Predicting On-Time Shipment Performance Using Machine Learning

Instructor
Dr. Michael Hudson
Student
Ngan Huynh
Date: Oct 28th, 2025

Abstract. This research project focused on applying machine learning techniques, Logistics Regression, and Support Vector Machine (SVM) to predict the likelihood of on-time shipments. Using the real-world dataset of Amazon delivery operations, I engineered geographic features using the Haversine formula and defined on-time performance relative to the median delivery time (<125 minutes). The performance of the two models was evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC. The SVM model achieved a slightly higher accuracy and F1-score compared to Logistic Regression, while both obtained an identical AUC, demonstrating similar classification capability. Subsequently, K-Means clustering combined with Principal Component Analysis (PCA) was employed to segment the dataset and identify underlying behavior patterns. The Elbow Method suggested four optimal clusters, and PCA visualization in two dimensions revealed distinct groups reflecting variations in delivery speed, distance, and agent performance.

Keywords: shipment on time, logistics regression, SVM, k-means clustering, machine learning, model evaluation

1. Introduction

The rapid acceleration of the e-commerce sector has made on-time package delivery the single most important determinant of customer loyalty and logistical success. Consequently, logistics networks are under immense pressure to achieve high service reliability. Predicting the timeliness of a shipment is a highly dimensional challenge influenced by dynamic variables such as traffic, weather, area, route distance, and the performance history of the delivery agent.

To address this problem, actionable foresight into these operations, this project utilizes a hybrid machine learning approach. I deploy two supervised classification models, Logistic Regression and Support Vector Machine (SVM), trained on Amazon's historical delivery data to accurately predict delivery performance. Crucially, this predictive capability is paired with an unsupervised technique, K-Means clustering coupled with Principal Component Analysis (PCA), to identify and segment operational patterns. This combined methodology transforms raw data into a powerful tool for strategic resource allocation and continuous efficiency improvement.

2. Literature Review

The application of machine learning in logistics is well-established, focusing broadly on optimization, forecasting, and risk prediction. Machine learning models such as Logistic Regression, Decision Trees, Random Forests, and SVM have shown promising results in classification-based logistics problems. Logistic Regression provides interpretability, while SVM excels in handling nonlinear boundaries. Clustering methods such as K-Means, combined with dimensionality reduction techniques like PCA, have been effective in discovering hidden delivery behavior patterns. This project builds these foundations by integrating classification for predictive and clustering for descriptive techniques to enhance understanding and accuracy of delivery performance.

3. Business Goals and Problem Definition

3.1 Business Goals

This project is driven by a dual objective to enhance logistics operations. First, it seeks to develop highly accurate predictive models that reliably forecast the probability of on-time versus late shipment delivery. Second, it aims to leverage unsupervised learning to segment the delivery ecosystem, identifying distinct performance-based clusters among agents and orders. The ultimate objective is to utilize this combined intelligence (forecasting risk and characterizing behavior) to optimize resource allocation and generate actionable insights that drive efficiency improvements in the critical last-mile delivery phase.

3.2 Modeling Environment

The project leveraged data preprocessing, feature engineering, and exploratory data analysis using Pandas, Seaborn, and Matplotlib. Predictive modeling was performed with Scikit-learn implementations of Logistic Regression, SVM, and K-Means with PCA for dimensionality reduction. The dataset consists of 43,739 records and 16 attributes, with the target variable On_Time, where 1 indicates on-time delivery and 0 indicates late delivery. The selected features include Weather, Traffic, Distance, Area, Category, Agent Rating, Agent Age, and Vehicle.

4. Solution Methodology

4.1 Data models

The initial dataset contained 43,739 records and 16 attributes related to delivery operations, including agent demographics, location coordinates, weather, traffic, and delivery times. During the data cleaning phase, missing values in numerical attributes were imputed using the median, while categorical attributes were filled using the mode to preserve data consistency. All duplicate rows were identified and removed to prevent redundancy in the analysis.

