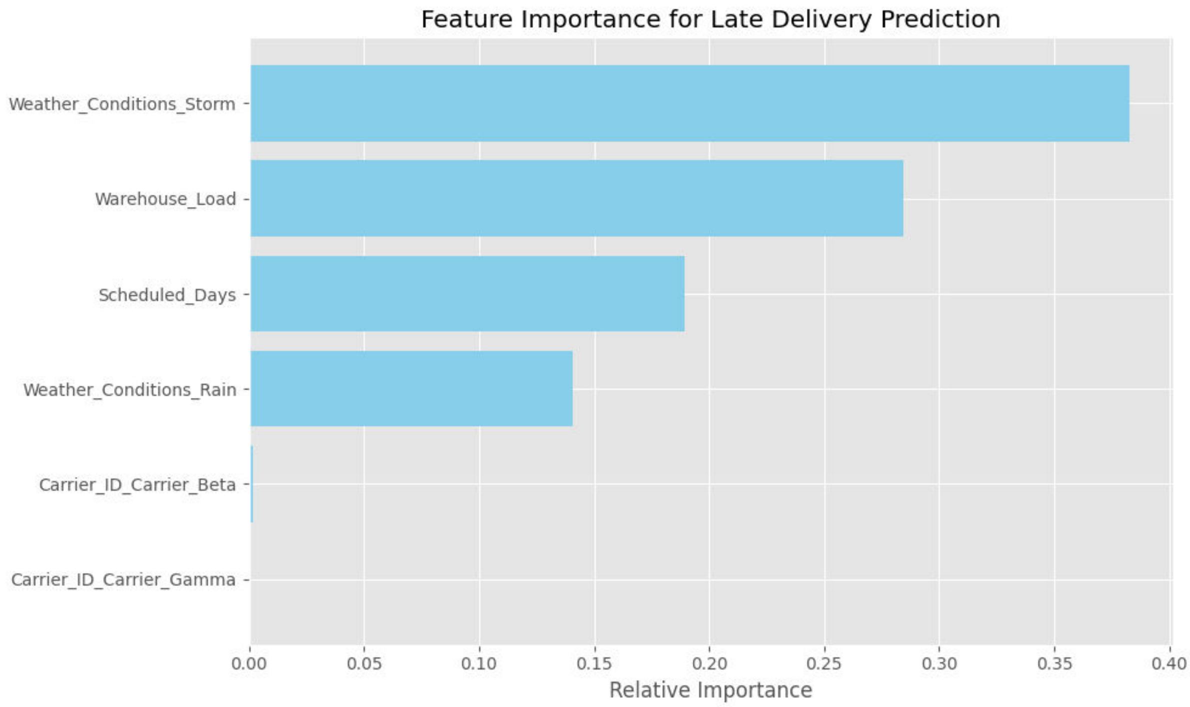
After preprocessing, feature selection is applied to identify the most relevant attributes that influence shipment delays. The processed data is then split into training and testing sets. Multiple machine learning models, including Decision Tree, Logistic Regression, Support Vector Machine (SVM), Random Forest, and Gradient Boosting, are trained using the training data.

