CS465 - Machine Learning

Saudi Arabia Supermarket Data Analysis Report

Nouf Almojel
222410007@psu.edu.sa

PREPARED FOR :
Dr. Abrar Wafa

PRINCE SULTAN UNIVERSITY

Table of Contents

Project Code

Phase 1: Introduction

This report presents an analysis of the supermarket data from Tamimi Markets, which operating within the Saudi Arabian market. The study aims to provide a comprehensive market understanding by preparing data and conducting analysis to engineer features for product pricing research and distribution pattern investigation.

Phase 2: Data Preparation

First, a dataset of 1,220 supermarket products collected from Tamimi Markets was used for the analytical work. Below findings are::

- Complete Data: All columns did not contain any missing values.
- No Duplicates: No duplicate records were present.
- Standardization:To make reliable measurement and comparison possible, all products were standardized to grams per milliliter units.
- Manual Intervention: One non-automated product size required human intervention for standardization.

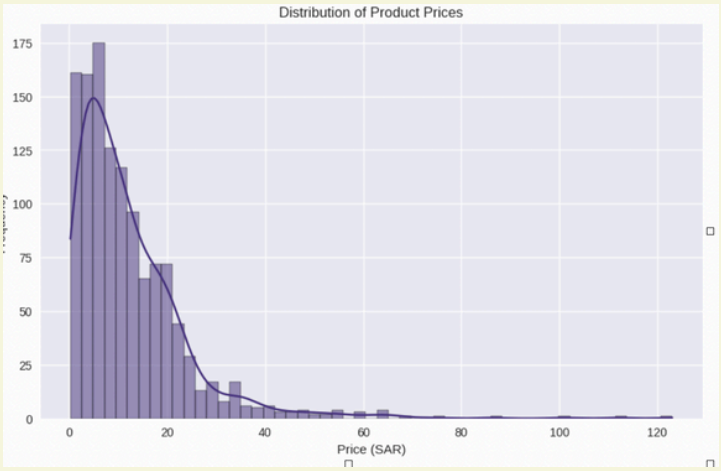
DataSet file: tamimimarkets.csv

Phase 3: Exploratory Data Analysis [EDA]

Price distribution histogram

This image shows the distribution of product prices ranging from 0.3 to 123 SAR

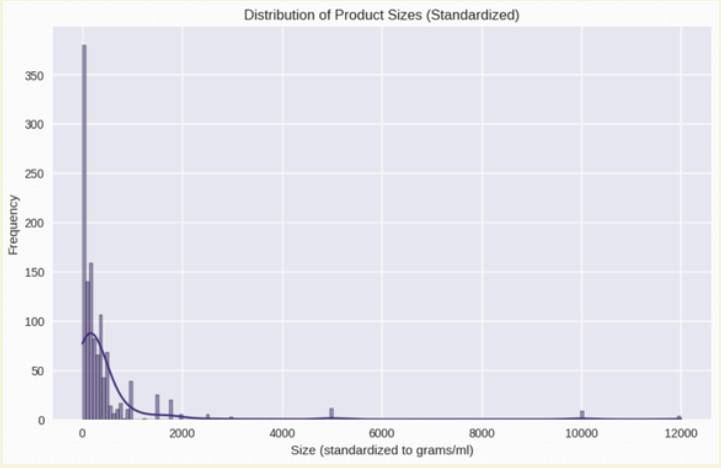
The price analysis shows most products fall in the lower price ranges. The median price is 9.5 SAR while the average is slightly higher at 12.30 SAR, indicating some higher-priced outliers. The standard deviation of 11.85 SAR confirms significant price variation.



