BUSINESS APPLICATIONS OF DATA SCIENCE - PROJECT 1

Predicting On-Time Delivery in Supply Chain

Aadya Sood, Arvind Yadav, Gaurav Dixit,
Radhika Swaroop, Yashasvee Singh

INTRODUCTION

Timely delivery of orders is an important factor in supply chain management, influencing customer satisfaction and operational efficiency. Delays in delivery can lead to increased costs, disrupted operations, and decreased reliability. The aim of this project is to analyze operational data and predict whether an order will be delivered on time, using various machine learning algorithms.

To achieve this, we have explored four different algorithms: K-Means Clustering, K-Nearest Neighbors (KNN), Naive Bayes, and the Classification and Regression Tree (CART) algorithm. We compared the workings of these algorithms to see which one would give us the highest accuracy, hence being the most suitable model. After evaluating their performance, the CART algorithm emerged as the best-performing model for this problem.

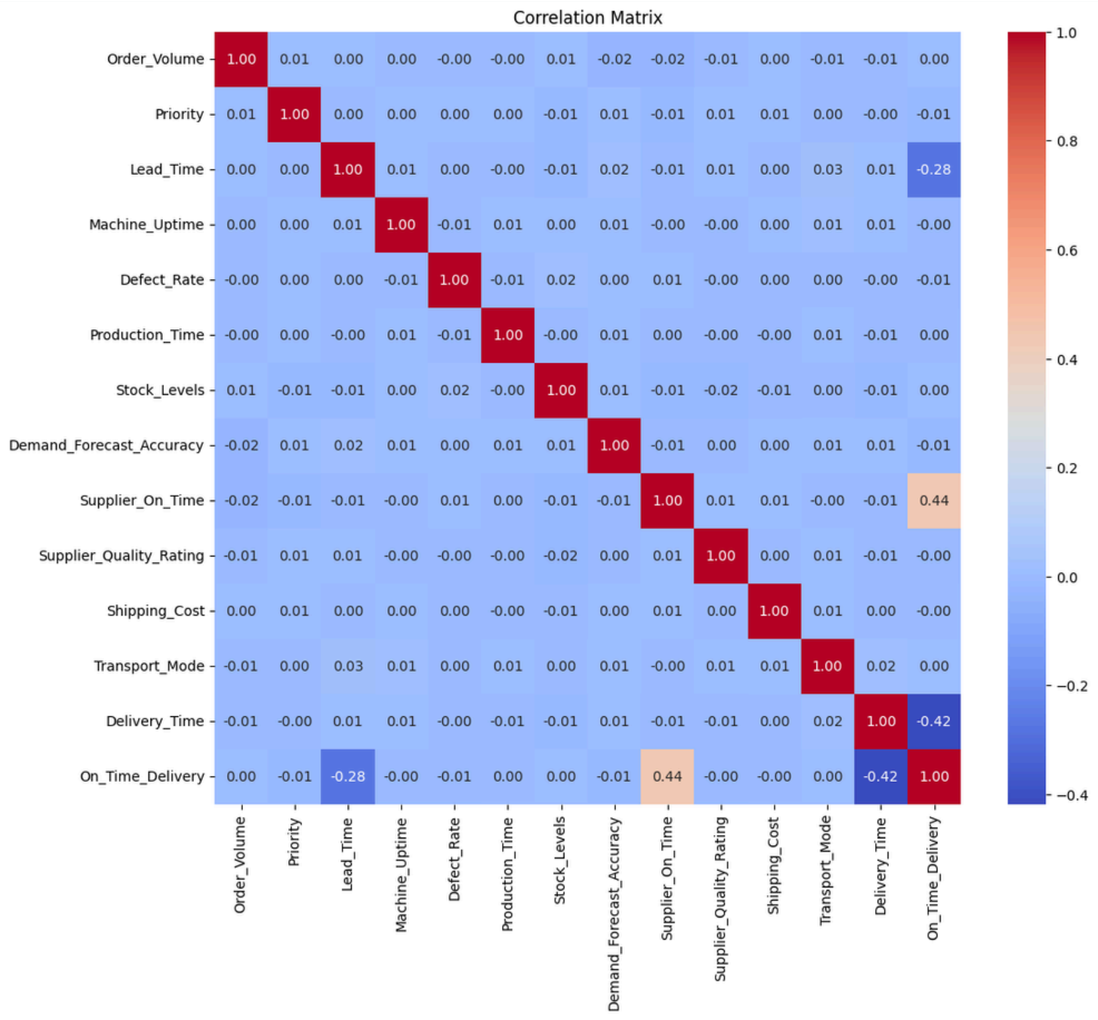
DATASET OVERVIEW

The dataset provided includes operational metrics spanning different aspects of the supply chain. Each of these features plays a crucial role in determining the likelihood of an order arriving on time.

Dataset Details:

- **Order Details**: Order volume, priority, lead time.
- **Production Efficiency**: Machine uptime, defect rate, production time.
- **Inventory Management**: Stock levels, demand forecast accuracy.
- **Supplier Performance**: Supplier on-time percentage, quality rating.
- **Logistics & Shipping**: Shipping cost, transport mode, delivery time.
- **Target Variable**: On_Time_Delivery (1 = On-time, 0 = Delayed).

CORRELATION MATRIX



- There is a moderate positive correlation between supplier reliability and on-time delivery. This shows that supplier performance will play. animportant part in the predictive model.
- A strong negative correlation between Delivery Time and On-Time Delivery suggests that longer delivery times reduce the likelihood of on-time deliveries.
- There is a mild negative correlation between Lead Time and On-Time Delivery, indicating that longer lead times can negatively impact on-time delivery.
- Demand forecast accuracy does not show a strong correlation with on-time delivery, suggesting that other operational factors have a more direct impact.