The Haversine formula was applied to compute the distance between store and drop locations, creating a new feature (Distance_km). Categorical variables (e.g, Weather, Traffic, Area, Vehicle, Category) were encoded using One-Hot Encoding. Numerical variables (Agent Age, Agent Rating, Distance) were standardized using StandardScaler.

4.2 Statistical Summary and Correlation Analysis

A statistical overview revealed that the median delivery time was 125 minutes, consistent with the threshold for defining on-time performance.

Table 1: Statistical Summary and Percentile Analysis of Delivery Attributes

	Agent_Age	Agent_Rating	Store_Latitude	Store_Longitude	Drop_Latitude	Drop_Longitude	Delivery_Time
count	43739.000000	43685.000000	43739.000000	43739.000000	43739.000000	43739.000000	43739.000000
mean	29.567137	4.633780	17.210960	70.661177	17.459031	70.821842	124.905645
std	5.815155	0.334716	7.764225	21.475005	7.342950	21.153148	51.915451
min	15.000000	1.000000	-30.902872	-88.366217	0.010000	0.010000	10.000000
25%	25.000000	4.500000	12.933298	73.170283	12.985996	73.280000	90.000000
50%	30.000000	4.700000	18.551440	75.898497	18.633626	76.002574	125.000000
75%	35.000000	4.900000	22.732225	78.045359	22.785049	78.104095	160.000000
max	50.000000	6.000000	30.914057	88.433452	31.054057	88.563452	270.000000

Figure 1: Delivery Time Distribution

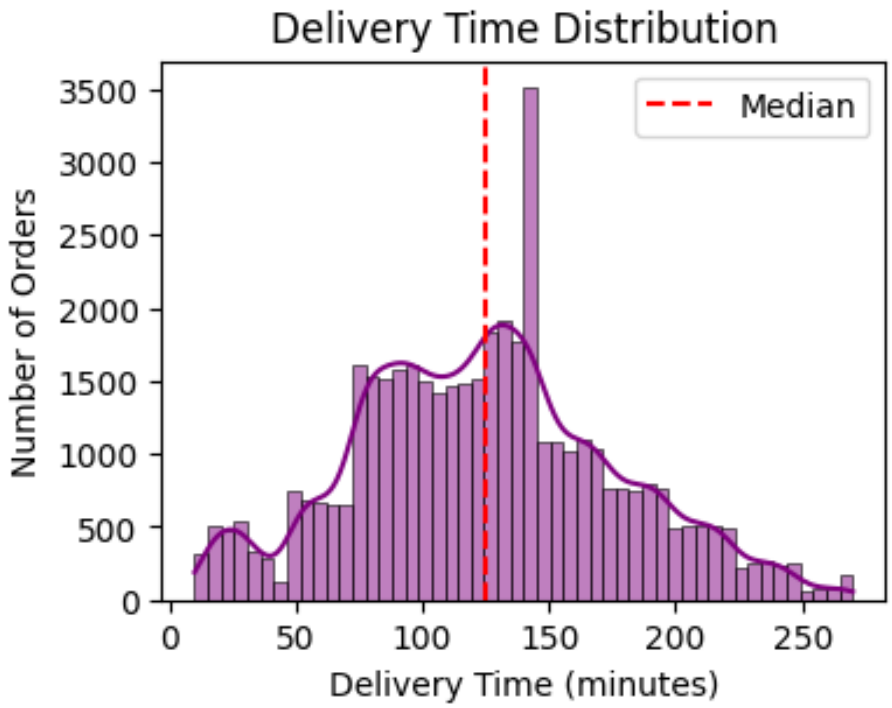
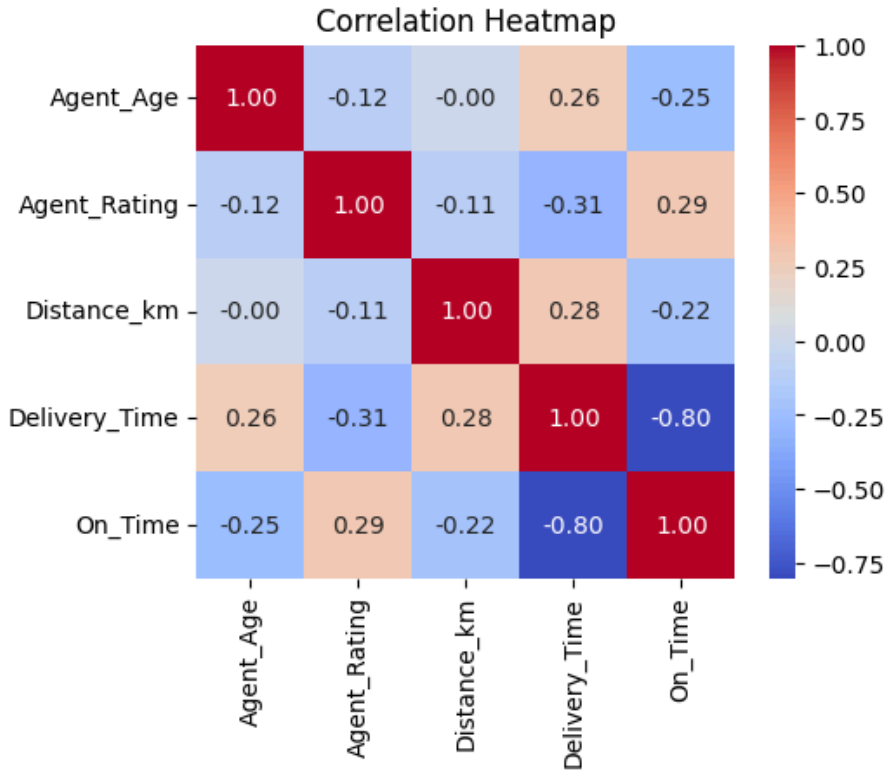