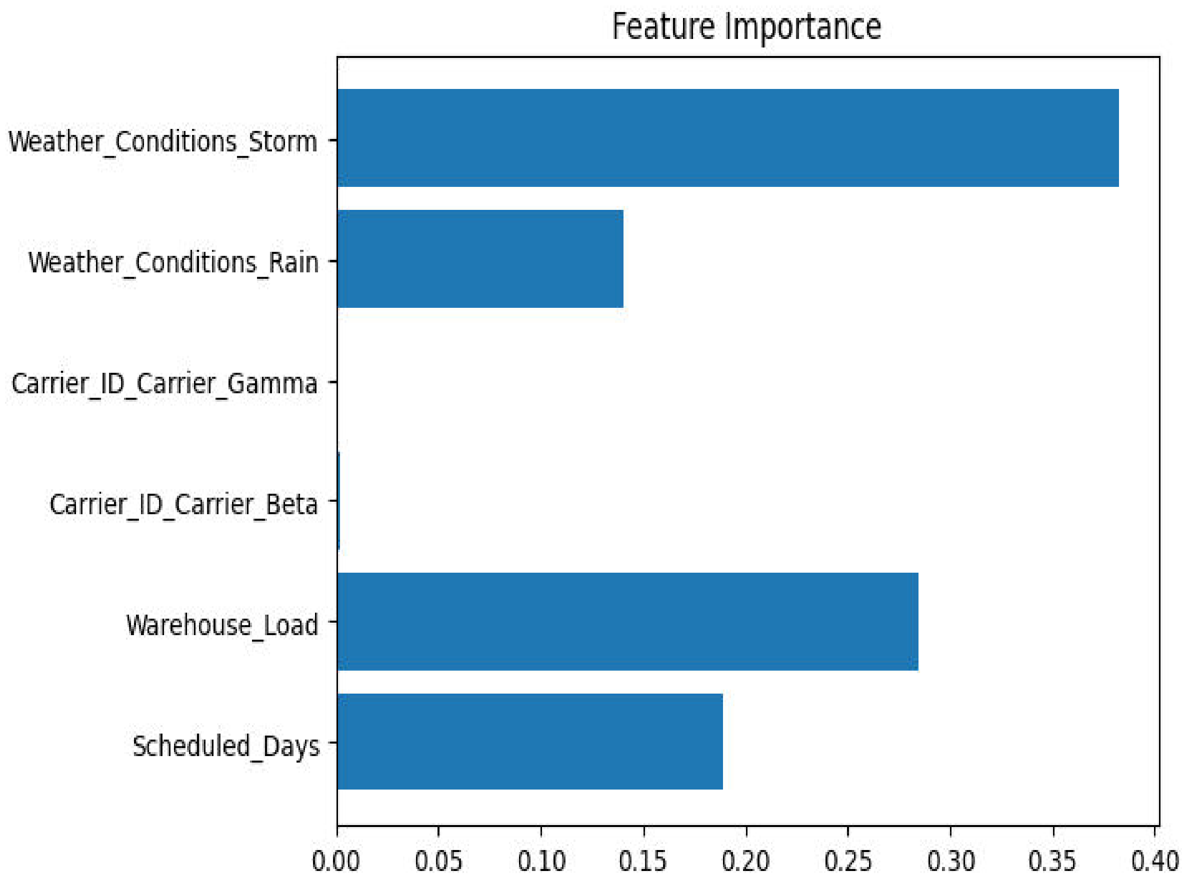


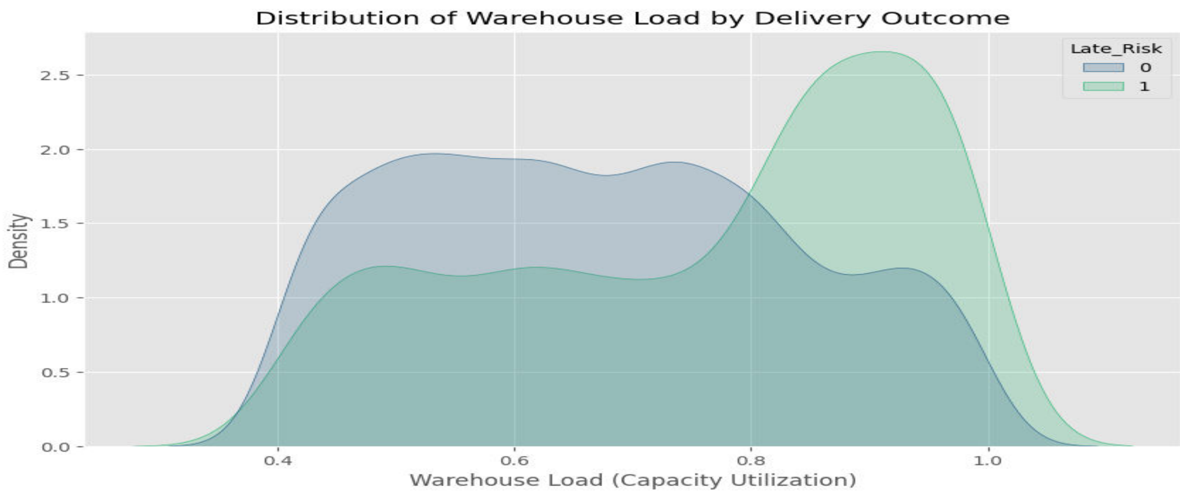
V. Result and Output

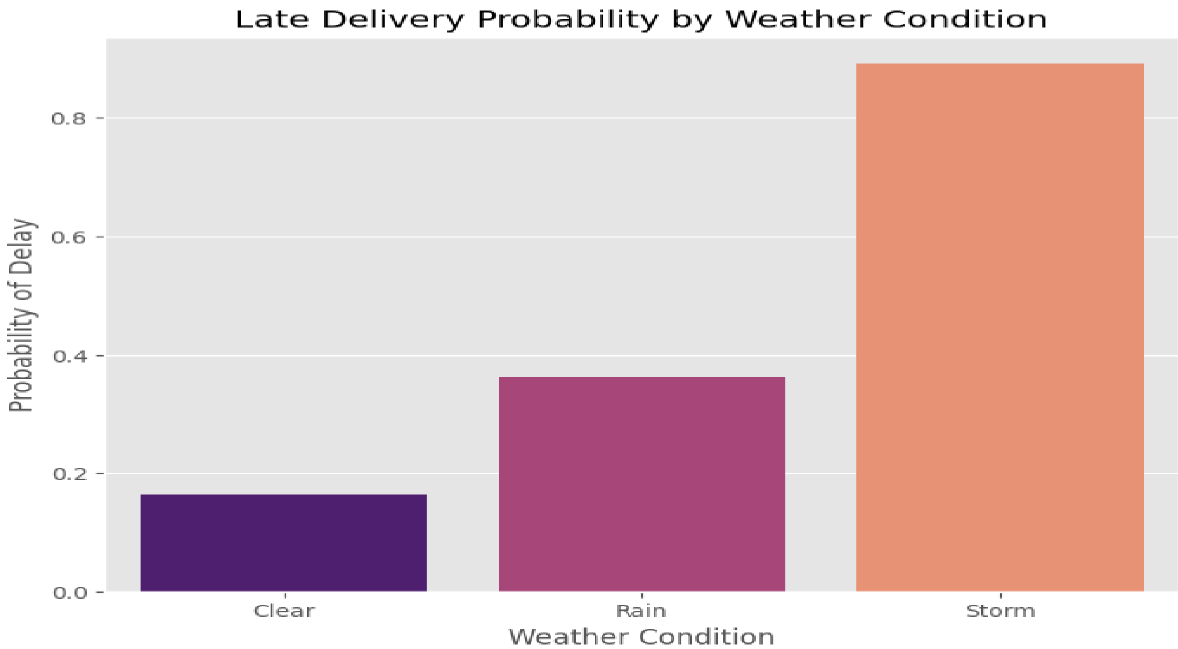
```
Model Performance Summary
-------------------------
Achieved Accuracy: 90.58%

Classification Report:
              precision    recall  f1-score   support

           0       0.91      0.97      0.94       906
           1       0.90      0.69      0.78       294

    accuracy                           0.91      1200
   macro avg       0.90      0.83      0.86      1200
weighted avg       0.91      0.91      0.90      1200
```




Feature Importance for Late Delivery Prediction
Weather_Conditions_Storm
Warehouse_Load
Scheduled_Days
Weather_Conditions_Rain
Carrier_ID_Carrier_Beta
Carrier_ID_Carrier_Gamma
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Relative Importance


Feature Importance
Weather_Conditions_Storm
Weather_Conditions_Rain
Carrier_ID_Carrier_Gamma
Carrier_ID_Carrier_Beta
Warehouse_Load
Scheduled_Days
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40






```
Enter Carrier ID (Carrier_Alpha, Carrier_Beta, Carrier_Gamma):
Enter Scheduled Delivery Days (2, 3, 5, 7): 2
Enter Weather Conditions (Clear, Rain, Storm): Rain
Enter Warehouse Load (0-1): 1
⚠ Shipment will be DELAYED
```

```
Model Performance
------------------------
Accuracy : 0.91
Precision: 0.9
Recall   : 0.69
F1 Score : 0.78

Confusion Matrix
[[883  23]
 [ 90 204]]
```

VI. Conclusion

This project demonstrates the effectiveness of machine learning techniques in predicting shipment delays in logistics and supply chain systems. By implementing and comparing multiple classification algorithms such as Decision Tree, Logistic Regression, Support Vector Machine, Random Forest, and Gradient Boosting, the study identifies the most suitable model for accurate delay prediction. The results show that ensemble learning methods, particularly Random Forest and Gradient Boosting, provide higher accuracy and better performance due to their ability to capture complex patterns in data.

The proposed system improves decision-making by enabling early identification of potential delays, helping logistics companies optimize delivery planning and reduce operational risks. It also reduces manual effort and enhances the efficiency of supply chain operations. Although the system cannot eliminate uncertainties completely, it serves as a powerful support tool for predictive analytics in logistics.

In the future, integrating real-time data, advanced deep learning models, and IoT-based tracking systems can further improve prediction accuracy and provide more dynamic and scalable solutions for modern logistics challenges.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.