Assessment Report

on

"Market Basket Analysis"

submitted as partial fulfillment for the award of

BACHELOR OF TECHNOLOGY DEGREE

SESSION 2024-25

in

CSE AIML

By

Yash Jangid (202401100400216)

KIET Group of Institutions, Ghaziabad

April, 2025

Introduction

In the highly competitive landscape of retail, understanding consumer purchasing patterns is paramount. Market basket analysis, which investigates associations between items purchased together, provides valuable insights into product affinities and customer behavior at the point of sale. By leveraging both supervised and unsupervised machine learning techniques, retailers can not only predict outcomes based on historical data but also uncover latent groupings within product assortments, leading to more informed decisions in inventory planning, layout optimization, and targeted promotions.

This report delves into a two-pronged analytical approach. First, we demonstrate how classification methods—if ground-truth labels are available—can be evaluated using standard performance metrics (accuracy, precision, recall) and visualized through confusion matrix heatmaps. Second, recognizing that real-world transaction data may lack explicit labels, we apply unsupervised segmentation techniques on aisle descriptions to reveal semantic clusters of product categories. We employ TF-IDF to convert textual aisle names into numerical vectors, cosine similarity to measure pairwise likeness, and KMeans clustering combined with PCA for grouping and visualization. The findings aim to showcase a robust framework for both classification evaluation and aisle segmentation, equipping stakeholders with actionable insights to enhance retail strategy.

In summary, this analysis offers:

1. **A rigorous classification evaluation pipeline** that quantifies model performance through widely recognized metrics and interpretable heatmaps.
2. **An unsupervised clustering methodology** that exploits text analytics for aisle grouping, revealing natural segments in product categorization.
3. **Comprehensive visualizations** and concise interpretations to facilitate decision-making in retail operations and marketing.

Methodology

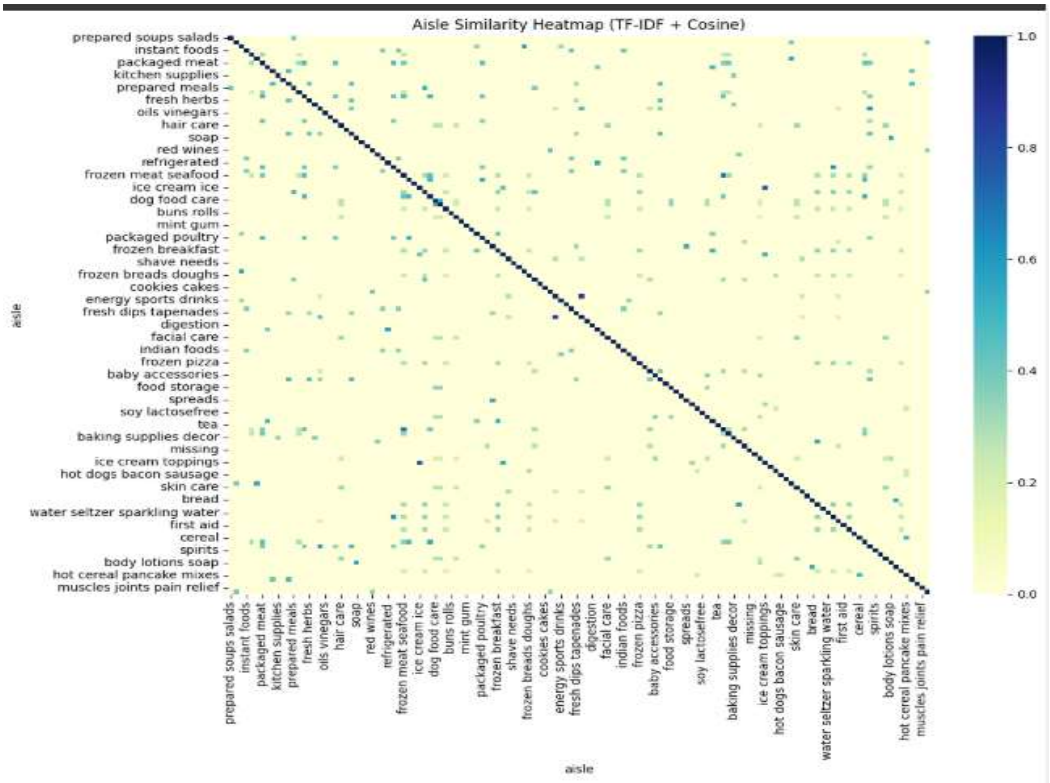
1. **Data Loading & Exploration**:
    - Load the provided Market Basket Analysis.csv containing aisle_id and aisle fields.
    - Inspect data types, count missing values, and examine basic distributions of aisle names.
2. **Classification Workflow (Conditional)**:
    - Check for the existence of true_label and predicted_label columns.
    - If present, compute a confusion matrix, visualize it with a heatmap, and calculate accuracy, precision, and recall (weighted average).
3. **Unsupervised Segmentation & Clustering**:
    - Transform aisle names into TF-IDF vectors to capture semantic information.
    - Compute cosine similarity to quantify pairwise textual similarity between aisles.
    - Generate a heatmap of the similarity matrix for visual inspection.
    - Apply KMeans clustering to

