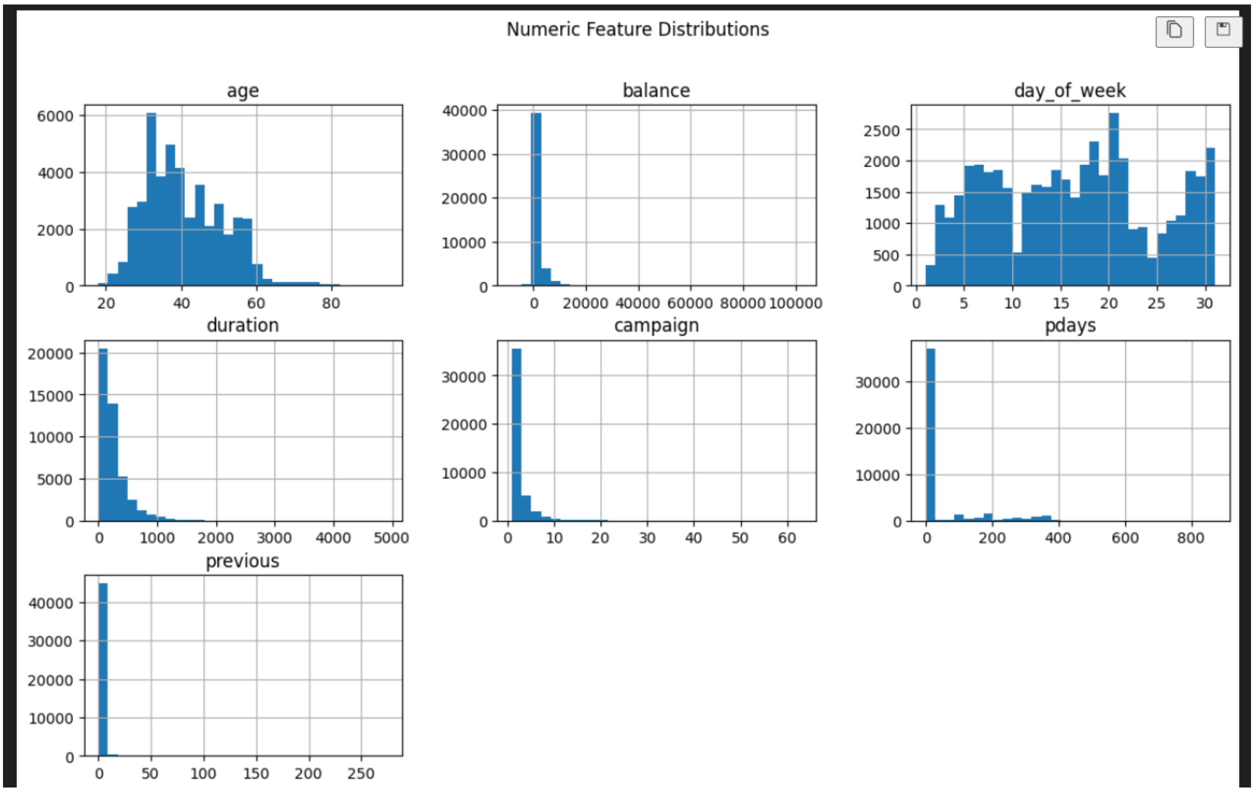
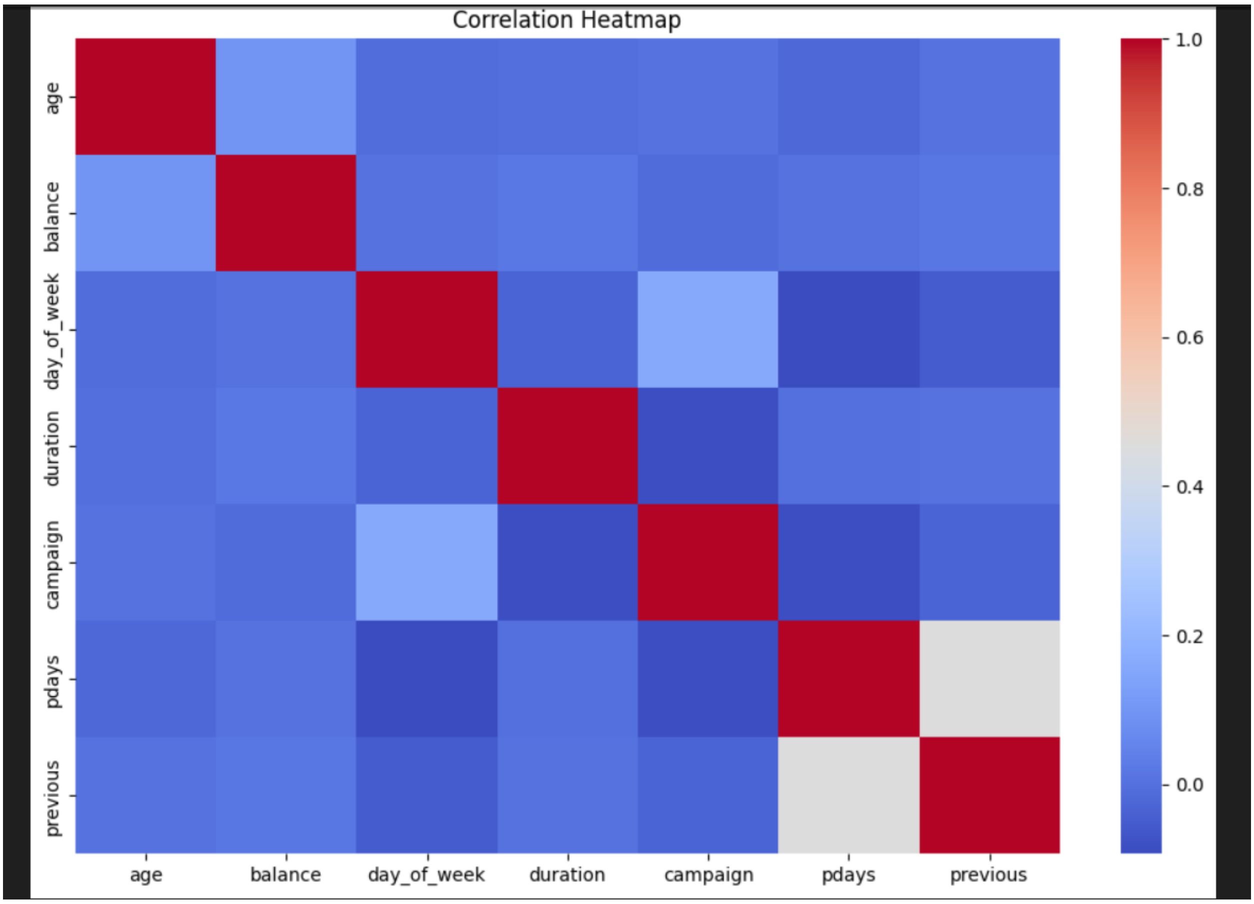
