KINGDOM OF SAUDI ARABIA
Ministry of Higher Education
Taibah University
College of Computer Science and Engineering (Girls Section)

المملكة العربية السعودية
وزارة التعليم العالي
جامعة طيبة
كلية علوم وهندسة الحاسبات
(قسم الطالبات)

A project for

Data Mining & Data Warehouse IS-372

"Dates Export in Saudi Arabia"

Areej Hilal Alsehli	**F10**	**4150201**
Rand Abdulaziz Al-subeeh	**F 4**	**4258567**

Supervised BY dr.Alaa Alharbl

Introduction

Objective:

This project aims to achieve several objectives, including studying the development and growth of the Saudi date export market and helping to understand the most influential countries in this sector's prosperity. From other perspectives, this project was selected for implementation to support and contribute to Vision 2030, which aims to strengthen Saudi Arabia's position as the world's largest exporter of dates and raise the value of non-oil exports

The importance of data analysis and mining:

The importance of data analysis and mining in the current era lies fundamentally in making accurate decisions and making predictions based on previous studies. It helps in understanding patterns and relationships, which improves the way data is handled and leads to things such as: increased competitiveness, improved innovations, and increased efficiency.

Data Selection

The data was retrieved from the Open Data Platform [1], a platform developed by the Saudi Data and Artificial Intelligence Authority (SDAIA). The official publisher of the data is the national center for palms and dates [2].

The data covers exports, including the year, product classification, countries, export value, and shipped quantity. Each row represents an export transaction to a specific destination, reflecting the global distribution of dates

The total number of the dataset is 1911 rows, from the period 2016-2023.

This data set was selected to understand the date trading sector and gain a broader perspective in the past and current years, to influence the efficiency of choices and discover patterns through analysis in order to improve decision making.

Datasets Description

Data	Data type	Description
Year	numeric	Export year
Weight (kg)	numeric	Weight of the exported shipment
H.S code	numeric	A standardized numerical method of classifying traded products [3]
Commodity Description AR	text	Type of dates exported in Arabic
Commodity Description EN	text	Type of dates exported in English
Value (S.R)	numeric	Shipment price in riyals
Country AR	text	The country to which it was exported to, written in Arabic
Country EN	text	The country to which it was exported to, written in English

```
[1] from google.colab import files
    uploaded = files.upload()
```

```
Choose Files datesExport.csv
• datesExport.csv(text/csv) - 183559 bytes, last modified: 4/12/2025 - 100% done
Saving datesExport.csv to datesExport.csv
```

```
import pandas as pd
import io
df = pd.read_csv("datesExport.csv", thousands=',')
df.head()
```

	Year	H.S Code	Commodity Description AR	Commodity Description EN	Country AR	Country EN	Value (S.R)	Weight (kg)
0	2016	8041010	تمر طازج	FRESH DATES	اثيوبيا	ETHIOPIA	438457	431527
1	2016	8041010	تمر طازج	FRESH DATES	اريتيريا	ERITREA	84000	20000
2	2016	8041010	تمر طازج	FRESH DATES	اسبانيا	SPAIN	8020	1230
3	2016	8041010	تمر طازج	FRESH DATES	استراليا	AUSTRALIA	692235	96568
4	2016	8041010	تمر طازج	FRESH DATES	افغانستان	AFGHANISTAN	795996	150068

- Reading the file and printing the first 5 rows.

Data Preprocessing

In this section, the most important work necessary to prepare the dataset before starting the project was carried out, which included the following:

Cleaning:

