**Post Graduate Studies**

**Njala University**

**Freetown Campus**

**Pattern Recognition Project**

***Performance Analysis using Apriori and Machine Learning Algorithms***

*Group 1*

*Alhaji Kanneh*

*Francis Jr. Hazley*

Contents

[Dataset: Global Music Streaming Trends & Listener Insights 3](#_Toc192861369)

[About Dataset: 3](#_Toc192861370)

[Column Descriptions: 3](#_Toc192861371)

[Analysis Done 3](#_Toc192861372)

[Steps Followed 3](#_Toc192861373)

[Outcome 4](#_Toc192861374)

[Dataset: Supermarket Sales 5](#_Toc192861375)

[About Dataset 5](#_Toc192861376)

[Column Description 5](#_Toc192861377)

[Analysis Done 5](#_Toc192861378)

[Steps Followed 5](#_Toc192861379)

[Outcome 6](#_Toc192861380)

[Dataset: Final Post College Salaries 6](#_Toc192861381)

[About Dataset 6](#_Toc192861382)

[Key Column Description 6](#_Toc192861383)

[Analysis Done 6](#_Toc192861384)

[Steps Followed 6](#_Toc192861385)

[Outcome 7](#_Toc192861386)

[Dataset: MM19\_CSDB\_DS.csdb 7](#_Toc192861387)

[About Dataset 7](#_Toc192861388)

[Key Column Description 7](#_Toc192861389)

[Analysis Done 7](#_Toc192861390)

[Steps Followed 7](#_Toc192861391)

[Outcome 8](#_Toc192861392)

# Dataset: Global Music Streaming Trends & Listener Insights

## About Dataset:

This dataset provides insights into music streaming trends from 2018 to 2024 across multiple platforms like Spotify, Apple Music, and YouTube. It includes listener demographics, streaming habits, genre preferences, and engagement metrics

## Column Descriptions:

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| User\_ID | Unique identifier for each listener |
| Age | Age of the listener |
| Country | Country where the user resides |
| Streaming Platform | The platform used (Spotify, Apple Music, etc.) |
| Top Genre | The most streamed music genre |
| Minutes Streamed Per Day | Average daily listening time |
| Number of Songs Liked | Total liked songs by the user |
| Most Played Artist | The most played artist by the user |
| Subscription Type | Free or Premium subscription |
| Listening Time (Morning/Afternoon/Night) | When the user listens the most |
| Discover Weekly Engagement (%) | Percentage of auto-generated playlists played |
| Repeat Song Rate (%) | Percentage of songs repeated frequently |

## Analysis Done

Apriori Algorithm was used to find the associated rules for top genre listened on various platforms.

## Steps Followed

1. Group the data by streaming platform and top genre
2. Generate frequent sets using minimum support level of 0.6 and column names
3. Generate the overall associated rules using confidence thresh hold of 0.7
4. Eliminate from the rules any rule whose lift is not more than 1.2 to come up with strong rules set
5. Display the rules.

## Outcome

36 rule sets were generated with the top 3 been

1. (Hip-Hop, Country) -> (Jazz, Classical)

Explanation: if you listen to Hip-Hop together with Country music, you will most likely listen to Jazz and classical music.

1. (Hip-Hop, Classical) -> (Jazz, Country)

Explanation: if you listen to Hip-Hop together with classical music, you will most likely listen to Jazz and Country music.

1. (Jazz, Country) -> (Hip-Hop, Classical)

Explanation: if you listen to Jazz and Country music, you will most likely listen to Hip-Hop together with classical music.



# Dataset: Supermarket Sales

## About Dataset

"Supermarket\_sales" is a public dataset on Kaggle.

## Column Description

|  |  |
| --- | --- |
| Column Name | Description |
| Invoice id | Computer generated sales slip invoice identification number |
| Branch | Branch of supercenter (3 branches are available identified by A, B and C). |
| City | Location of supercenters |
| Customer type | Type of customers, recorded by Members for customers using member card and Normal for without member card |
| Gender | Gender type of customer |
| Product line | General item categorization groups - Electronic accessories, Fashion accessories, Food and beverages, Health and beauty, Home and lifestyle, Sports and travel |
| Unit price | Price of each product in $ |
| Quantity | Number of products purchased by customer |
| Tax | 5% tax fee for customer buying |
| Total | Total price including tax |
| Date | Date of purchase (Record available from January 2019 to March 2019) |
| Time | Purchase time (10am to 9pm) |
| Payment | Payment used by customer for purchase (3 methods are available – Cash, Credit card and Ewallet) |
| COGS | Cost of goods sold |
| Gross margin percentage | Gross margin percentage |
| Gross income | Gross income |
| Rating | Customer stratification rating on their overall shopping experience (On a scale of 1 to 10) |

## Analysis Done

Apriori Algorithm was used to find association between product lines bought by customers in an invoice.

