DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

FINAL REVIEW

TITLE:
SPOTIFY PODCAST ANALYSIS AND PREDICTION MODEL

TEAM MEMBERS:

DEEPANA D	927622BAD006
JANANI V	927622BAD021
RITHANI KS	927622BAD045

GUIDED BY

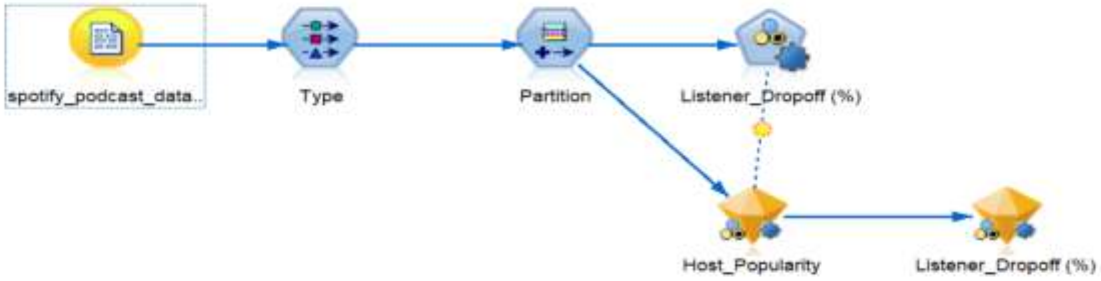
Ms. SUBHASRI S IBM

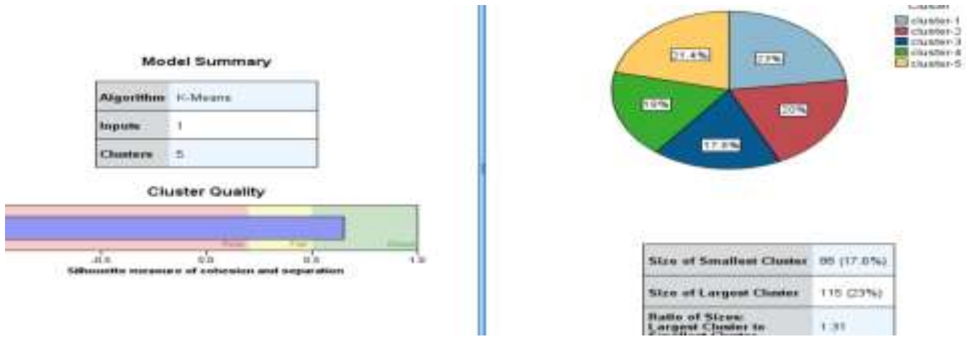
ABSTRACT

- In the fast-growing podcast industry, retaining listener attention is essential. This project aims to predict listener drop-off and generate podcast recommendations using **IBM SPSS Modeler**. The analysis uses podcast-related features such as episode duration, host popularity, ad positioning, and user behavior.
- The dataset is processed through SPSS Modeler where classification models like **C5.0** and **CHAID** are applied to predict drop-off behavior, while **K-Means clustering** is used to group listeners based on engagement patterns. The **Auto Classifier** aids in selecting the best-performing model for deployment.
- The goal is twofold: first, to predict whether a listener will drop off before completing a podcast episode; and second, to recommend suitable podcast content for users based on behavioral clustering and preference patterns.

IMPLEMENTATION:

- The implementation of the Spotify Podcast analysis model was carried out using **IBM SPSS Modeler**, a powerful data mining tool that supports CRISP-DM methodology. The project workflow followed a clear, structured process involving data input, preparation, modeling, and evaluation nodes.




Model Summary

Algorithm	K-Means
Inputs	1
Clusters	5

Size of Smallest Cluster	89 (17.8%)
Size of Largest Cluster	115 (23%)
Ratio of Sizes: Largest Cluster to	1.31

OUTPUT DESCRIPTION:

1. Source Node – spotify_podcast_data

- **Function:** Imports the podcast dataset containing all relevant listener and episode attributes.
- **Output:** Raw data including fields such as Listener_Dropoff, Host_Popularity, Category, Ad_Position, etc.

2. Type Node

- **Function:** Assigns the correct metadata role (Input, Target, None) to each variable.
- **Output:**
  - Target: Listener_Dropoff (%)
  - Inputs: Features like Host_Popularity, Category, Duration, Device, etc.

3. C5.0 Model Node

- **Function:** Predicts whether a listener will drop off using the C5.0 decision tree algorithm.
- **Output:**
  - A high-accuracy classification tree.
    - Rules such as “If Host_Popularity > 8 and Ad_Count < 2 → No Dropoff.”
    - Helps identify key influencers of drop-off.

4. CHAID Model Node

- **Function:** Builds a multi-way decision tree using chi-square statistics for predicting listener drop-off.
- **Output:**
    - Tree with multi-level branches showing combinations like: “Free User + Mid-roll Ads → High Drop-off Risk”
    - Easily interpretable and ideal for business use.