It was noticed that the country and product description columns were in two different languages, and Arabic was removed to improve the data entered the model and analysis, and the English columns were renamed for uniqueness.

```
#preview columns to delete the required one
for i, col in enumerate(df.columns):
    print(f"Column {i}: {col}")
```

```
Column 0: Year
Column 1: H.S Code
Column 2: Commodity Description AR
Column 3: Commodity Description EN
Column 4: Country AR
Column 5: Country EN
Column 6: Value (S.R)
Column 7: Weight (kg)
```

```
col_index = [2,4]
df = df.drop(df.columns[col_index], axis=1)
print(df)
```

```
      Year  H.S Code Commodity Description EN  Country EN   Value (S.R)  \
0     2016   8041010              FRESH DATES     ETHIOPIA       438457
1     2016   8041010              FRESH DATES      ERITREA        84000
2     2016   8041010              FRESH DATES        SPAIN         8020
3     2016   8041010              FRESH DATES    AUSTRALIA       692235
4     2016   8041010              FRESH DATES  AFGHANISTAN       795996
...    ...       ...                      ...          ...          ...
1905  2023   8041029       OTHER DRIED DATES   MAURITANIA       921103
1906  2023   8041029       OTHER DRIED DATES    MAURITIUS       265900
1907  2023   8041029       OTHER DRIED DATES      NIGERIA        20269
1908  2023   8041029       OTHER DRIED DATES  NETHERLANDS      1799410
1909  2023   8041029       OTHER DRIED DATES    HONG KONG       152067

      Weight (kg)
0          431527
1           20000
2            1230
3           96568
4          150068
...           ...
1905       229150
1906        29346
1907         2780
1908       174803
1909         8510

[1910 rows x 6 columns]
```

```
# renaming columns
df.columns = ['Year', 'H.S Code', 'Commodity Description', 'Country', 'Value (S.R)', 'Weight (kg)']
```

handling missing and duplicates values:

As part of the Preprocessing, a missing and duplicates value was checked, and it was found that there was none in the data

```
print("missing valuse:")
print(df.isnull().sum())
```

```
missing valuse:
Year                    0
H.S Code                0
Commodity Description    0
Country                 0
Value (S.R)             0
Weight (kg)             0
dtype: int64
```

```
print("duplicates valuse:")
print(df.duplicated().sum())
```

```
duplicates valuse:
0
```

transforming data into a suitable format for analysis:

The commas were changed so that Python reads them as INT and not as STRING, using (thousands=',') which considered Data Parsing because It tells pandas how to interpret the data directly when it is read, without the need for subsequent manual cleanup

```
[ ]  import pandas as pd
     import io
     df = pd.read_csv("datesExport.csv", thousands=',')
     df.head()
```

Descriptive Statistics and Visualization

In this section descriptive statistical analysis has been Conduct on the dataset to gain insights into its characteristics.

Descriptive analysis provides us with a broader understanding of data and identifies problems and patterns. It presents us with a numerical summary of the data, such as (median, mean, and standard deviation). It helps us understand the distribution and deal with extreme values. Then, we present it through graphical visualization.

descriptive statistical analysis

```
[ ]  df.describe()
```

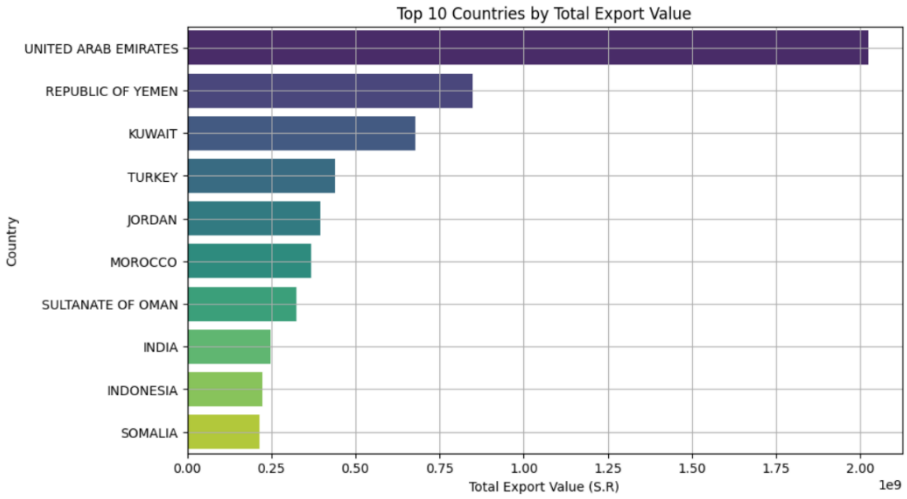
	Year	H.S Code	Value (S.R)	Weight (kg)
count	1,910	1,910	1,910	1,910
mean	2,020	8,041,019	4,096,767	945,083
std	2	7	15,620,508	4,374,352
min	2,016	8,041,010	1	1
25%	2,018	8,041,010	23,011	4,740
50%	2,020	8,041,021	264,998	43,072
75%	2,022	8,041,021	1,605,464	295,474
max	2,023	8,041,029	189,337,551	72,631,242

From the statistical analysis shown above, it can be concluded that the data shows a concentration around a specific commodity imported or exported in very different quantities and values. It is possible that a few records have a significant impact on the overall averages due to their outliers, which will be Visualized beneath.