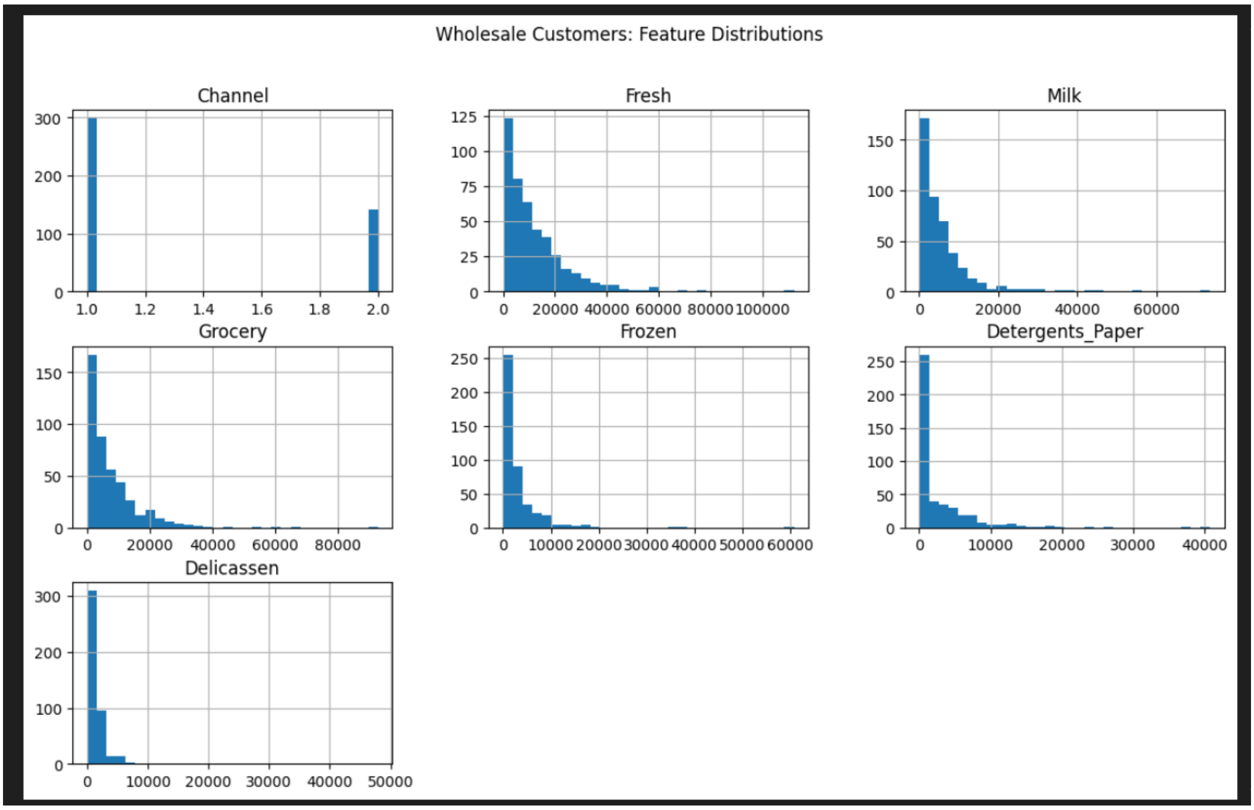