- There is minimal impact from Machine Uptime, Defect Rate, and Production Time. These variables have very low correlation with on-time delivery, implying that internal manufacturing issues may not be the primary driver of delays.

METHODOLOGY

1 Data Preprocessing

We checked for any missing values , encoded categorical variables, and standardized numerical features. Our data had no missing values. We then divided the dataset into training and testing sets (80-20 split) for model evaluation.

2 Model Implementation

We applied four machine learning models (K-Means, KNN, Naive Bayes and CART) to analyze operational data and make a prediction for on-time delivery.

3 Performance Comparison

We evalutated the model performance using accuracy and interpretability as our variables of concern. A confusion matrix was made for the models - it provides a comprehensive overview of the model's classification performance by displaying the counts of actual versus predicted classifications.

4 Results and Analysis

We analysed the results and observed that the CART algorithm showed the best model performance. We then gave recommendations for business strategy according to our interpretation of the model evalutation.

DATA PREPROCESSING

Before applying machine learning models, the dataset underwent several preprocessing steps which helped in preparing data for better model performance. These were:

- **Handling Missing Values:**

Checked for missing or null values and imputed or removed them where necessary.

- **Encoding Categorical Variables:**

Converted categorical variables (e.g., Priority, Transport Mode) into numerical form using label encoding and one-hot encoding.

- **Feature Scaling:**

Standardized numerical features to ensure uniformity across the dataset.

- **Data Splitting:**

The dataset was divided into an 80-20 train-test split to train and evaluate models effectively.

K MEANS CLUSTERING

K-Means is an unsupervised learning algorithm used to identify patterns and clusters in the data. It was applied for exploratory data analysis to visualize how orders grouped based on their delivery status.

Approach

- The algorithm grouped orders into clusters based on similarities in their features.
- The optimal number of clusters (k) was determined using the elbow method.

Limitations

- K-Means does not directly predict delivery status.
- It is useful for understanding data patterns but lacks classification capabilities.