Visualization

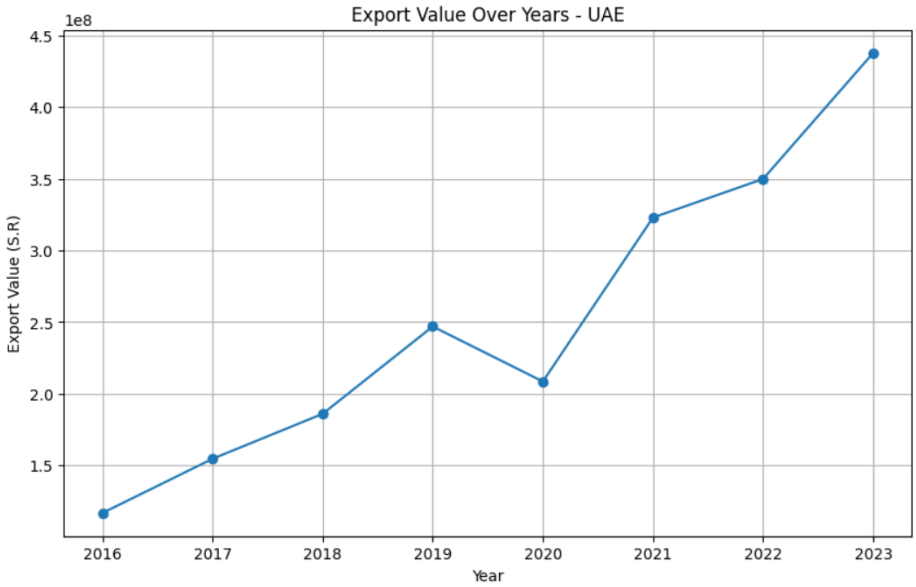
BAR PLOT

This chart represents the top 10 countries in terms of export value in Saudi Riyals. It clearly shows that the United Arab Emirates comes in first place by a very large margin over the rest of the countries, followed by Yemen, then Kuwait. At the end of the list, Somalia and Indonesia are shown with the lowest export value, indicating the market's dependence mainly on UAE exports, which highlights the opportunities available to strengthen the market towards a specific country.



LINE PLOT

In this graph, the study of the value of UAE exports during the years available in the data list was determined after noting that the UAE has the highest list of total exports, as the curve in the graph shows a continuous increase over the years starting from 2016 and reaching 2023, with a clear decrease in the year 2020 compared to 2019, and this may be due to the occurrence of the Covid-19 pandemic.



Multiple Line Chart

This graph was created to understand the most in-demand date varieties during the years 2016-2023, allowing for a greater understanding of market demand. As shown, the following is analyzed:

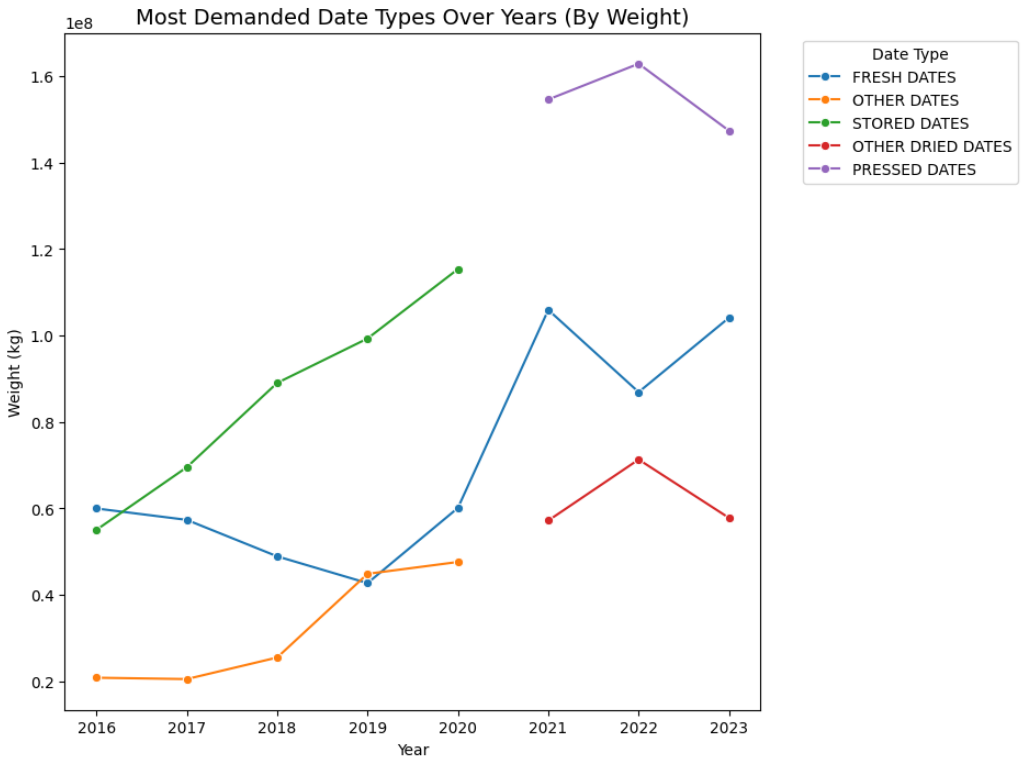
First: Highest Demand

- Pressed dates are the most in demand by weight.
- Fresh dates are the most in demand by timeframe, as they are the only variety that has been exported continuously from 2016 to 2023, while the remaining varieties vary over time.

Second: Conclusions

Pressed dates were the most in-demand product in the period from .2020 to 2023

Fresh Dates showed a decline in 2019 but rebounded in 2020, indicating a revitalized market

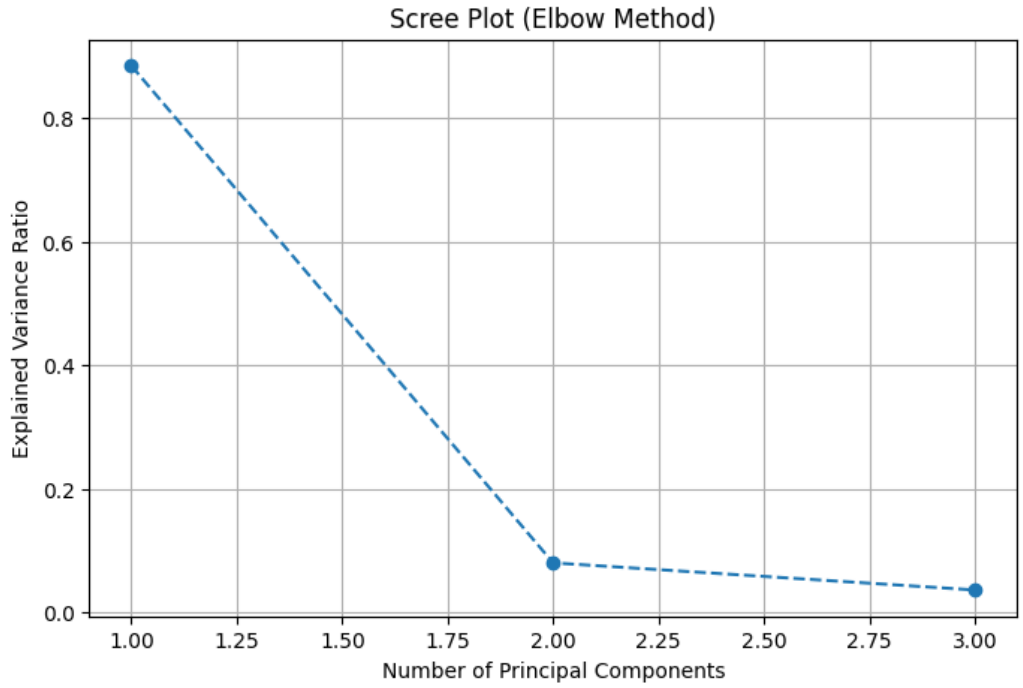


Principal Component Analysis (PCA) Implementation

In this part of the project, PCA was implemented to reduce the dimensions of the data while preserving as much information as possible. This provides us with a simplification of complex data, improves algorithms, and reduces noise resulting from low-variance components. The elbow method was implemented and displayed using a scree plot, and then PCA SCATTER PLOT was implemented based on that.

