Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan.

Hassan Rais	2022212
M. Hamza Mehmood Zaidi	2022379

Course: DS341 – Data Mining
Instructor: Dr Ayaz Umer

Project Report: Customer Behavior Analytics & Predictive Insights

Dataset Overview

Dataset: Online Shoppers Purchasing Intention Dataset
Source: UCI Machine Learning Repository
Number of Records: 12,330
Number of Features: 18 features + 1 targets (Revenue)

```
🔍 Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 19 columns):
```

Problem Statement: To analyze customer behavior data to uncover buying patterns, predict purchases, and segment customers using data mining techniques. The goal is to help online businesses optimize marketing strategies and improve conversion rates.

3. Preprocessing Details

Handling Missing Values: No missing values were found in the dataset.

Feature Encoding:

'Month' and 'VisitorType' encoded using LabelEncoder.

'Weekend' and 'Revenue' converted to binary (0 = No, 1 = Yes).

Transformations:

Scaled all features using `StandardScaler` for classification and clustering models.

4. Exploratory Data Analysis

Key Statistics:

Revenue is highly imbalanced (~85% No Purchase, ~15% Purchase)

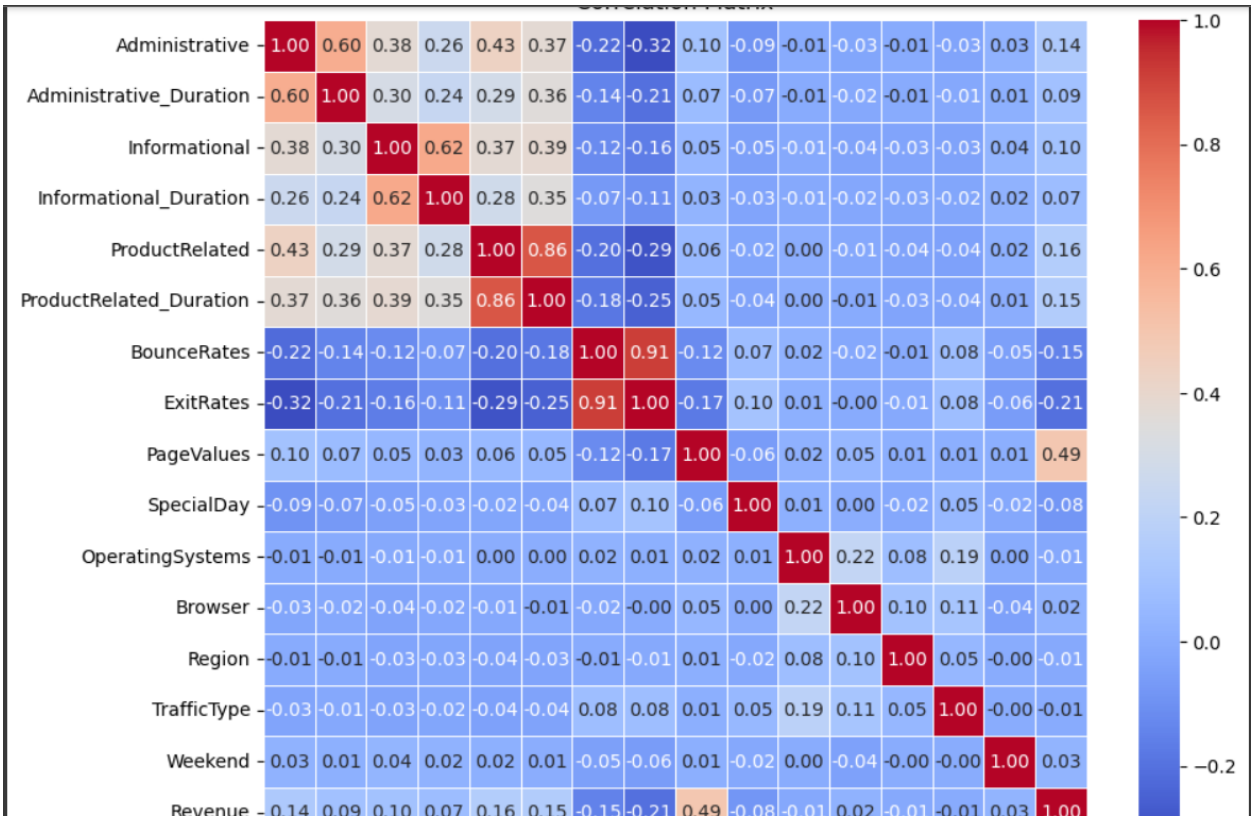
High PageValues and ProductRelated_Duration are positively correlated with purchases.