5. K-Means Node

- **Function:** Groups listeners into behavior-based clusters using unsupervised learning.
- **Output:**
    - Clusters such as:
        - Cluster 1: High-engagement, premium users
        - Cluster 2: Low-engagement, mobile users
    - Used to personalize recommendations based on cluster behavior.

6. Recommended_Next CHAID Node

- **Function:** Predicts what type of podcast should be recommended next, based on user behavior.
- **Output:**
  - Rules like: "If Cluster = 2 AND Category = Entertainment → Recommend Short, Host-Driven Content"
  - Helps in generating content suggestions based on listener profiles.

7. Partition Node

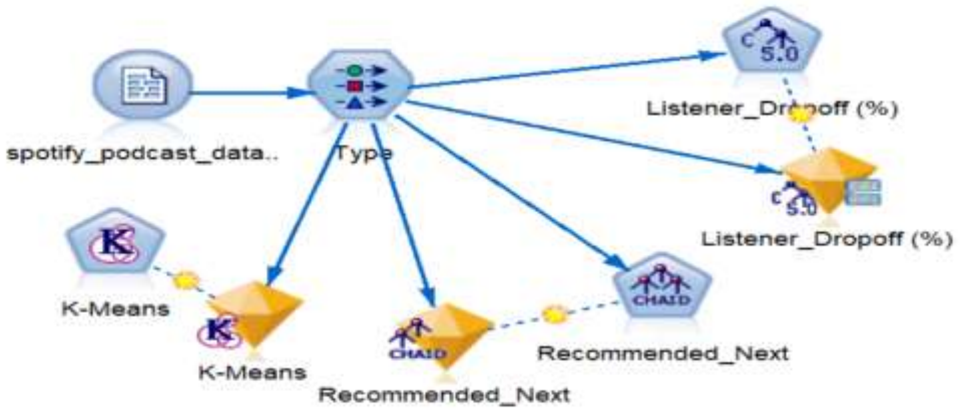
- **Function:** Splits data into training and testing sets to validate model performance.
- **Output:**
  - Ensures model evaluation is unbiased and performance metrics are reliable.

8. Derived Field (Listener_Dropoff % Calculation)

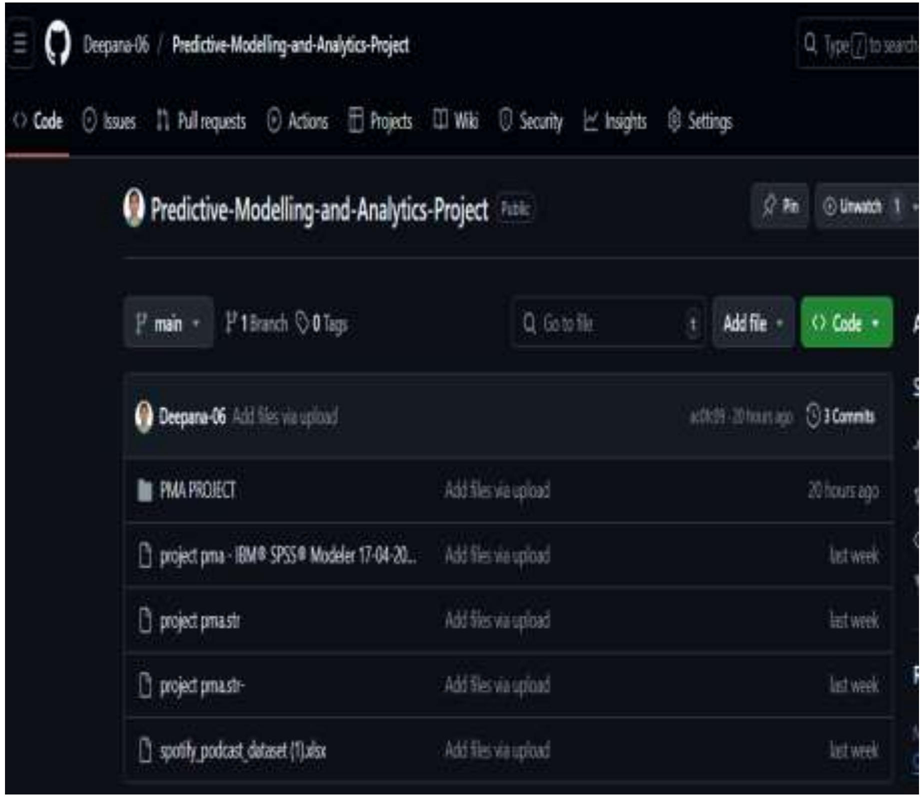
- **Function:** Computes drop-off percentage if not originally in the dataset.
- **Output:** New field Listener_Dropoff (%) for modeling.