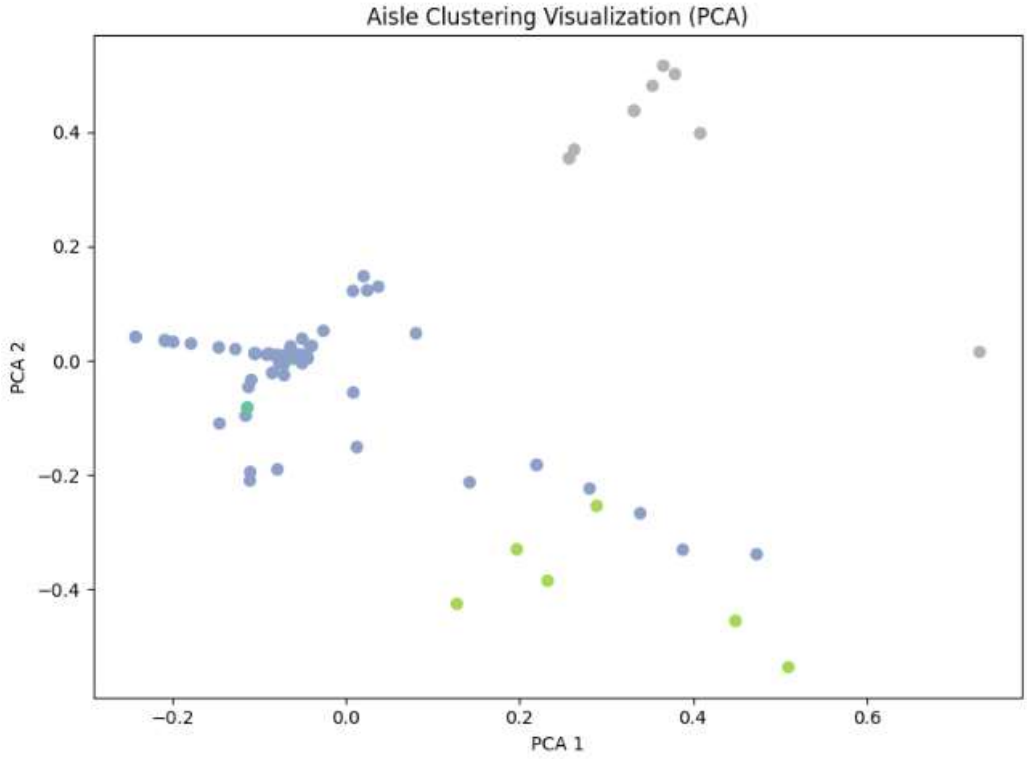
CODE

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.metrics import confusion_matrix, accuracy_score, precision_score, recall_score

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.metrics.pairwise import cosine_similarity

from sklearn.cluster import KMeans

from sklearn.decomposition import PCA


# Load data

df = pd.read_csv("10. Market Basket Analysis.csv")


# Check for classification columns

has_classification = {'true_label', 'predicted_label'}.issubset(df.columns)


if has_classification:

    # Classification block

    y_true = df['true_label']

    y_pred = df['predicted_label']

    labels = sorted(set(y_true) | set(y_pred))

    cm = confusion_matrix(y_true, y_pred, labels=labels)

    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=labels, yticklabels=labels)

    plt.title('Confusion Matrix')

    plt.show()

    print("Accuracy:", accuracy_score(y_true, y_pred))

    print("Precision (weighted):", precision_score(y_true, y_pred, average='weighted'))

    print("Recall (weighted):", recall_score(y_true, y_pred, average='weighted'))

else:
```

```
# Clustering block

tfidf = TfidfVectorizer()

X = tfidf.fit_transform(df['aisle'])

sim = cosine_similarity(X)

sns.heatmap(sim, cmap="YlGnBu", xticklabels=df['aisle'], yticklabels=df['aisle'])

plt.title('Aisle Similarity Heatmap')

plt.show()

kmeans = KMeans(n_clusters=5, random_state=42)

clusters = kmeans.fit_predict(X)

df['cluster'] = clusters

pca = PCA(n_components=2)

coords = pca.fit_transform(X.toarray())

plt.scatter(coords[:,0], coords[:,1], c=clusters, cmap='Set2')

for i, name in enumerate(df['aisle']): plt.annotate(name, coords[i])

plt.title('Aisle Clusters (PCA)')

plt.show()

print(df[['aisle','cluster']])
```

OUTPUT


Aisle Similarity Heatmap (TF-IDF + Cosine)
aisle
prepared soups salads
instant foods
packaged meat
kitchen supplies
prepared meals
fresh herbs
oils vinegars
hair care
soap
red wines
refrigerated
frozen meat seafood
ice cream ice
dog food care
buns rolls
mint gum
packaged poultry
frozen breakfast
shave needs
frozen breads doughs
cookies cakes
energy sports drinks
fresh dips tapenades
digestion
facial care
indian foods
frozen pizza
baby accessories
food storage
spreads
soy lactosefree
tea
baking supplies decor
missing
ice cream toppings
hot dogs bacon sausage
skin care
bread
water seltzer sparkling water
first aid
cereal
spirits
body lotions soap
hot cereal pancake mixes
muscles joints pain relief
1.0
0.8
0.6
0.4
0.2
0.0


Aisle Clustering Visualization (PCA)
PCA 2
PCA 1
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6

References

- W3Schools. (n.d.). Python Data Analysis. Retrieved from https://www.w3schools.com/python/pandas