Figure 2: Correlation Heatmap



A strong negative correlation (-0.80) between On_Time and Delivery_Time confirmed that longer delivery durations were associated with delays.
A moderate positive correlation (0.29) between Agent_Rating and On_Time suggested that higher-rated agents are more likely to deliver on time.
A weak negative correlation (-0.25) between Agent_Age and On_Time implied that younger agents may perform slightly faster deliveries.

5. Machine Learning Models

5.1 Train/Test Split and Scaling

The dataset was divided into 80% for training and 20% testing. Numerical columns were scaled using StandardScaler, ensuring all features contributed equally to model training. Two models were implemented and evaluated: Logistic Regression and SVM.

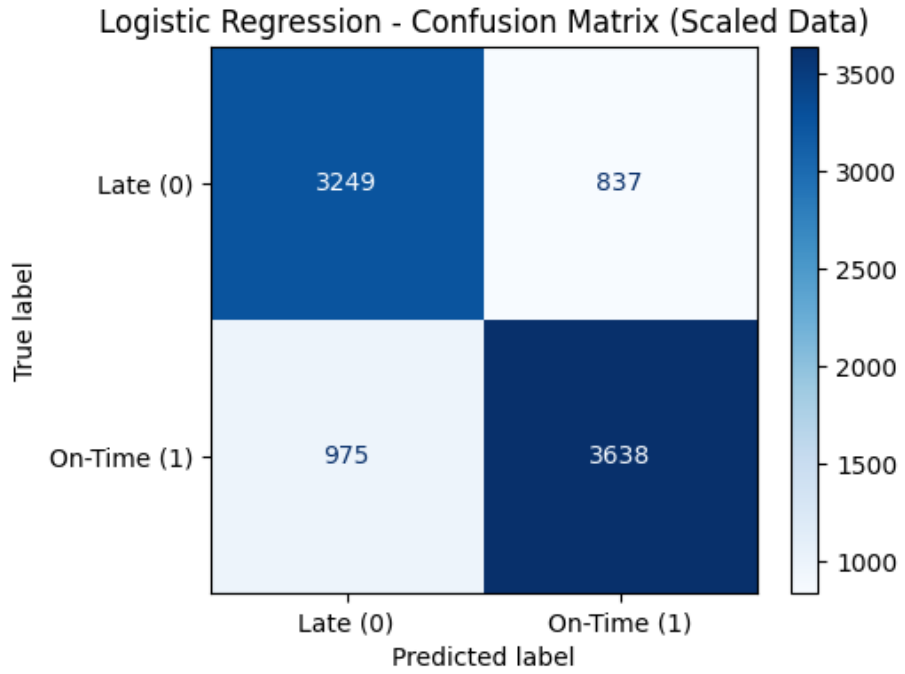
5.2 Logistic Regression

Table 2: Logistic Regression Performance Metrics (Scaled Data)