Product size distribution

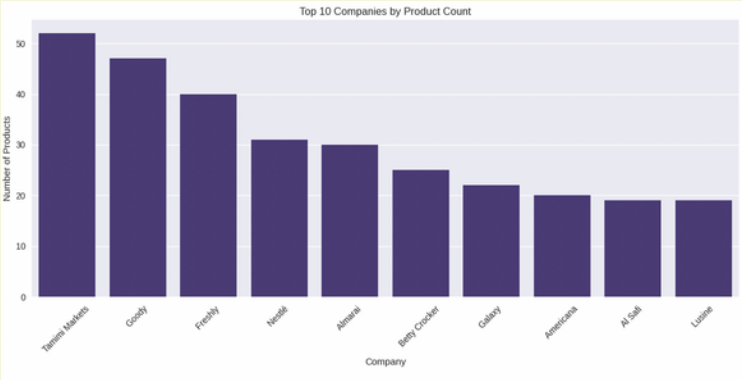
This image displays the standardized sizes of products

The size standardization allows for meaningful comparison between products sold in different units (grams, milliliters, etc.).



Top companies bar chart

This image shows the companies with the most products in the dataset Using the dataset analyzed 304 unique companies were found, with some being represented much more than the others.

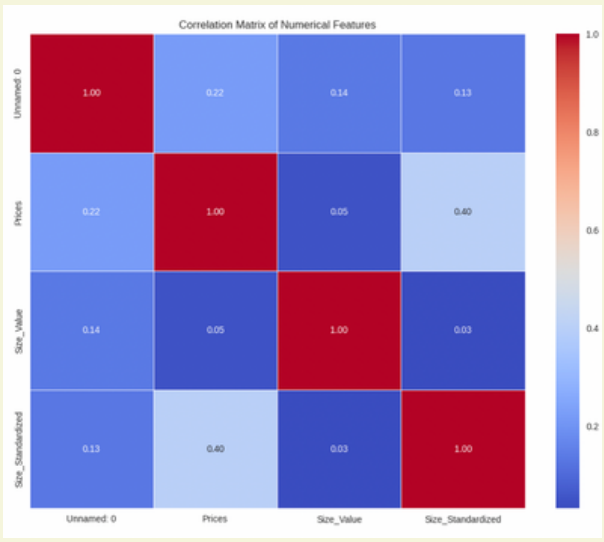


Average price by company

This image displays companies with the highest average product prices Volvic has the most expensive product (Mineral Water at 123 SAR), while Tamimi Markets' own brand has the cheapest product (Bottled drinking water at 0.3 SAR).

Correlation matrix

This image shows relationships between numerical variables



Phase 4: Feature Engineering

We created several new features for the analysis purpose :

1. **Price Per Unit:** The price per unit is measured as price divided by the standardized size. This shows a correlation of 0.2065 with price.
2. **Premium Product Flag:** Products priced above 16.95 SAR (75th percentile) were labeled as premium. This identified 302 premium products and showed the strongest correlation with price (0.7350).
3. **Product Categories:** Items were classified based on keywords in product names:
   - Other: 891 products
   - Dairy: 128 products
   - Meat: 71 products
   - Beverages: 70 products
   - Grains/Bakery: 60 products
4. **Feature Importance:** Most strongly correlated with price were:
   - Is_Premium: 0.7350
   - Size_Standardized: 0.4009
   - Unnamed: 0: 0.2205
   - Price_Per_Unit: 0.2065
   - Size_Value: 0.0515

```
Phase 4: Feature Engineering

Price per unit feature created.
Premium product flag created (threshold: 16.95 SAR)

Product Categories Created:
Product_Category
Other            891
Dairy            128
Meat              71
Beverages         70
Grains/Bakery     60
Name: count, dtype: int64

Feature Correlation with Price:
Prices              1.000000
Is_Premium          0.734973
Size_Standardized   0.400923
Unnamed: 0          0.220491
Price_Per_Unit      0.206517
Size_Value          0.051486
Name: Prices, dtype: float64

Top Features for Prediction (based on correlation):
2. Is_Premium: 0.7350
3. Size_Standardized: 0.4009
4. Unnamed: 0: 0.2205
5. Price_Per_Unit: 0.2065
6. Size_Value: 0.0515


Analysis Summary:

1. Data Preparation:
   - Dataset: 1220 records with 12 features
   - Preprocessed size information and created standardized size measurements
   - Handled any missing values and duplicates

2. Exploratory Data Analysis:
   - Analyzed price distribution (avg: 12.30 SAR)
   - Examined top 10 companies by frequency
   - Explored relationship between price and other features

3. Feature Engineering:
   - Created price per unit feature
   - Added premium product flag (302 premium products)
   - Created 5 product categories
```

Phase 5: Model Building

Data Preparation

Our dataset went through preparation steps before model development involved:

- Selecting relevant features for prediction
- A single vacant value underwent treatment during missing values handling process (1 value was addressed).
- The data was divided into two sections for training purposes (976 samples) and testing operations (244 samples).
- A standardization process applied to all variables for maintaining equivalent comparison possibilities.