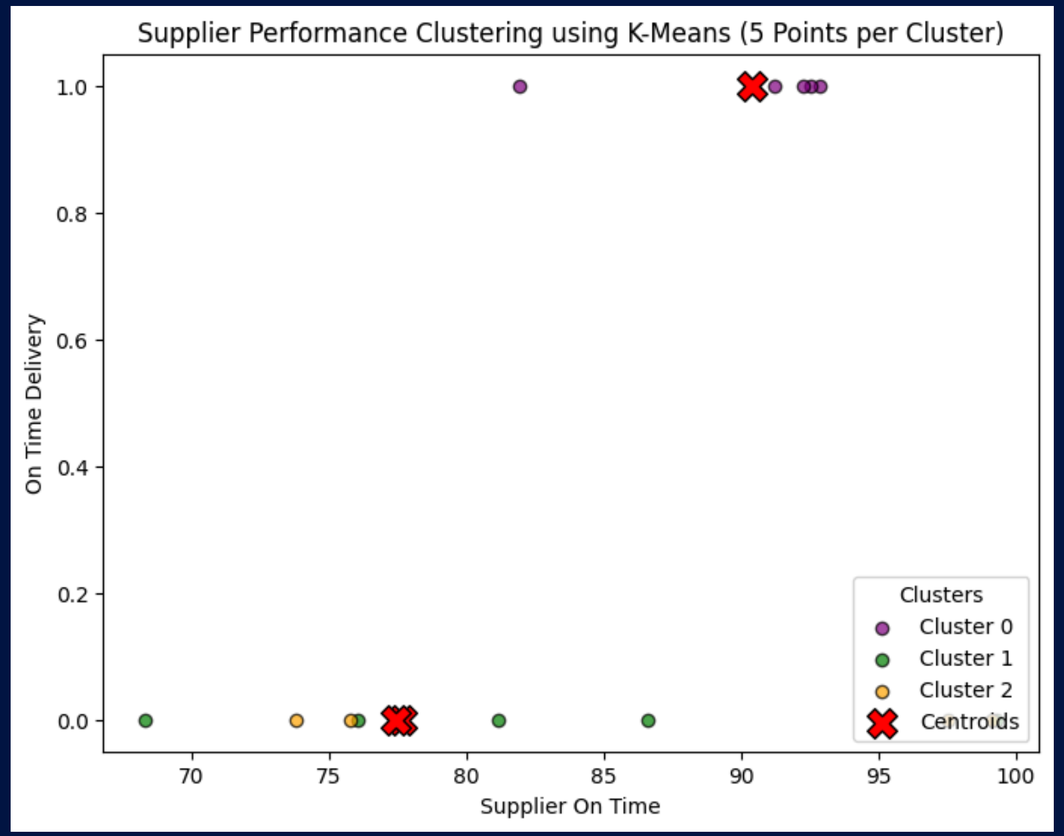
K MEANS CLUSTERING

This figure represents the clusters applied to assess supplier performance based on two variables: Supplier On Time and On Time Delivery.

Three distinct clusters were identified and plotted, with each cluster represented by a different colours. The centroids of each cluster, depicted as red 'X' markers, represent the central points of the clusters.

The plot includes 5 representative data points from each cluster, randomly sampled.

- Cluster 0 primarily includes suppliers with higher on-time delivery rates, indicating reliable performance.
- Cluster 1 consists of suppliers with relatively low on-time delivery, suggesting potential operational inefficiencies.
- Cluster 2 features suppliers with moderate performance, possibly indicating variability in delivery consistency.



K-NEAREST NEIGHBOUR

KNN is a supervised learning algorithm that classifies data points based on their proximity to labeled training examples.