9. Model Evaluation

- Each model (C5.0, CHAID, K-Means) is evaluated based on:
  - **Accuracy**
  - **AUC (Area Under Curve)**
  - **Interpretability**
- Best-performing models used for final recommendations and predictions.



GITHUB SCREENSHOT:



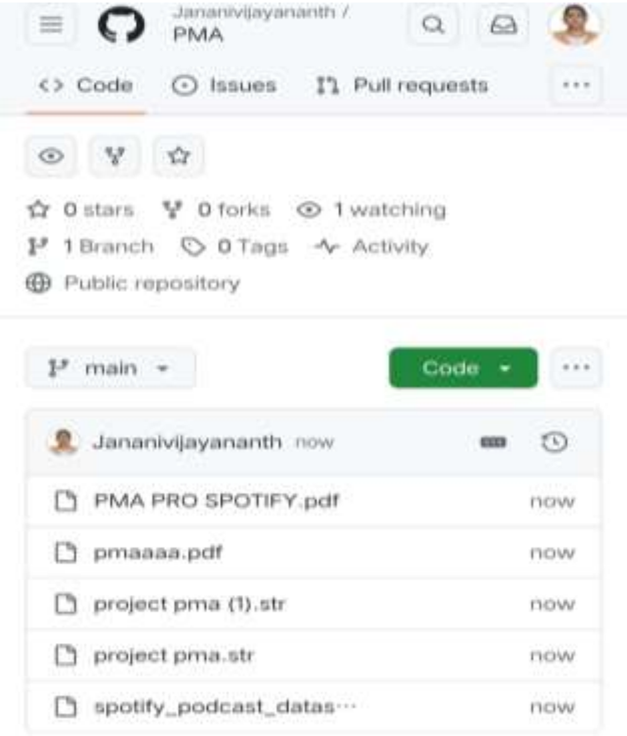
Deepana-06 / Predictive-Modelling-and-Analytics-Project

Predictive-Modelling-and-Analytics-Project

main 1 Branch 0 Tags

Deepana-06 Add files via upload a0fc09 · 20 hours ago 3 Commits

PMA PROJECT	Add files via upload	20 hours ago
project pma - IBM® SPSS® Modeler 17-04-20...	Add files via upload	last week
project pma.str	Add files via upload	last week
project pma.str-	Add files via upload	last week
spotify_podcast_dataset (1).xlsx	Add files via upload	last week



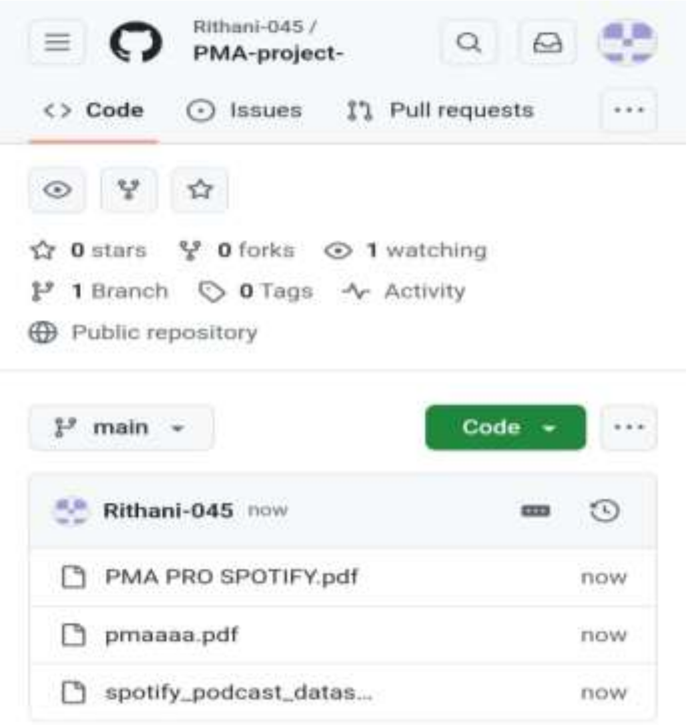
Jananivijayananth / PMA

0 stars 0 forks 1 watching
1 Branch 0 Tags Activity
Public repository

main

Jananivijayananth now

PMA PRO SPOTIFY.pdf	now
pmaaaa.pdf	now
project pma (1).str	now
project pma.str	now
spotify_podcast_datas···	now



Rithani-045 / PMA-project-

0 stars 0 forks 1 watching
1 Branch 0 Tags Activity
Public repository

main

Rithani-045 now

PMA PRO SPOTIFY.pdf	now
pmaaaa.pdf	now
spotify_podcast_datas...	now

CONCLUSION:

□ Drop-off behavior is predicted with high accuracy using Decision Trees.

□ Users are segmented into clusters for tailored podcast recommendations.

□ Rules and tree paths give explainable insights for content optimization.