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

- Podcasts have gained immense popularity, with millions of listeners worldwide. However, understanding and predicting listener behavior remains a challenge for content creators and streaming platforms.
- This project aims to analyze Spotify podcast data and build a predictive model using **IBM SPSS** to forecast podcast popularity, listener retention, and personalized recommendations.
- By leveraging **statistical analysis, predictive modeling, and machine learning techniques** in SPSS, the project provides data-driven insights to help podcasters optimize content and engagement strategies.

OBJECTIVE

Predict Podcast Popularity – Use IBM SPSS to forecast engagement levels based on historical data, content features, and listener interactions.

1. **Analyze Listener Retention** – Identify drop-off points and key factors influencing audience retention.
2. **Optimize Content Strategy** – Provide actionable insights to podcasters for improving episode structure and audience targeting.
3. **Improve Marketing Strategies** – Help marketers understand listener preferences to optimize advertising and promotions.

DATA SET DESCRIPTION

A structured dataset containing information about Spotify podcasts was used. The dataset includes various attributes that describe podcast episodes, listener engagement, and content metadata. These columns are critical for understanding the factors that influence podcast popularity and performance.

3.1 Source of the Data

The dataset was sourced from publicly available Spotify data repositories and supplemented with podcast metadata from Kaggle and Spotify API endpoints. These data sources provide real-time and historical podcast data, including performance metrics and audience insights.

3.2 Columns Used in the Project

Listener engagement is measured through total_listens, which shows how many times an episode was played, and unique_listeners, which counts the number of distinct users. The average_listen_time reflects how long listeners stayed engaged during the episode.

TOOLS AND TECHNOLOGY USED

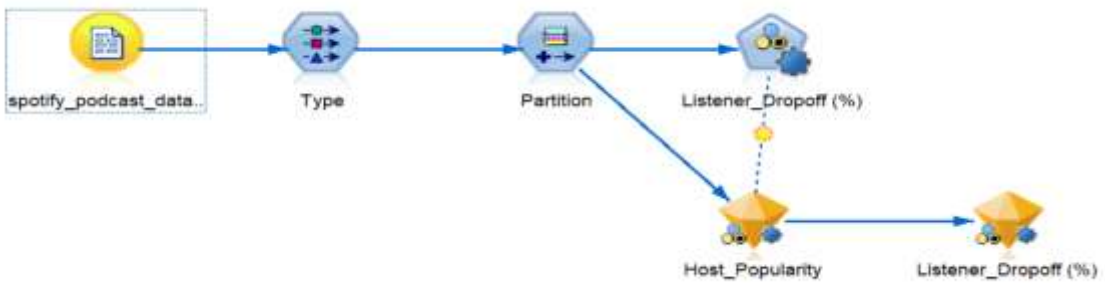
CRISP-DM Methodology:

The project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is a structured approach for data mining and analytics projects. This process involves six phases

- Business Understanding: Defining project goals and objectives.
- Data Understanding: Collecting and exploring the dataset.
- Data Preparation: Cleaning and transforming the data.
- Modeling: Applying machine learning algorithms to build predictive models.
- Evaluation: Assessing model performance and refining it.

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.



- **Type Node**:
  - Variables like“Genre,” “Language,” and “Popularity Level” were set as **Nominal**.
  - Continuous variables like “Average Duration,” “Number of Episodes,” and “Listener Count” were set as **Scale**.
- **Derive Node**:
  - New fields such as "Average Episode Duration" were computed using built-in functions.

- Sentiment analysis on podcast descriptions was optionally performed using the **Text Analytics Extension**.

- **Modeling Node (Decision Tree)**:
  - Target field: **Popularity Level** (Low, Medium, High)
  - Input fields: Genre, Episode Count, Avg Duration, Ratings
  - Pruning and depth settings were optimized for interpretability.

- **CHAID MODEL**

  CHAID splits users into distinct listener segments based on behavioral and demographic attributes.