Approach

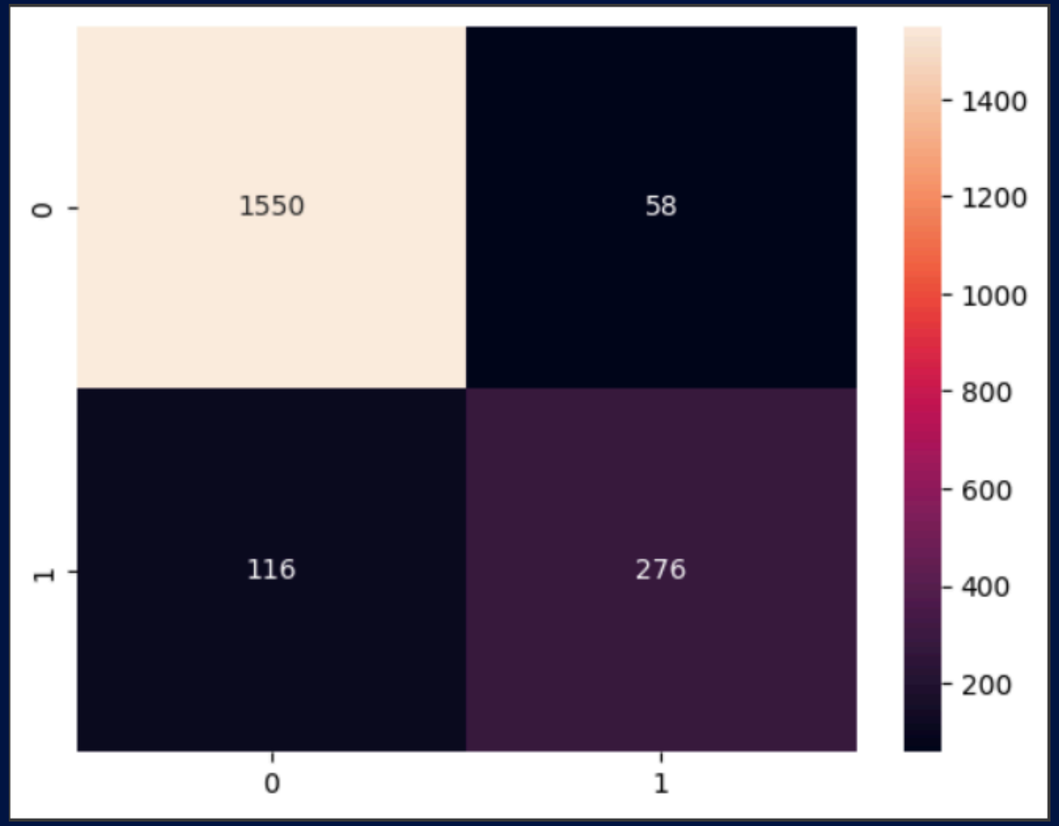
- Chose k=3 for classification.
- Used Euclidean distance to determine the similarity between orders.
- Required feature scaling due to distance-based calculations.

Limitations

- Performance degrades with large datasets.
- Sensitive to irrelevant features and requires careful tuning of k.

K-NEAREST NEIGHBOUR

- The model exhibits a high number of true negatives, indicating good performance in identifying negative cases.
- The presence of 58 false positives suggests some misclassifications where suppliers may have been incorrectly classified as positive.
- The 116 false negatives indicate that some actual positive cases were missed, which could impact the overall sensitivity of the model.



THIS MODEL ACHIEVED AN ACCURACY OF 91.3%.

NAIVE BAYES

Naïve Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence between features.

Approach

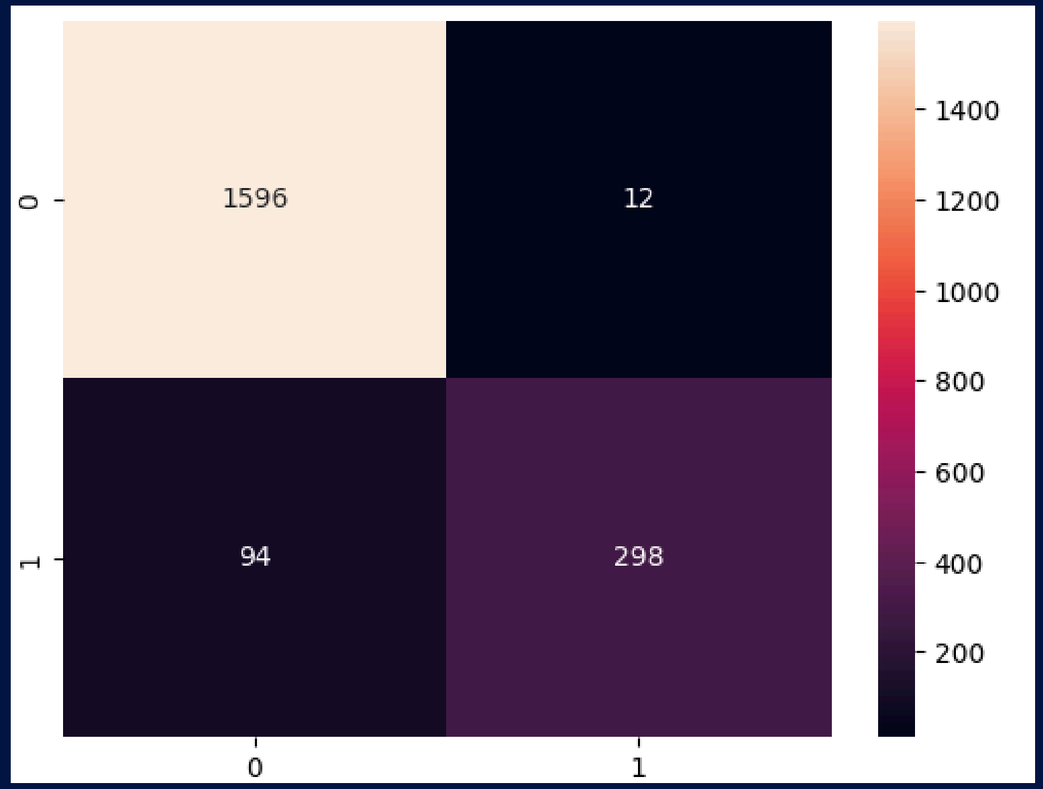
- Calculated conditional probabilities for each feature.
- Classified orders as on-time or delayed based on likelihood estimation.