Metric	Class 0 (Late)	Class 1 (On-Time)	Average
Precision	0.77	0.81	-

Recall	0.80	0.79	-
F1-Score	0.78	0.80	-
Accuracy	-	-	0.79

Figure 3: Logistic Regression - Confusion Matrix (Scaled Data)



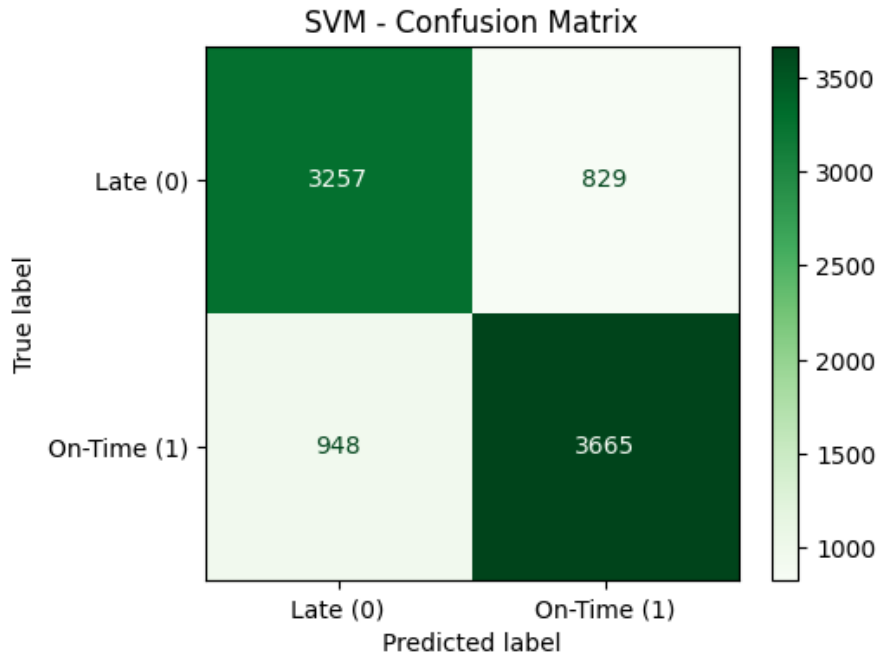
The confusion matrix, shown in Figure 3, indicated balanced performance across both classes, confirming the model's ability to generalize well.

5.3 Support Vector Machine (SVM)

Table 3: Support Vector Machine (SVM) Performance Metrics (Scaled Data)