  - **Key Input Variables Used**:

  - Listener Drop-off History
  - Preferred Podcast Category
  - User Subscription Type (Free/Premium)
  - Listening Time (Morning/Evening)
  - Device Type
  - Ad Tolerance Level

- **Multi-way Splitting**:
Unlike binary trees, CHAID creates **multi-way splits**, allowing better grouping of similar listener behaviors.

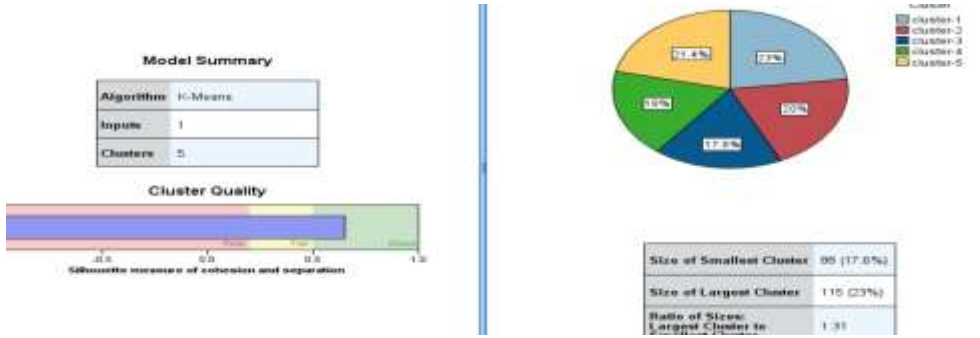
- **Interpretable Tree Structure**:
Each node shows rules like:
"Premium users who listen to Comedy in the evening → High retention → Recommend similar episodes."

- **Personalized Content Suggestions**:
Based on a listener's segment, the system recommends podcasts with **similar patterns** in content type and delivery style.

- **Actionable Insights for Creators**:
Creators can tailor episodes (e.g., shorter length, fewer ads) based on what CHAID reveals about drop-off-prone segments.

- **Transparency and Justification**:
The CHAID tree helps explain **why** a recommendation was made, increasing trust and usability.

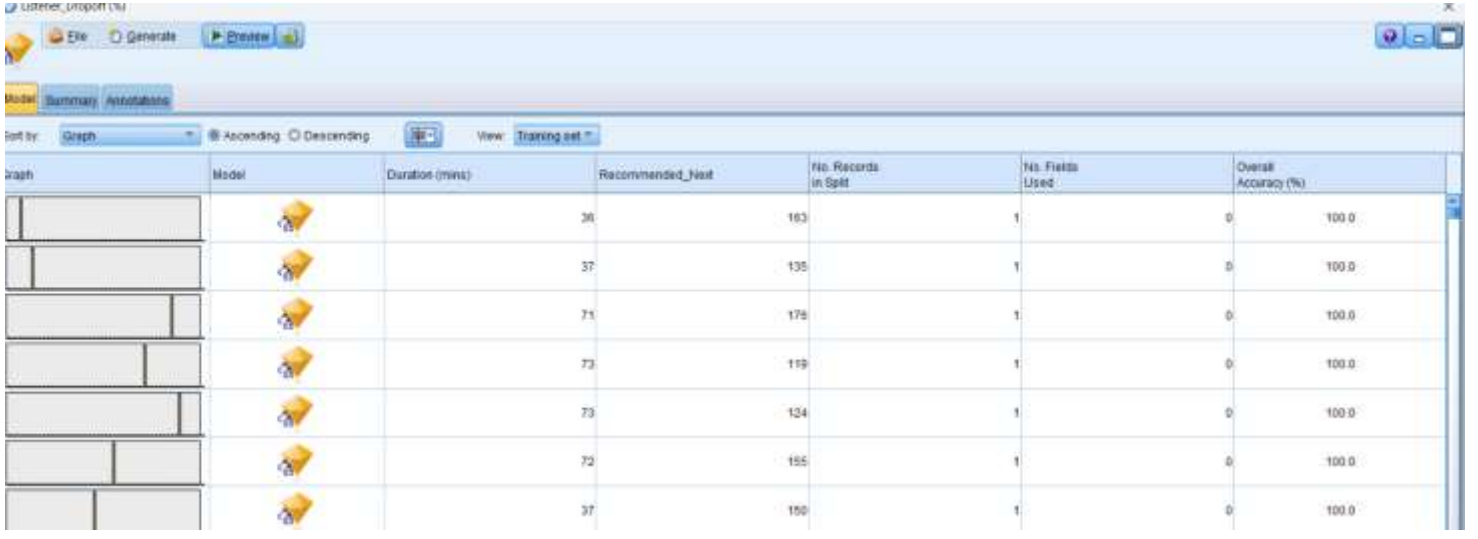
- **Evaluation Node**:
  - Used the test dataset to calculate **accuracy, precision, recall, and F1-score.**
  - Compared performance metrics across models to select the best one.