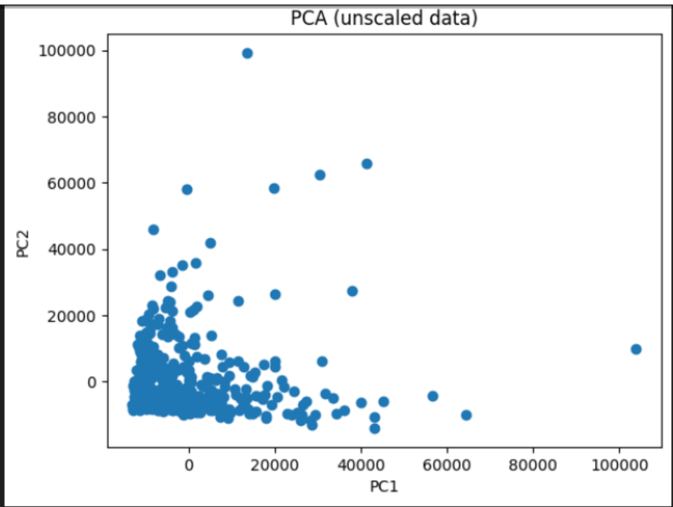
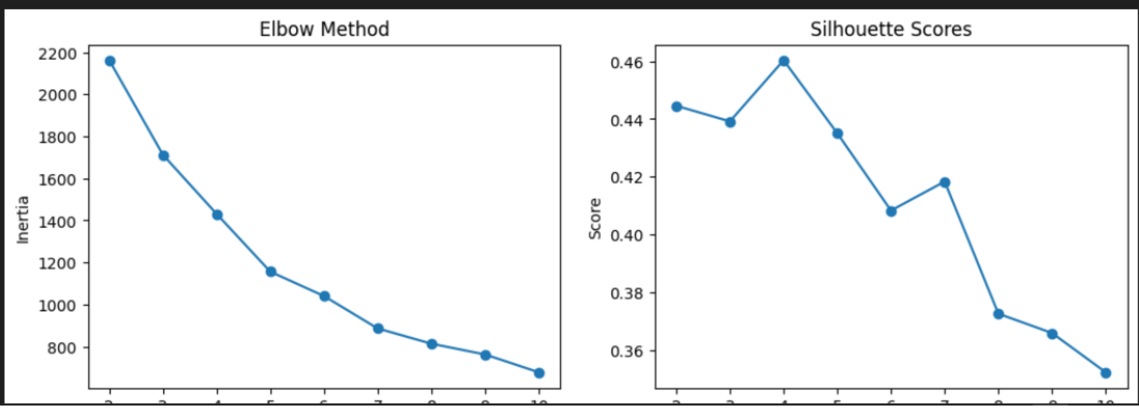