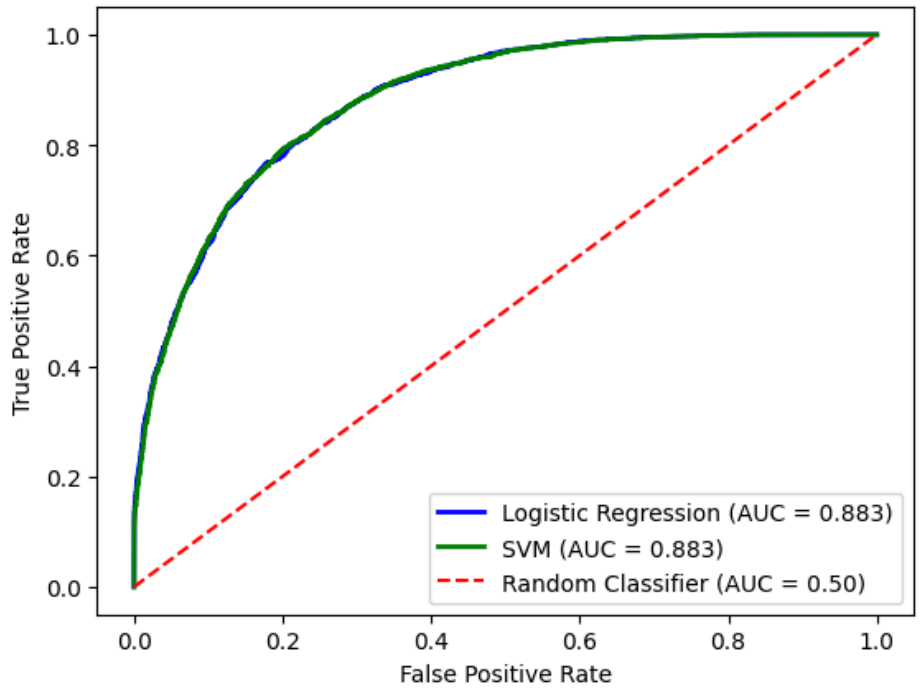
Metric	Class 0 (Late)	Class 1 (On-Time)	Average
Precision	0.77	0.82	-
Recall	0.80	0.79	-
F1-Score	0.78	0.80	-
Accuracy	-	-	0.80

Figure 3: SVM - Confusion Matrix (Scaled Data)



The SVM model slightly outperformed Logistic Regression in both accuracy and precision for on-time predictions.