Visualizations:

Distribution plots for numeric variables

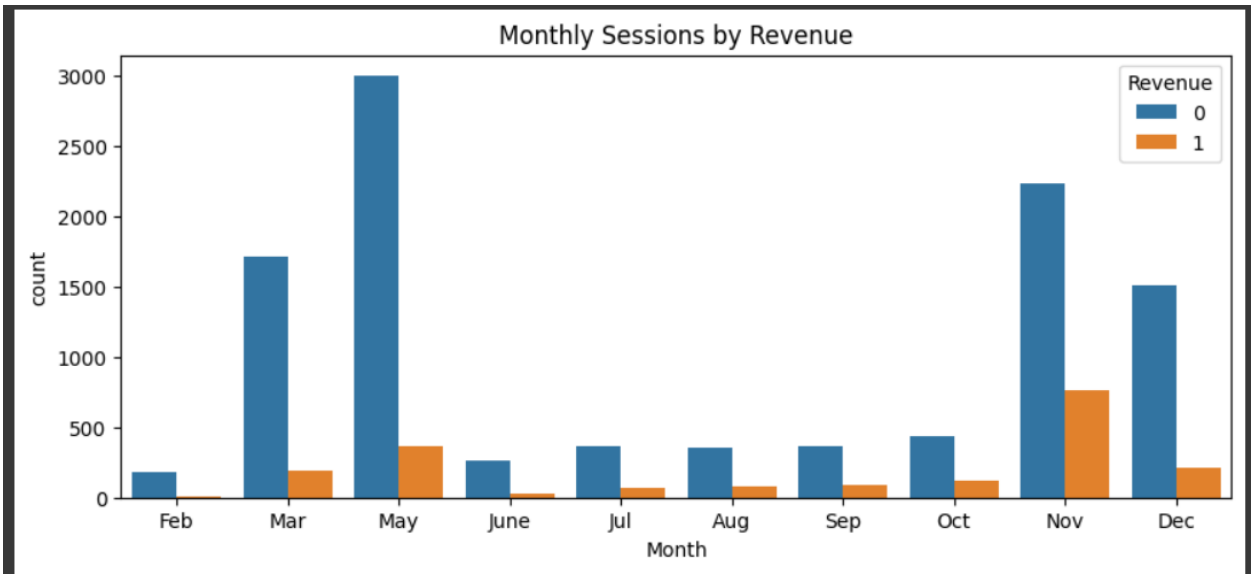
Countplots for categorical features

Heatmap showing feature correlations

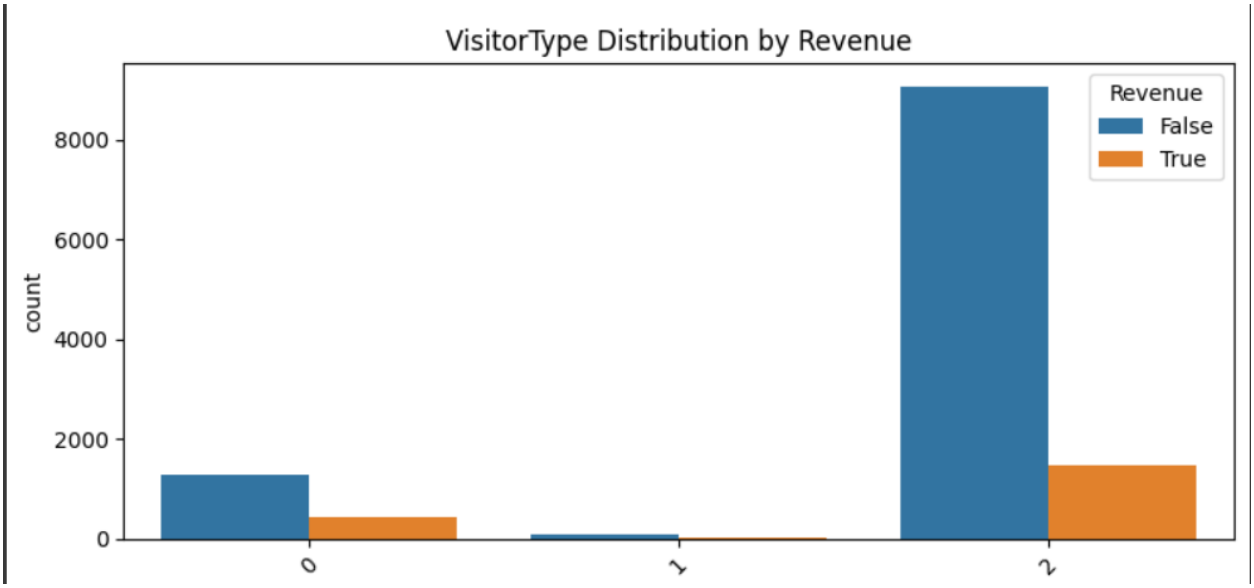


Observed Trends:

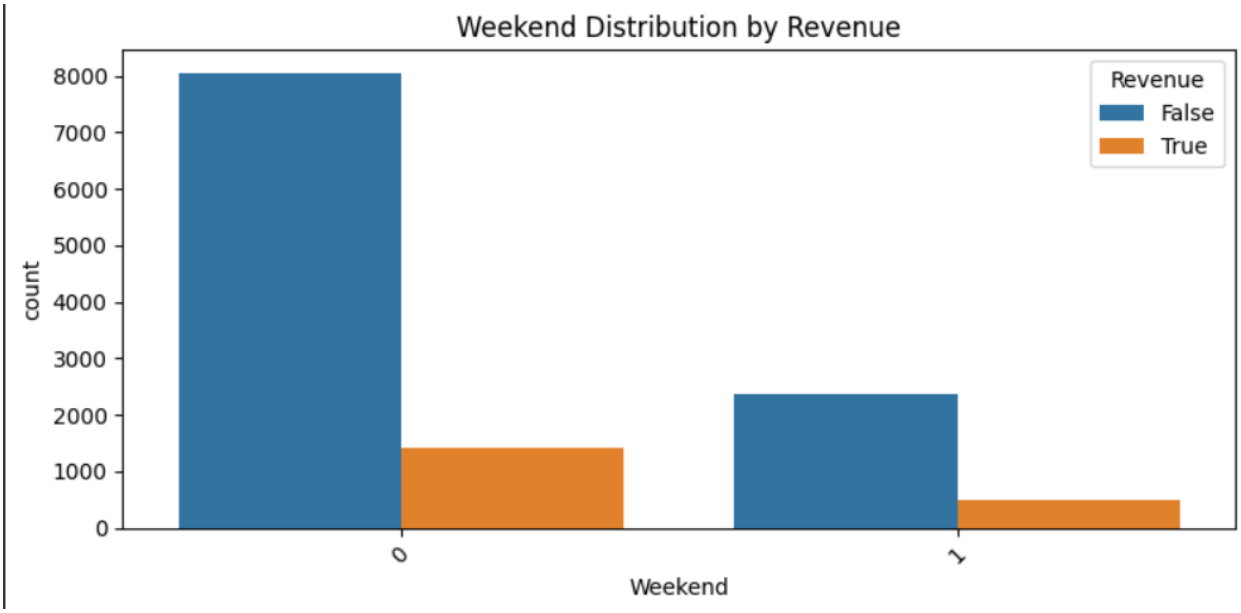
Most purchases occur in Nov–Dec



New visitors with longer durations tend to convert



Returning visitors on weekdays usually don't purchase



5. Association Rule Mining

Techniques Applied: Apriori, fpgrowth

Thresholds: min_support = 0.05, min_confidence = 0.5

Strong Rules:

(PageValueGroup_NoValue, VisitorType_Returning_Visitor, Weekend_Weekday) → Revenue_NoPurchase

(PageValueGroup_HasValue, VisitorType_New_Visitor) → Revenue_Purchase

Interpretation:

Returning users with no high-value pages almost never purchase.

New users who visit value-rich pages show strong purchase intent.

```
🔹 Top 5 Apriori Rules:
                                          antecedents  \
90  (Weekend_Weekday, PageValueGroup_NoValue, Visi...
40  (PageValueGroup_NoValue, VisitorType_Returning...
45          (Weekend_Weekday, PageValueGroup_NoValue)
84  (Weekend_Weekday, VisitorType_New_Visitor, Rev...
6                          (PageValueGroup_NoValue)

                consequents   support  confidence      lift
90      (Revenue_NoPurchase)  0.504461    0.969602  1.147112
40      (Revenue_NoPurchase)  0.644201    0.966537  1.143485
45      (Revenue_NoPurchase)  0.579562    0.963722  1.140155
84  (PageValueGroup_NoValue)  0.070073    0.962138  1.235746
6       (Revenue_NoPurchase)  0.748581    0.961458  1.137477

🔸 Top 5 FP-Growth Rules:
                                           antecedents  \
37   (Weekend_Weekday, PageValueGroup_NoValue, Visi...
7    (PageValueGroup_NoValue, VisitorType_Returning...
24          (Weekend_Weekday, PageValueGroup_NoValue)
115  (Weekend_Weekday, VisitorType_New_Visitor, Rev...
2                          (PageValueGroup_NoValue)

                consequents   support  confidence      lift
37      (Revenue_NoPurchase)  0.504461    0.969602  1.147112
7       (Revenue_NoPurchase)  0.644201    0.966537  1.143485
24      (Revenue_NoPurchase)  0.579562    0.963722  1.140155
115 (PageValueGroup_NoValue)  0.070073    0.962138  1.235746
2       (Revenue_NoPurchase)  0.748581    0.961458  1.137477
```