Limitations

- Less effective when features are interdependent.
- Not the best choice for complex relationships in supply chain data.

NAIVE BAYES

- Out of all the negative instances, 1596 were correctly classified (True Negatives), while 12 were misclassified as positive (False Positives).
- Among the positive instances, 298 were correctly identified (True Positives), but 94 were misclassified as negative (False Negatives).
- The model achieved a high accuracy, indicating that it performed well in distinguishing between the two classes.
- While the false positive rate is low, the false negatives suggest that some positive cases were missed. This may indicate areas for improvement, particularly if correctly identifying positive instances is crucial in the given context.
- Overall, the Naive Bayes classifier demonstrated strong predictive performance on the dataset.



THIS MODEL ACHIEVED AN ACCURACY OF 94.7%.

CLASSIFICATION AND REGRESSION TREE (CART)

CART is a decision tree-based model that splits data into branches based on feature importance.

Approach

- Constructed a decision tree to classify orders.
- Used Gini index for optimal split selection.
- Provided interpretable decision-making rules for business insights.
- A depth of 4 was chosen - at a depth of 7, the model achieved 100% accuracy. But, as it would lead to overfitting of the data, the optimum max-depth is kept at the value over which the maximum accuracy is achieved with significant marginal improvement over the previous value.

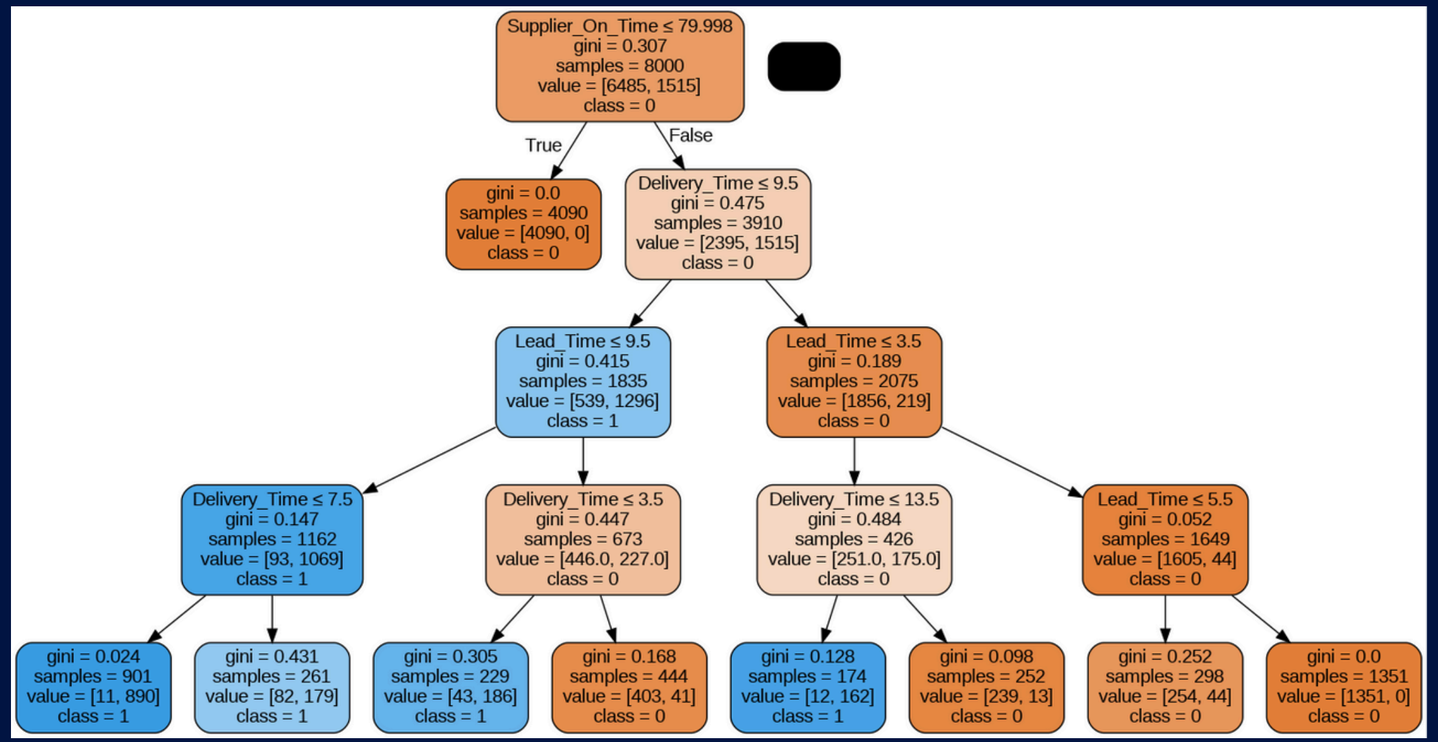
Limitations

- Can overfit, especially with complex datasets, reducing generalisation to new data.
- Sensitive to small data changes, leading to unstable results.
- May become biased toward dominant classes in imbalanced datasets.

CART – DECISION TREE

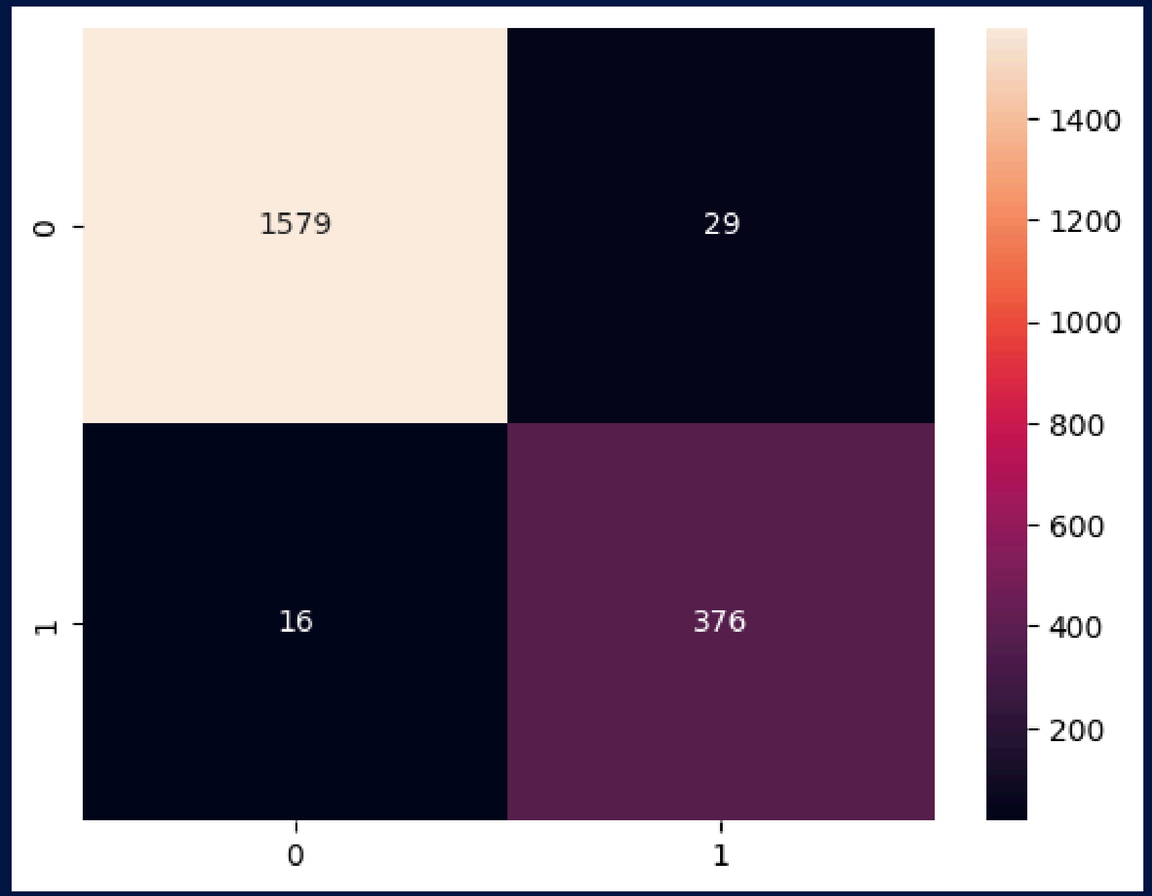
- The decision tree provides a clear and interpretable view of how decisions are made based on the dataset. It highlights that Supplier On-Time Performance is the most significant factor influencing the classification of orders. This suggests that suppliers meeting deadlines play a critical role in ensuring successful order delivery.

