Inspiring Excellence

CSE422 Lab Project Report

Project Title: E-commerce Shipping Performance Analysis
Course: CSE422 – Data Mining & Machine Learning
Student Name: Md. Maruf Hassan, Md. Tasnim Muttaki
ID: 22301348, 21101216
Instructor: Rafeed Rahman, Nafiz Imtiaz Rafin
Semester: Fall 2025
Submission Date: 4 January 2026

Table of Contents

1. Introduction

This research discusses the on-time delivery performance of an e-commerce firm using historical delivery information. The primary purpose is to anticipate whether shipments will arrive on time for consumers by combining several customer, product, and logistical data.
This is a binary classification problem using the target variable. Furthermore, unsupervised learning (KMeans clustering) is used to find trends in delivery performance without labeling.
Motivation: Improve logistics, identify variables generating interruptions, and use machine learning techniques on real-world company information.

2. Dataset Description

Dataset Overview

- **Number of Features:** 12
- **Number of Data Points:** 10,000+
- **Target Variable:** Reached_on_time_Y_N
- **Problem Type:** Classification

Feature Types

- **Quantitative/Numerical Features:**
    - Customer_care_calls, Customer_rating, Cost_of_the_Product, Prior_purchases, Discount_offered, Weight_in_gms
- **Categorical Features:**
    - Warehouse_block, Mode_of_Shipment, Product_importance, Gender

Correlation Analysis

- The heatmap shows a mild correlation of some numerical features with on-time delivery.
- Categorical variables have weak linear correlation but are important for clustering and ML models.

Imbalanced Dataset

- Slight imbalance in Reached_on_time_Y_N.
- Bar chart visualization helps understand class distribution.

3. Dataset Pre-processing & Theory

Problem 1: Null / Missing Values

- **Theory:** ML algorithms cannot handle missing values natively. Imputation or row removal is required.
- **Observation:** No missing values present.

Problem 2: Categorical Variables

- **Theory:** Most ML models require numeric input. Categorical features must be converted to numbers.

Solution: Used LabelEncoder to convert categories (e.g., Gender, Mode_of_Shipment) to numeric codes.

```
data[col] = LabelEncoder().fit_transform(data[col].astype(str))
```

-

Problem 3: Feature Scaling

- **Theory:** Features with different scales can bias distance-based models (like K-Means) and gradient-based models (like Neural Networks).

Solution: Applied MinMaxScaler to normalize all features to [0,1].

```
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)
```

4. Dataset Splitting & Theory

- **Theory:** Splitting ensures models are trained and tested on separate data to evaluate generalization.

Method: 80% training, 20% testing, stratified by target to maintain class proportions.

```
X_train, X_test, y_train, y_test = train_test_split(
    X_scaled, y, test_size=0.2, random_state=42, stratify=y)
```

5. Model Training & Testing

Unsupervised Learning: K-Means Clustering

- **Theory:** K-Means partitions data into k clusters by minimizing within-cluster variance.
- **Application:** Used n_clusters=2 to detect clusters corresponding to on-time vs delayed shipments.

PCA: Applied Principal Component Analysis to reduce dimensions to 2 for visualization.

```
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)
```

- PCA helps visualize clusters and variance explained by the first two components.

Supervised Learning Models

1. **Logistic Regression** – Suitable for binary classification; predicts probability of on-time delivery.
2. **Decision Tree Classifier** – Creates hierarchical rules to classify instances; interpretable.
3. **Neural Network (MLP)** – Handles nonlinear relationships between features; requires scaling.

Training: Models are fitted on X_train and y_train. Predictions are generated on X_test.

Evaluation Metrics & Theory

- **Accuracy:** Proportion of correct predictions.
- **Precision:** Fraction of true positives among predicted positives.
- **Recall:** Fraction of true positives among actual positives.
- **AUC-ROC:** Area under ROC curve; measures model's ability to distinguish classes.
- **Confusion Matrix:** Shows TP, TN, FP, FN counts to evaluate model performance.

```
results = {
   "Logistic Regression": {...},
   "Decision Tree": {...},
   "Neural Network (MLP)": {...}
};
```

6. Model Selection / Comparison Analysis

Accuracy Comparison

- Neural Network performed best, followed by Logistic Regression and Decision Tree.
- Bar chart visualizes accuracy differences.

Confusion Matrices

- Provide insight into misclassification patterns.

ROC Curve Comparison

- Neural Network shows highest AUC (~0.92).

Figures:

- kmeans_pca.png – KMeans PCA visualization
- accuracy_comparison.png – Accuracy of models
- roc_curve_comparison.png – ROC curve for each model
- Confusion matrices for each model

7. **Conclusion & Theory**

In this study, the Neural Network and Logistic Regression models outperformed the other models in terms of capturing dataset trends. K-Means clustering revealed underlying patterns but is not appropriate for direct predictions. The challenges included a little class imbalance, category encoding, and the necessity for scalability, particularly for Neural Networks. For future enhancements, feature engineering, sophisticated models such as Random Forest and XGBoost, and pipeline automation are advised.