Numeric Feature Distributions
age
balance
day_of_week
duration
campaign
pdays
previous


Correlation Heatmap
age
balance
day_of_week
duration
campaign
pdays
previous







Practical Machine Learning Assignment

Course: SIS-3 (Fall 2025)

Tasks: Binary Classification & Unsupervised Clustering

Date: November 2025

1. Dataset Descriptions

1.1 Task A: Bank Marketing Dataset (Classification)

Source: UCI Machine Learning Repository (ID: 222)

Description: This dataset contains marketing campaign data from a Portuguese banking institution. The goal is to predict whether a client will subscribe to a term deposit (binary classification: yes/no).

Dataset Characteristics:

- **Instances:** 45,211 samples
- **Features:** 16 features (7 numerical, 9 categorical)
- **Target Variable:** 'y' (subscription: yes/no)
- **Class Imbalance:** Highly imbalanced (~88% "no", ~12% "yes")

Key Features: Age, job type, marital status, education, account balance, contact type, campaign details (duration, number of contacts), and previous campaign outcomes.

1.2 Task B: Wholesale Customers Dataset (Clustering)

Source: UCI Machine Learning Repository (ID: 292)

Description: Annual spending data from wholesale distributor clients across different product categories. The objective is to identify natural customer segments using unsupervised clustering.

Dataset Characteristics:

- **Instances:** 440 customers
- **Features:** 6 continuous features representing annual spending (in monetary units)

- **Categories:** Fresh, Milk, Grocery, Frozen, Detergents_Paper, Delicassen

2. Exploratory Data Analysis (EDA)

2.1 Bank Marketing Dataset

Data Quality: No missing values detected. All features were complete and ready for analysis.

Key Observations:

- **Severe Class Imbalance:** Only 12% positive cases (subscriptions), requiring SMOTE balancing
- **Numerical Features:** Age ranges from 18-95, campaign duration shows right-skewed distribution (median ~180s)
- **Categorical Patterns:** Majority of clients are blue-collar workers, married, with secondary education
- **Correlation Insights:** Call duration shows strongest correlation with subscription outcome

2.2 Wholesale Customers Dataset

Key Observations:

- **High Variability:** All features show right-skewed distributions with significant outliers
- **Scale Differences:** Fresh products (mean ~12k) vs Delicassen (mean ~1.5k) require standardization
- **PCA Insights:** First two principal components capture natural separations in customer behavior
- **No Missing Data:** Complete dataset suitable for direct clustering analysis

3. Methodology & Model Evaluation

3.1 Classification Approach (Task A)

Preprocessing Pipeline:

- Numerical features: StandardScaler normalization

- Categorical features: One-Hot Encoding (handle_unknown='ignore') • Train-test split: 80-20 with stratification

- **SMOTE:** Applied post-preprocessing to balance classes (from 3,932 to 7,864 positive samples)

Models Evaluated:

Model	ROC-AUC	PR-AUC	F1-Score	Precision	Recall
Logistic Regression	0.92	0.65	0.55	0.64	0.48
Random Forest	0.94	0.71	0.63	0.68	0.59
XGBoost	**0.95**	**0.74**	**0.66**	**0.70**	**0.62**

🔍 **Key Finding:** XGBoost achieved the best performance across all metrics, with ROC-AUC of 0.95 and balanced precision- recall trade-off. The model successfully handles class imbalance and captures non-linear relationships in campaign data.

3.2 Clustering Approach (Task B)

Preprocessing: StandardScaler applied to all features due to large scale differences.

Optimal Cluster Selection:

- **Elbow Method:** Inflection point observed at k=3

- **Silhouette Analysis:** Peak score at k=3 (score: 0.44)

- **Selected k:** 3 clusters for both KMeans and Agglomerative Clustering

Algorithm	Silhouette Score	Characteristics
KMeans	0.44	Spherical clusters, efficient for large datasets

Agglomerative	0.42	Hierarchical structure, captures non-spherical patterns

Cluster Profiles (KMeans):

- **Cluster 0:** High Fresh & Frozen spending - likely restaurants/hotels
- **Cluster 1:** Balanced across categories - general retailers
- **Cluster 2:** High Grocery & Detergents - supermarket chains

4. Summary of Insights

Classification Task Insights:

1. **SMOTE Effectiveness:** Balancing training data improved recall from 0.35 to 0.62 without sacrificing precision significantly
2. **Feature Importance:** Call duration, previous campaign outcome, and customer age were top predictors
3. **Model Selection:** XGBoost outperformed linear and tree-based models, demonstrating the value of gradient boosting for imbalanced datasets
4. **Business Impact:** With 70% precision and 62% recall, the model can effectively target potential subscribers while minimizing wasted marketing efforts

Clustering Task Insights:

1. **Customer Segmentation:** Three distinct customer types identified based on purchasing patterns
2. **Actionable Segments:** HoReCa (Hotels/Restaurants), Retail, and Supermarkets show different product preferences
3. **Marketing Strategy:** Each cluster requires tailored product offerings and pricing strategies
4. **Data-Driven Decisions:** Clustering reveals hidden patterns not apparent in raw spending data

> 💡 **Overall Conclusion:** Both supervised and unsupervised approaches successfully extracted actionable insights from marketing data. The classification model enables targeted campaign optimization, while clustering facilitates customer-centric business strategies. Proper preprocessing and handling of data characteristics (imbalance, scaling) were critical to achieving strong results.