5.4 ROC-AUC Comparison
Figure 4: ROC-AUC Comparison

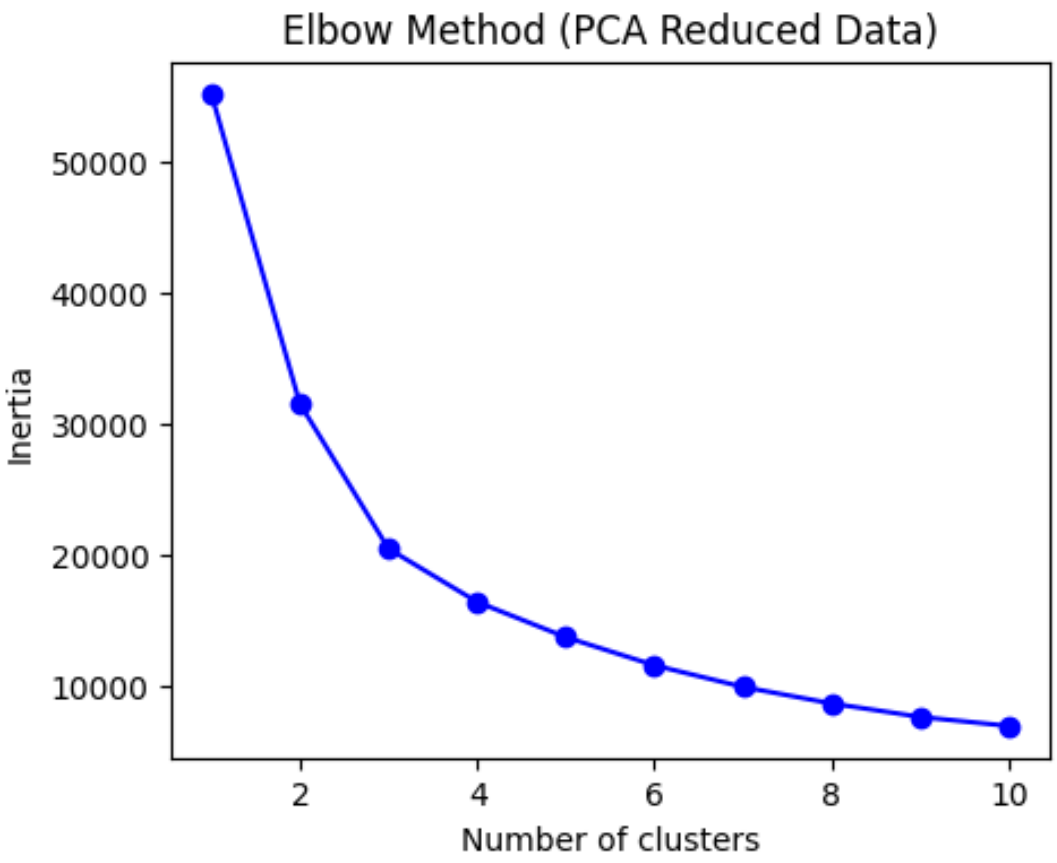


Both models achieved an identical AUC of 0.883, indicating strong discriminative power in predicting whether a shipment would be on time or late.
Although SVM showed marginally better accuracy, Logistic Regression remains more interpretable and computationally efficient.

6. Unsupervised Clustering: K-Means with PCA
To further analyze operational dynamics, K-Means Clustering was applied to four numerical features: Delivery_Time, Distance_km, Agent_Rating, and Agent_Age.

6.1 PCA Dimensionality Reduction
The dataset was scaled using RobustScaler, and PCA was used to reduce dimensionality from 4D to 2D, improving visualization without losing significant variance.

6.2 Elbow Method
Figure 5: Elbow Method (PCA Reduced Data)



The Elbow Method showed the optimal cluster number as k = 4, balancing simplicity and cluster separation.

6.3 Cluster Analysis

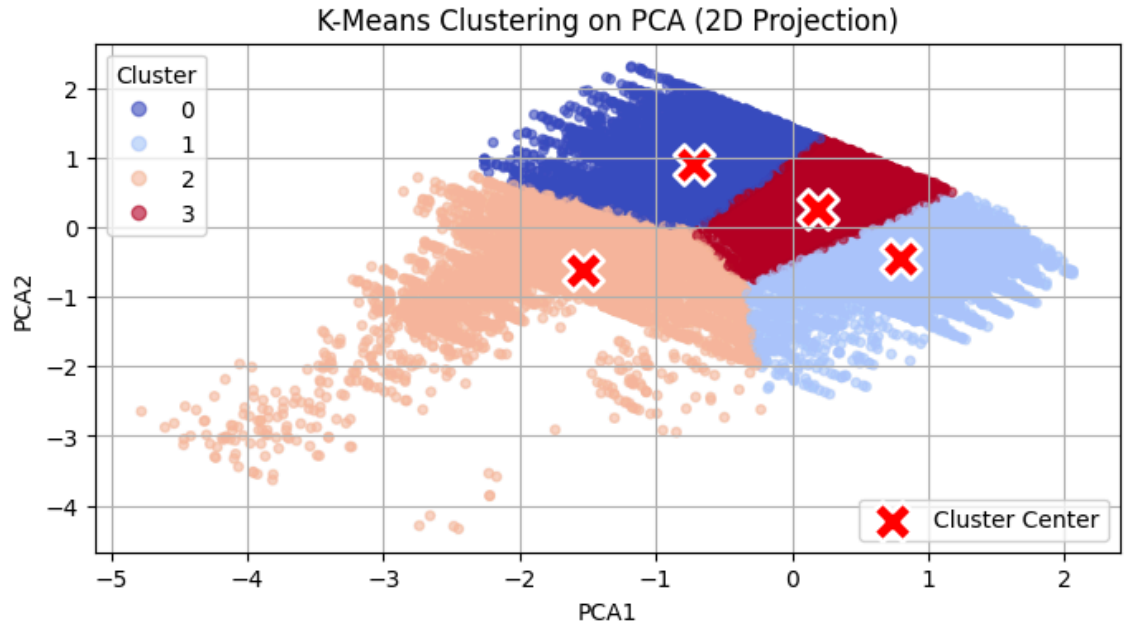
Table 4: K-Means Cluster Centers (After PCA)

Cluster	Delivery_Time (min)	Distance_km	Agent_Rating	Agent_Age
0	191.51	15.18	4.70	31.98
1	72.89	6.02	4.72	26.88
2	174.85	10.73	4.65	31.80
3	124.01	10.34	4.76	29.96