Models Implemented

Our project utilized four machine learning algorithms during the implementation stage as per requirements.

- Linear Regression serves as a basic algorithm which establishes price-feature relations using linear mathematical expressions.
- The Decision Tree represents a tree-like model which obtains decisions through applying feature thresholds.
- The Support Vector Machine (SVM) algorithm determines optimal boundary lines which split price ranges.
- Neural Network represents a mathematical structure that identifies complex non-linear feature interrelations between various variables.

Training Results

Most of the model predictions during the first training phase exhibited promising findings.

Model	Training R^2 Score
Linear Regression	0.6510
Decision Tree	1.0000
SVM	0.5585
Neural Network	0.7160

It is likely the Decision Tree reached a perfect score (1.0000) because it learned the training data excessively which results in poor generalizability to newly provided data.

Model Optimization

The Decision Tree model received optimization by performing hyper parameter tuning and cross-validation procedures.

- Maximum tree depth: 10
- The split condition for a node requires a minimum number of 2 samples in the training data.
- Minimum samples per leaf: 2

Our model achieved better balance after parameter optimization because it earned a cross-validation score of 0.8160.

Phase 6: Model Evaluation

Testing Model Performance

The evaluation of all models occurred on test data to understand their practical outcome.

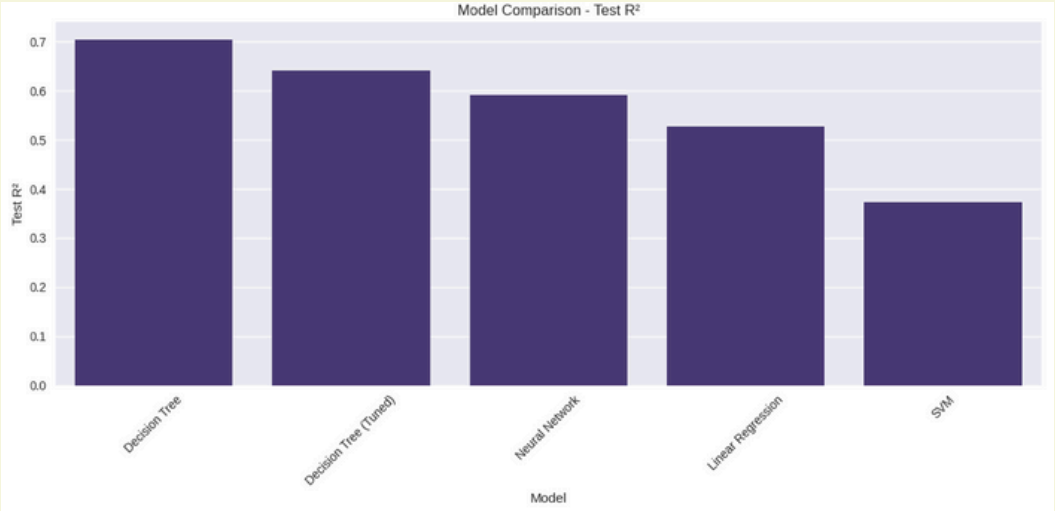
Model	Test R^2 Score	RMSE
Decision Tree	0.7044	7.7896
Decision Tree (Tuned)	0.6413	8.5803
Neural Network	0.5913	9.1591
Linear Regression	0.5273	9.8496
SVM	0.3740	11.3353

Performance Analysis

- The R^2 Score measures the percentage of price variations which our predictive model effectively interprets. Higher is better.
- The RMSE calculation provides an average prediction error measurement in Saudi Riyals (SAR) terms. Lower is better.