6. Classification Models

Models Used:

Decision Tree Classifier

Gaussian Naive Bayes

K-Nearest Neighbors (K=5, optimized at K=7 via elbow method)

Decision Tree Report:

	precision	recall	f1-score	support
No Purchase	0.87	0.84	0.86	2084
Purchase	0.27	0.33	0.30	382
accuracy			0.76	2466
macro avg	0.57	0.58	0.58	2466
weighted avg	0.78	0.76	0.77	2466

Naive Bayes Report:

	precision	recall	f1-score	support
No Purchase	0.87	0.89	0.88	2084
Purchase	0.30	0.26	0.28	382
accuracy			0.79	2466
macro avg	0.59	0.58	0.58	2466
weighted avg	0.78	0.79	0.79	2466

K-Nearest Neighbors Report:

	precision	recall	f1-score	support
No Purchase	0.86	0.96	0.91	2084
Purchase	0.40	0.13	0.19	382
accuracy			0.83	2466
macro avg	0.63	0.55	0.55	2466
weighted avg	0.79	0.83	0.80	2466

Performance (Class: Purchase):

Model	Precision	Recall	F1-Score
Decision Tree	0.27	0.33	0.30
Naive Bayes	0.30	0.26	0.28
KNN (k=5)	0.34	0.14	0.20

Conclusion:

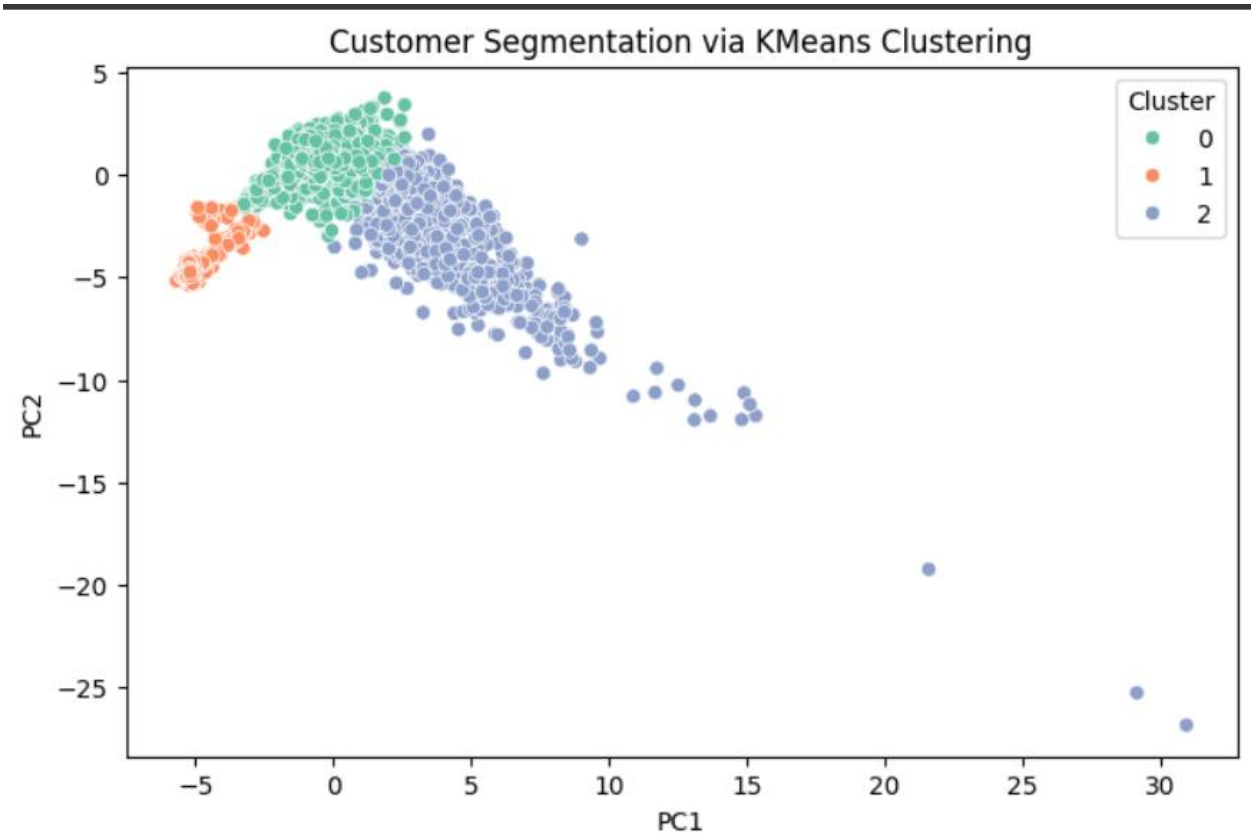
Decision Tree provided the best balance.