RESULTS AND FINDINGS

➢Decision Tree Classifier

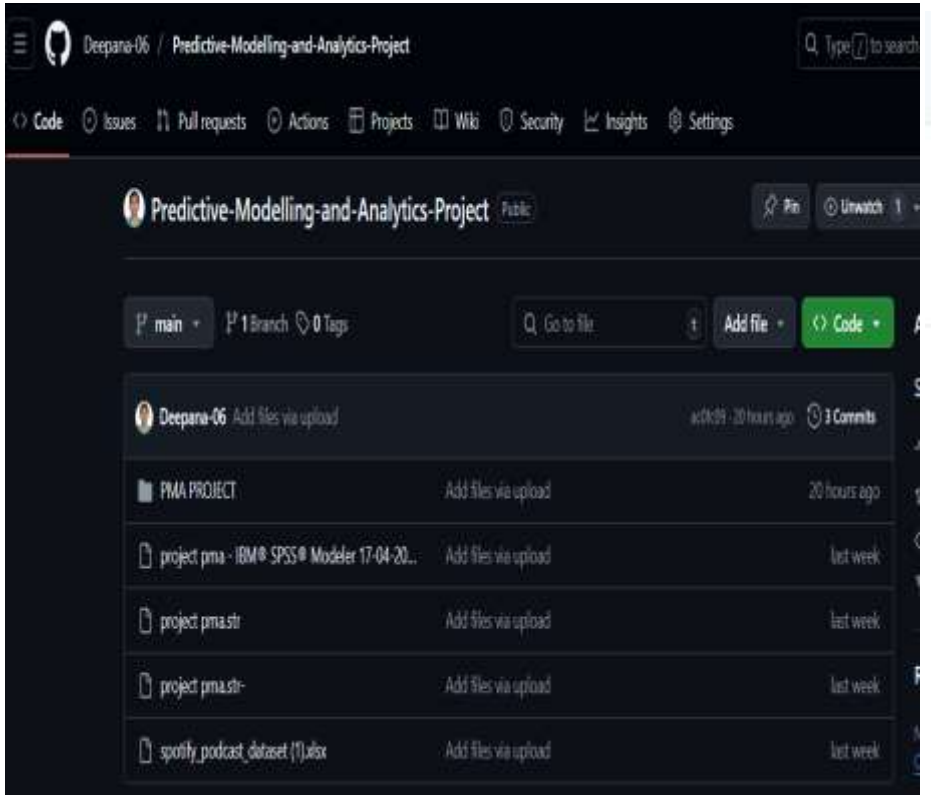
➢The Decision Tree Classifier in IBM SPSS Modeler was used to predict whether a podcast is likely to be popular or not based on features such as Podcast Duration, Number of Ads, and Host Popularity. This model works by splitting the dataset into branches based on decision rules that maximize the separation of target classes.