Our initial Decision Tree model achieved the best results because it explained 70.4% of product price variations through its R^2 score of 0.7044. Our price predictions through this model demonstrate an RMSE of 7.79 SAR which means they fall within a range of approximately 8 SAR from actual prices.

.



Phase 7: Insights & Reporting

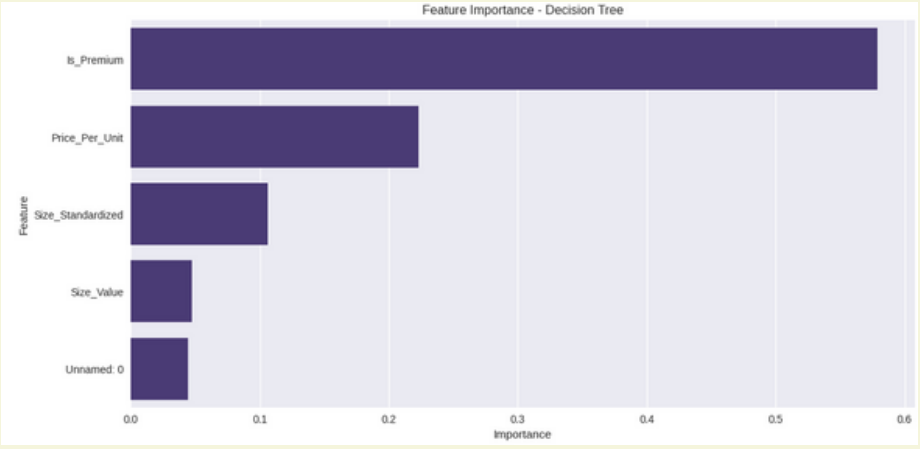
Feature Importance

A analysis identified which elements have the most effect on product pricing.

Feature	Importance
Is_Premium	57.9%
Price_Per_Unit	22.3%
Size_Standardized	10.6%
Size_Value	4.7%
Unnamed: 0	4.4%

This shows that:

- Whether a product is premium or not is the strongest predictor of its price
- Price per unit (value) is the second most important factor
- Product size also plays a significant role in determining price

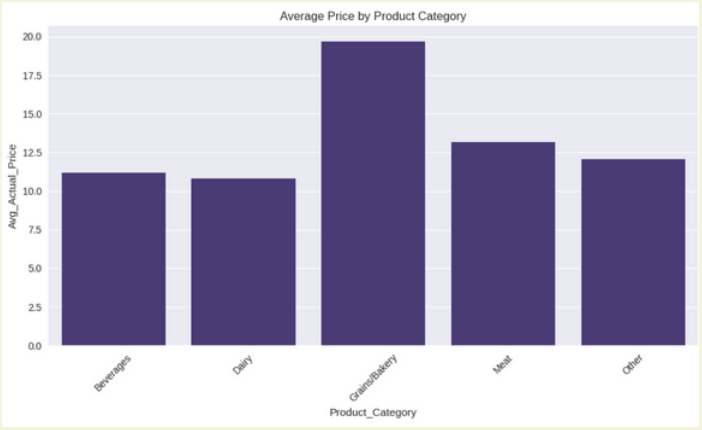


Product Category Analysis

In this section we finally observed the average prices and cost of different product categories:

Product Category	Average Price (SAR)
Grains/Bakery	19.69
Meat	13.18
Other	12.04
Beverages	11.16
Dairy	10.80

Food products made with grains and bakeries cost the most but dairy items are normally priced the lowest among other items.



Key Findings

1. **Price Distribution:**
   - Low price range (below 4.50 SAR): 304 products
   - Mid price range (4.50 - 16.95 SAR): 614 products
   - High price range (above 16.95 SAR): 302 products
2. **Size-Price Relationship:**
   - Moderate positive correlation (0.4009) between product size and price
   - Larger products tend to be somewhat more expensive, but the relationship isn't extremely strong
3. **Value Analysis:**
   - Average price per unit: 0.39 SAR
   - Lowest price per unit-Best: Drinking Water
   - Lowest value (highest price per unit): Green Tea & Mint