KNN required parameter tuning but underperformed due to class imbalance.

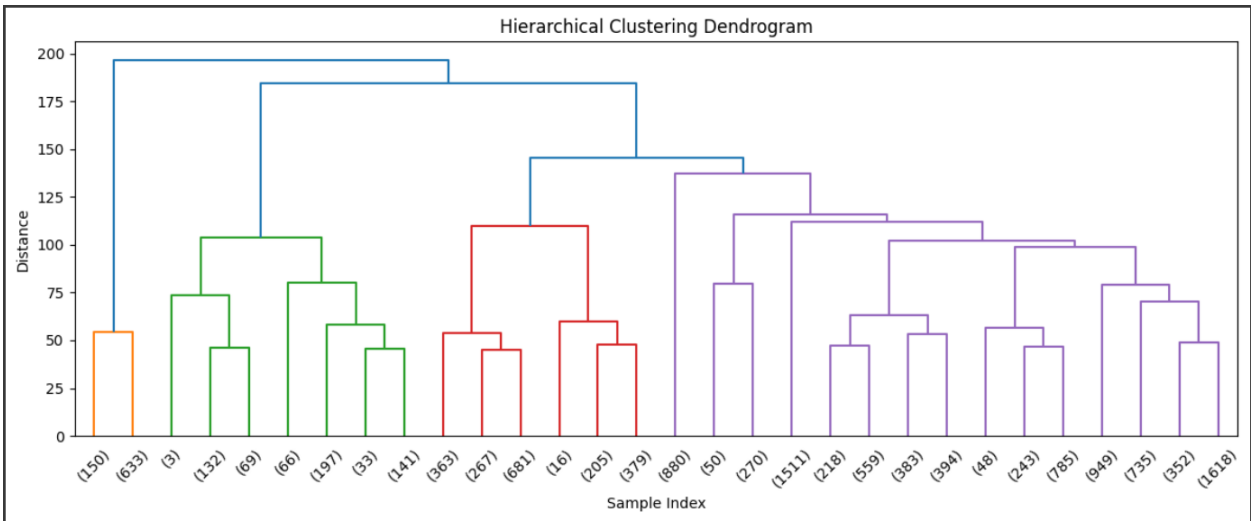
7. Clustering & Customer Segmentation

Clustering Methods:

K-Means (K=3)



Hierarchical Clustering



Cluster Insights:

Cluster 0: Low interaction, no purchase behavior

Cluster 1: High engagement, likely purchasers

Cluster 2: Long sessions, mixed outcomes

Visualizations:

PCA used to reduce dimensions and visualize clusters

Dendrogram generated from hierarchical clustering

Evaluation:

KMeans segmentation was intuitive and aligned with revenue patterns

Hierarchical clustering helped validate natural groupings

8. Insights & Recommendations

Business Recommendations:

1. Personalize content for returning weekday visitors.
2. Retarget high-engagement clusters with focused promotions.
3. Use exit-intent offers on sessions without product interaction.
4. Enhance product pages to increase PageValue early.
5. Schedule campaigns around peak conversion months (e.g., Nov).

Business Value:

- Segmentation improves targeted marketing
- Predictive modeling allows early churn prevention
- Rules can automate content/promo delivery

9. Challenges & Reflections

Class Imbalance: Required stratified sampling and evaluation beyond accuracy.

Leakage: PageValues caused inflated model performance. Addressed by excluding classification.

Model Sensitivity: KNN required careful K tuning; elbow method used.

10. Tools Used

Languages: Python

Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, mlxtend, scipy

Team Contribution Table

Member	Contribution
Hassan Rais	Data preprocessing, clustering, association rules
M. Hamza Zaidi	Classification models EDA, evaluation metrics