Scree Plot (Elbow Method)

This diagram below represents the elbow method, where the elbow is the point after which the variance curve begins to flatten, i.e.: adding any additional components no longer adds important information. In this diagram, the elbow is located at point 2 (i.e.: when the number of principal components = 2) and represents the best balance where the error rate becomes more stable.



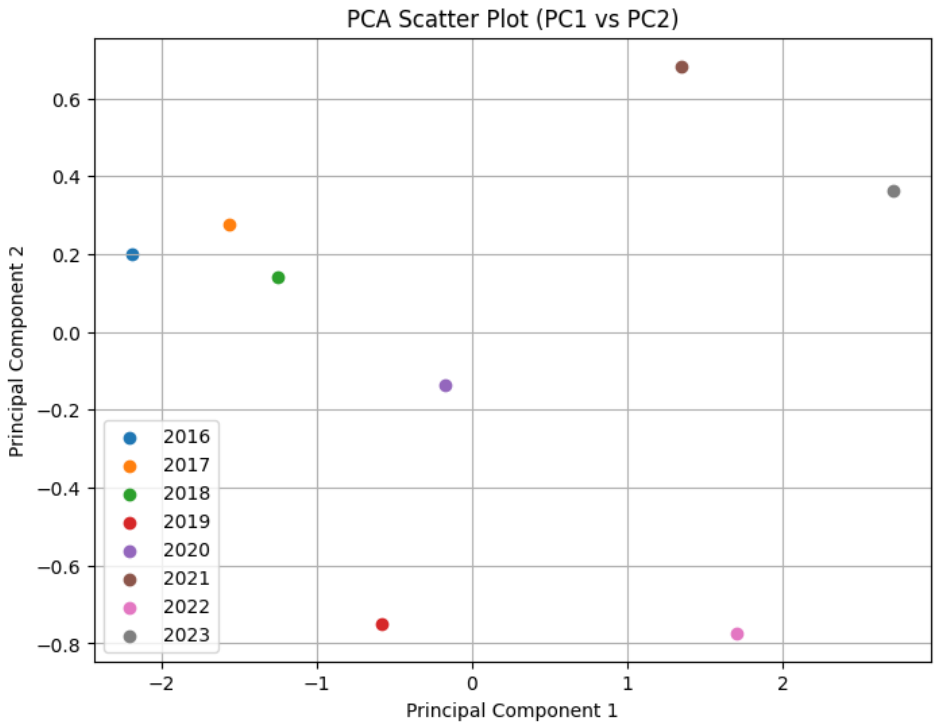
PCA Scatter plot

THIS SCATTER PLOT provides visual representation from 2016 to 2023 after dimension reduction to 2 main components where every year is represented as a dot, and Their positioning in the graph reflects the data pattern for that year based on principal components.

Conclusion:

Convergent data, such as 2016-2017, indicates similar characteristics.

Divergent data, such as 2016-2021, indicates significant differences in the data pattern.