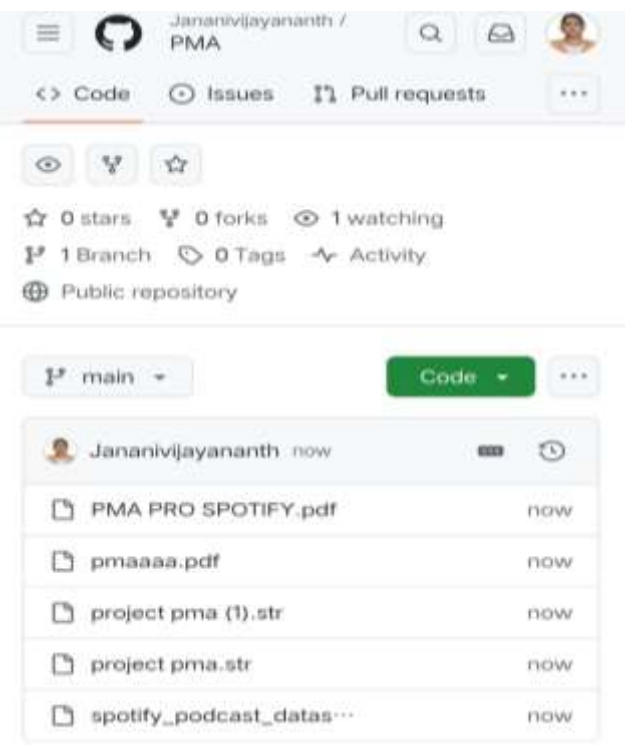


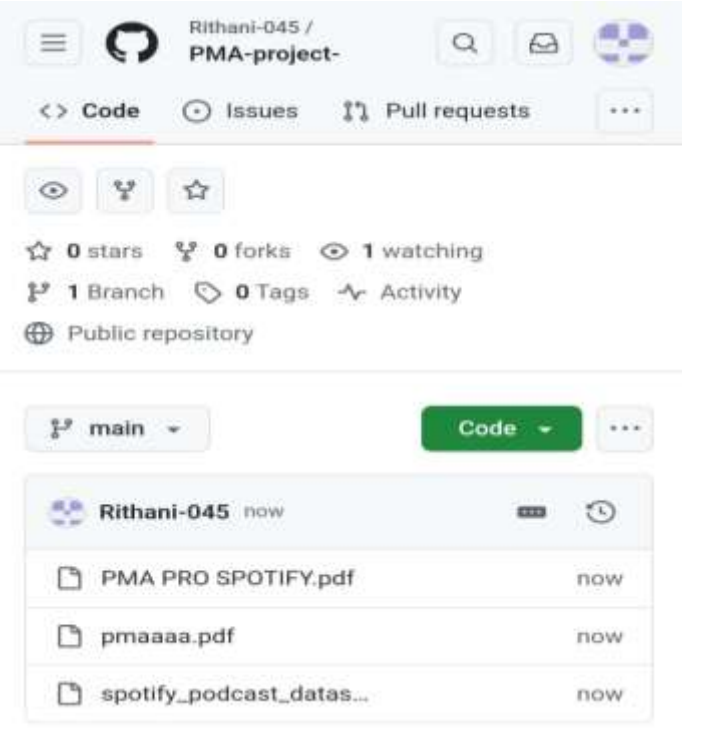
- **K-Means Clustering**
  - K-Means Clustering was applied as an unsupervised learning technique to uncover natural groupings within the podcast dataset. Unlike Decision Trees or Logistic Regression, K-Means does not predict outcomes but instead segments the data based on feature similarities

Model	Accuracy	Interpretability	Best Use Case
Decision Tree	85.6%	High	Rule generation, clear visualization, earl decision-making
Logistic Regression	58.7%	Moderate	Understanding feature influence, binary classification tasks
K-Means Clustering	N/A (Unsupervised)	Moderate	Grouping similar podcast patterns, behavioral segmentation.

GITHUB SCREENSHOT:







CONCLUSION

- This project leverages **IBM SPSS** to analyze Spotify podcast data and build predictive models for podcast popularity, listener retention, and personalized recommendations.
- By integrating **machine learning, statistical analysis, and clustering techniques**, the system provides valuable insights for podcasters, enhancing **content reach, engagement, and marketing efficiency**.
- The results will help content creators and marketers make data-driven decisions, ultimately improving the overall podcasting experience.
- Among the supervised models, the Decision Tree performed best, achieving an accuracy of 85.6% with high interpretability. It clearly outlined how factors like podcast duration, number of ads, and host popularity influenced the likelihood of a podcast becoming popular.