- Subsequent nodes refine the classification based on Delivery Time and Lead Time, indicating that logistical factors are also essential for predicting outcomes. Shorter lead and delivery times tend to correlate with successful orders, while delays increase the likelihood of unsuccessful ones.



CART - CONFUSION MATRIX

- The low number of False Positives (29) suggests that the model rarely misclassifies an unsuccessful order as successful, which is crucial for minimizing risk in business decisions.
- Similarly, with only 16 False Negatives (FN), the model demonstrates excellent sensitivity in capturing true unsuccessful orders. This makes it a reliable tool for identifying potential risks in the supply chain.
- Overall, the CART model shows a strong balance between accuracy and interpretability.



THIS MODEL ACHIEVED AN ACCURACY OF 97.75%.

WHY CART FOR FINAL ANALYSIS?

High Accuracy with Balanced Complexity

- 97.75% accuracy at depth=4 represents an excellent balance between predictive power and model simplicity.
- Prevents overfitting (which would occur at a higher depth with 100% accuracy) while maintaining substantially better performance than simpler models

Transparent Decision Rules

- Unlike KNN and Naive Bayes, CART provides explicit decision paths that supply chain managers can understand and implement.
- Since CART builds a decision tree, it reveals the most influential factors affecting delays (e.g., supplier performance, transport mode, lead time).

Excellent Performance Metrics

- High true positive and true negative rates indicate balanced performance across both on-time and delayed deliveries.
- Only 45 misclassifications out of 2000 orders demonstrates superior reliability

RECOMMENDATIONS FOR BUSINESS STRATEGY

Prioritize Supplier Performance Management:

- Since Supplier On-Time Performance is the most critical factor, businesses should establish strict supplier management protocols.
- Implement performance contracts and introduce penalties for delays while offering incentives for consistent on-time delivery.
- Maintain a database of supplier performance and use predictive insights to choose reliable suppliers.

Optimise Lead Time Management:

- The tree highlights Lead Time as a significant factor. Companies should collaborate closely with suppliers to reduce lead times by optimizing procurement and production schedules.
- Implement real-time tracking systems for early detection of potential delays and take proactive measures.

RECOMMENDATIONS FOR BUSINESS STRATEGY

Enhance Delivery Time Efficiency:

- For shorter delivery times, streamline logistics operations by working with third-party logistics providers with proven performance records.
- Invest in route optimisation technology and real-time monitoring to reduce transportation delays.

Develop Contingency Plans:

- For orders flagged as high-risk for delays, businesses should have contingency plans, including expedited shipping options or maintaining buffer inventory.
- Predictive insights from CART can help in identifying vulnerable orders in advance.

THANK YOU!