The clustering results reveal four distinct delivery segments:

Cluster 1 (Fastest/Youngest): Characterized by the shortest delivery time (~ 73 min) and shortest average distance; contains the youngest agents. This represents optimal performance.
Cluster 0 (Slowest/Longest Distance): Defined by the longest average delivery time (~ 92 min) and longest average distance; tends to have older agents. This represents high-risk, delayed deliveries.
Cluster 3 (On-Time/High Rated): Delivery time is near the median (~ 125 min) and contains the highest rated agents. Performance is reliable and on-schedule.
Cluster 2 (Intermediate Delay): Slower than cluster 3 but significantly faster than cluster 0; moderate distance. This represents a potential operational bottleneck where delays are moderate.

Figure 6: K-Means Clustering on PCA (2D Projection)



The PCA 2D scatter plot shows four distinct clusters with clear separation, and red “X” markers indicate cluster centers. This visual evidence confirms the meaning segmentation of delivery performance.

7. Result Summary
The two supervised models proved effective in predicting on-time delivery. The SVM model's marginally superior performance, combined with the detailed insights from the clustering analysis, provides a comprehensive view of delivery dynamics.

Table 5: Supervised Model Performance Summary

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.79	0.79	0.79	0.79	0.883

SVM	0.80	0.80	0.80	0.80	0.883

Both supervised models achieved strong overall accuracy, demonstrating strong predictive capability. The clustering analysis provided descriptive insights into delivery dynamics, supporting performance segmentation and strategic resource allocation.

8. Conclusion and future works

This study successfully applied Logistic Regression, SVM, and K-Means Clustering with PCA to analyze and predict shipment delivery performance. The SVM model achieved the best predictive accuracy (80.0%) and equal AUC (0.883), confirming its robustness for classification tasks. K-Means Clustering provided complementary insights by revealing four distinct delivery patterns, helping to identify groups of agents or deliveries requiring strategic focus.

For future work, three main areas are recommended to enhance the model's practical utility:

1. **Real-Time Data Integration:** Incorporate real-time data such as live traffic and weather APIs for dynamic predictions that adapt to constantly changing conditions.
2. **Advanced Algorithms:** Apply more advanced ensemble algorithms such as Random Forest or XGBoost, or deep learning models like Neural Networks to potentially enhance accuracy and capture more complex non-linear relationships.
3. **Deployment:** Develop an interactive dashboard to monitor on-time delivery rates and cluster patterns in real time, allowing logistics managers to make immediate operational decisions.

Acknowledgments
This study was developed as part of an applied analytics project on predictive modeling for shipment delivery performance. I would like to thank the Kaggle community for providing open access to the Amazon Delivery dataset, which served as the foundation for this research. I also acknowledge the developers and contributors of open-source tools, such as Python, Pandas, scikit-learn, Matplotlib, and Seaborn, which enabled efficient data processing, visualization, and model development throughout this study.

References
Sujal S, (2024), Amazon Delivery Dataset. In
https://www.kaggle.com/datasets/sujalsuthar/amazon-delivery-dataset

Geeksforgeeks, (2025), Logistic Regression in Machine Learning

https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/

Geeksforgeeks, (2025), Support Vector Machine (SVM)
https://www.geeksforgeeks.org/machine-learning/support-vector-machine-algorithm/

Albers U, (2022), K-means Clustering and Principal Component Analysis in 10 Minuteshttps://towardsdatascience.com/k-means-clustering-and-principal-component-analysis-in-10-minutes-2c5b69c36b6b/

Appendix A. Code Snippets

Key functions for metrics calculation (Accuracy, Precision, Recall, F1-Score, AUC).
Example of the code used for training Logistic Regression and SVM.
K-Means Clustering with PCA.

Appendix B. Additional Figures and Tables

Training/ Testing performance (Tables 2 and 3)
K-Means Cluster Centers (Table 4)
ROC-AUC Comparison (Figure 4)
K-Means Clustering on PCA (Figures 5 and 6)