Republic of the Philippines
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES
College of Computer and Information Sciences
Sta. Mesa, Manila

In partial fulfillment of the requirements in the course

INTE - E2 - IT ELECTIVE 2
DATA MINING

FINAL PROJECT

<BSIT 3-2N / July 13, 2024>

Submitted by:

Name	Student Number
Aboy, Mark Angelo	2021-07488-MN-0
Belleca, Rochelle Anne	2022-18483-MN-1
Perez, Aldrich Chen Kenneth	2021-08551-MN-0
Rebulanan, Aaron Vergel	2021-08578-MN-0
Reyes, Cristian Jay	2021-08580-MN-0
Velasco, Ma. Cristina	2021-09204-MN-0

Submitted to:

MR. SEVERINO M. BEDIS JR.
Subject Instructor

Date Submitted:

July 13, 2024

Republic of the Philippines
POLYTECHNIC UNIVERSITY OF THE PHILIPPINES
College of Computer and Information Sciences
Sta. Mesa, Manila

I. TITLE

Analyzing On-Time Delivery in E-Commerce Shipping Data using Naïve Bayesian Algorithm

II. INTRODUCTION

Project Background:

The project involves applying data mining techniques to analyze an e-commerce dataset with the objective of gaining insights into customer behavior and shipping logistics. The data is sourced from an international e-commerce company that sells electronic products. The dataset, available on Kaggle, contains 10,999 observations of 12 variables, including customer ID, warehouse block, mode of shipment, customer care calls, customer rating, cost of the product, prior purchases, product importance, gender, discount offered, weight in grams, and whether the product was delivered on time.

The motivation for this project is to understand and predict the factors that influence timely delivery of products. By analyzing the data, the company aims to improve its logistics and customer satisfaction. The project will utilize the Naïve Bayesian algorithm, which is suitable for classification problems, to predict whether a shipment will reach on time based on various features.

Data mining is crucial in extracting valuable insights from large datasets. It helps businesses make informed decisions and improve their operations. This project demonstrates the application of data mining techniques to a real-world problem, showcasing its importance in business intelligence and predictive analytics.

III. OBJECTIVES:

To analyze the e-commerce shipping dataset using data mining techniques, specifically the Naïve Bayesian algorithm, to uncover insights and patterns that can improve the prediction of timely deliveries and optimize logistics strategies.

Research Questions:

1. What are the key factors that significantly influence the timely delivery of e-commerce products?
2. How accurately can the Naïve Bayesian algorithm predict the timely delivery of e-commerce shipments based on the given dataset?
3. How does the mode of shipment (Ship, Flight, Road) affect the likelihood of on-time delivery in e-commerce logistics?

Dataset Description:

The dataset contains the following variables:

- **ID**: Customer ID

- **Warehouse block**: Warehouse block (A, B, C, D, E)
- **Mode of shipment**: Shipment mode (Ship, Flight, Road)
- **Customer care calls**: Number of calls made by customers for inquiries
- **Customer rating**: Customer rating (1 to 5)
- **Cost of the product**: Cost of the product in USD
- **Prior purchases**: Number of prior purchases
- **Product importance**: Importance of the product (low, medium, high)
- **Gender**: Gender of the customer (Male, Female)
- **Discount offered**: Discount offered on the product
- **Weight in gms**: Weight of the product in grams
- **Reached on time**: Target variable indicating whether the product was delivered on time (1: No, 0: Yes)

Preprocessing Steps:

1. **Handling Missing Values**: Check for and handle any missing values in the dataset.
2. **Data Transformation**: Convert categorical variables into numerical format using techniques like one-hot encoding.
3. **Feature Selection**: Identify and select relevant features for the model.
4. **Data Normalization**: Normalize numerical features to ensure uniformity.

Challenges Encountered:

- Handling categorical data and transforming it into a suitable format for the Naïve Bayesian algorithm.
- Balancing the dataset if there is an imbalance between the classes of the target variable.

IV. METHODOLOGY:

The Naïve Bayesian algorithm is chosen for this project due to its simplicity and effectiveness in classification tasks. It is particularly useful for predicting categorical outcomes based on multiple features.

Naïve Bayesian is ideal for this dataset as it assumes feature independence and performs well with categorical data. It is computationally efficient and provides probabilistic predictions, which are valuable for decision-making.

Alternative Approaches Considered

- Decision Trees: Considered for their interpretability but not chosen due to potential overfitting.
- Support Vector Machine (SVM): Considered for its robustness but not chosen due to computational complexity.

V. EVALUATION PLAN:

Performance Metrics:

- **Accuracy**: Proportion of correctly predicted instances.
- **Precision**: Proportion of true positive predictions among all positive predictions.
- **Recall**: Proportion of true positive predictions among all actual positives.
- **F1 Score**: Harmonic mean of precision and recall.

Validation Methods:

- **Train-Test Split**: Splitting the dataset into training and testing sets to evaluate model performance.
- **Cross-Validation**: Using k-fold cross-validation to ensure robustness and prevent overfitting.