## Steps Followed

1. Group the data by invoice id and product line
2. Generate frequent sets using minimum support level of 0.6 and column names
3. Generate the overall associated rules using confidence thresh hold of 0.5
4. Eliminate from the rules any rule whose lift is not more than 1.2 to come up with strong rules set
5. Display the rules.

## Outcome

42 rule sets were generated with top three being:

1. (Electronic accessories) -> (Fashion accessories)
2. (Fashion accessories) -> (Electronic accessories)
3. (Electronic accessories) -> (Home and lifestyle)



# Dataset: Final Post College Salaries

## About Dataset

This dataset contains information on the salaries and job satisfaction of various college majors, categorized by their early career pay, mid-career pay, and the percentage of graduates who find their work to be highly meaningful. The data is ranked based on the mid-career pay, with the highest-paying majors listed first

## Key Column Description

|  |  |
| --- | --- |
| Column Name | Description |
| **Rank** | The ranking of the major based on mid-career pay |
| **Major** | The field of study or major |
| **Degree Type** | The type of degree (in this case, all are bachelor’s degrees) |
| **Early Career Pay** | The average salary for graduates in the early stages of their career (typically 0-5 years after graduation) |
| **Mid-Career Pay** | The average salary for graduates in the mid-career stage (typically 10+ years after graduation) |
| **% High Meaning** | The percentage of graduates who report that their work has high meaning or significance |

## Analysis Done

Use Machine Learning to predict mid-career pay based on major, degree type, early career pay, and % high meaning

## Steps Followed

1. Encode categorical variables (Major, Degree Type) using one-hot encoding
2. Normalize numerical features (Early Career Pay, % High Meaning)
3. Split data into training and testing sets (e.g., 80-20 split)
4. Apply Algorithms
   1. Linear Regression: Baseline model.
   2. Random Forest Regressor: Handles non-linear relationships.
   3. Gradient Boosting Regressor using XGBoost for improved accuracy.
5. Predict the metrics
   1. Linear Regression: Baseline model.
   2. Random Forest Regressor: Handles non-linear relationships.
   3. Gradient Boosting Regressor using XGBoost for improved accuracy.

## Outcome

The results were:

Linear Regression - MAE: 3,998.5552 MSE: 32,256,534.5435 R²: 0.948

Random Forest - MAE: 132.5098 MSE: 60,972.3268 R²: 0.9999

XGBoost - MAE: 456.5551 MSE: 1,085,706.9535 R²: 0.9983

# Dataset: MM19\_CSDB\_DS.csdb

## About Dataset

This dataset contains time-series data related to various industrial and economic indicators, primarily focused on manufacturing sectors such as aerospace, electronics, and metal products. The data is indexed to the year 2010 and includes both annual and quarterly data points.

## Key Column Description

|  |  |
| --- | --- |
| Column Name | Description |
| **Year/Quarter/Month** | The time for the data, ranging from 1996 to 2015 |
| **K33V** | Air & Spacecraft & related Machinery |
| **K386** | Fabricated Metal products, except Machinery & Equipment |
| **K387** | Computer, Electronic & Optical products |
| **K38B** | Other Transport Equipment |
| **K5NY** | Average Weekly Earnings Index for 24-25 |
| **K5NZ** | Combined Costs for Weapons & Ammunition |
| **K5O2** | Average Weekly Earnings Index for 26-30 |
| **K5O3** | Combined Costs for Electronic Components & Boards |
| **K5O4** | Combined Costs for Measuring, Testing Navigation Equipment |
| **K5O5** | Combined Costs for Air & Spacecraft & related Machinery |
| **K8D4** | General Expenses for MM19 (RPI Excluding Food) |
| **MB4S** | GSI (excl. CCL) - Inputs for Manufacture of Computer, Electronic & Optical products |
| **MC48** | GSI (excl. CCL) - Inputs for Manufacture of Weapons & Ammunition |
| **MC4A** | GSI (excl. CCL) - Inputs for Manufacture of Air/Spacecraft & Related Machinery |

## Analysis Done

Using Machine Learning to predict future values of industrial indicators (e.g., K33V, K386) based on historical data

## Steps Followed

1. Handle missing values (e.g., fill with mean or interpolate).
2. Normalize numerical features
3. Create lagged features for time-series prediction (e.g., use past 3 months' data to predict the next month)
4. Split data into training and testing sets (e.g., 80-20 split)

## Outcome

ARIMA - MAE: 9.2849 RMSE: 10.6903

LSTM - MAE: 6.7847 RMSE: 10.1565

Random Forest - MAE: 0.3876 RMSE: 0.8246