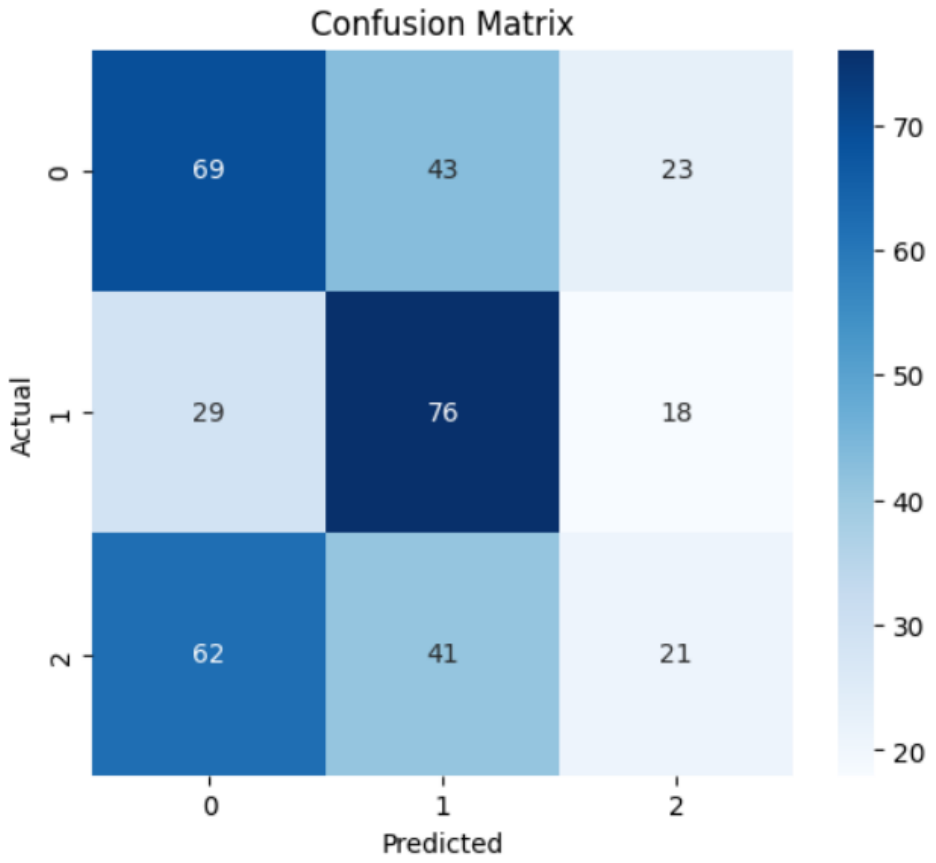


Classification

In this section we start training the Random Forest Classifier that aimed to Predict the 2024 Exports Based on Date Type, first by classify the export commodities into three **weight categories** (Low – Medium - High) based on two features: **Year** and **Commodity Description** (encoded as a numeric code)

And evaluating the classifier's performance using a **classification report**, **confusion matrix**, and **heatmap visualization**.

Confusion Matrix (Heatmap Visualization)



Conclusion:

1- The heatmap provides us with a visualization to the distribution of Actual vs predicted.
2- 76 correct "Low" classifications that are seen in the center is considered the strongest block.
3- There are misclassifications that are frequent and diverse.
4- Significant confusion between all three categories. For example:
   - 62 actual "Medium" samples were classified as "High".
   - 43 actual "High" were classified as "Medium"

	Commodity Description	Predicted Weight Category 2024
0	FRESH DATES	High
1	STORED DATES	Low
2	OTHER DATES	Medium
3	PRESSED DATES	Low
4	OTHER DRIED DATES	Medium

Evaluation and Results Interpretation

Classification Report:

	precision	recall	f1-score	support
High	0.43	0.51	0.47	135
Low	0.47	0.62	0.54	123
Medium	0.34	0.17	0.23	124
accuracy			0.43	382
macro avg	0.41	0.43	0.41	382
weighted avg	0.42	0.43	0.41	382

Classification Report Summary:

- Overall accuracy: 43%
- Macro average: precision: 41%, recall: 43%, F1-score: 41%

Conclusion

This project covers a comprehensive analysis of date export data from Saudi Arabia during the period 2016–2023. The goal is to understand the market and identify the most prominent importing countries and the most in-demand date varieties, supporting the Kingdom's efforts to meet Vision 2030's goal of boosting exports. The data was processed from various aspects, including cleaning, sorting, and analysis using descriptive statistical techniques and graphs, providing a deeper understanding of the market.

Finally, the random forest classifier was used to predict 2024 exports. The results showed that the UAE represents the largest export destination, with the most in-demand date varieties highlighted. The confusion matrix also provided an accuracy of 43%, highlighting the challenges of accurate classification

Recommendations

- Improving data quality:

such as adding new features such as packaging type or storage method.

- Seasonal analysis:

such as analyzing specific months, such as religious seasons (Ramadan and Hajj), or a global issue such as the COVID-19 pandemic.

Appendix:

https://colab.research.google.com/drive/18lWiX7-2AaGr44Y_L7LyPIZ9TJP51XRw?usp=sharing

References:

[1] data source: https://open.data.gov.sa/en/datasets/view/d0a9c6c5-ad58-40e7-be1e-772990d001ad

[2] publisher :https://ncpd.gov.sa/ar/reports-statistical

[3] https://www.trade.gov/harmonized-system-hs-codes