Expected Outcomes:

- A trained Naïve Bayesian model that can predict whether a shipment will be delivered on time.
- Insights into the key factors affecting timely delivery.

VI. SCREENSHOTS OF THE RESULT

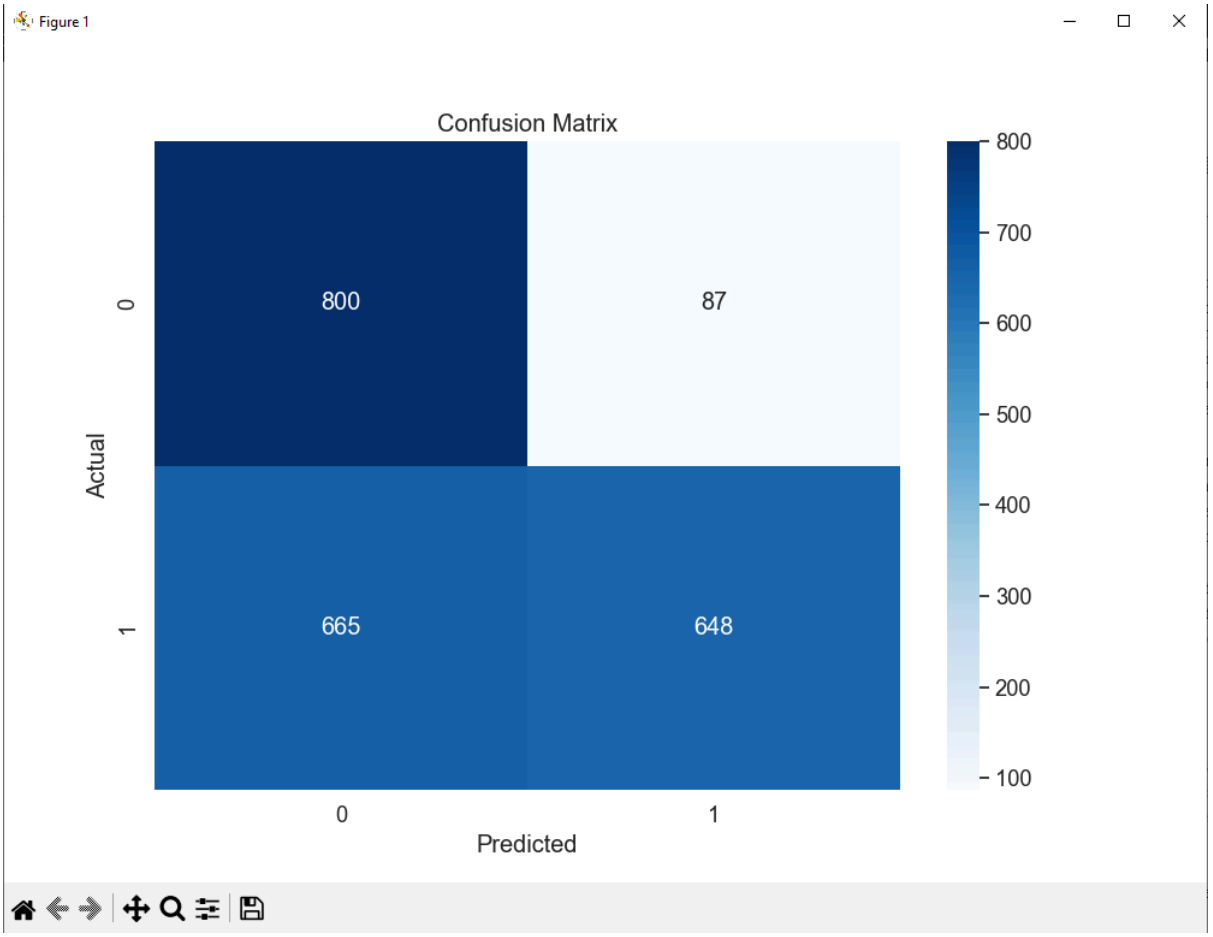
```
   ID Warehouse_block  ... Weight_in_gms  Reached.on.Time_Y.N
0   1               D  ...          1233                    1
1   2               F  ...          3088                    1
2   3               A  ...          3374                    1
3   4               B  ...          1177                    1
4   5               C  ...          2484                    1

[5 rows x 12 columns]
Optimal Prior: 0.2680000000000001
Accuracy: 0.6581818181818182
Classification Report:
              precision    recall  f1-score   support

           0       0.55      0.90      0.68       887
           1       0.88      0.49      0.63      1313

    accuracy                           0.66      2200
   macro avg       0.71      0.70      0.66      2200
weighted avg       0.75      0.66      0.65      2200


Process finished with exit code 0
```



VII. SOURCE CODES (make a Google Drive link then paste it here)
https://drive.google.com/drive/folders/1SVPe_p2kZSvtMXs3bF_fN_wRdMDVrq3q?usp=sharing

Naive Bayes