DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

TITLE:
SPOTIFY PODCAST ANALYSIS AND PREDICTION MODEL

TEAM MEMBERS:

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

GUIDED BY
Ms. SUBHASRI S IBM

INTRODUCTION

➢ Spotify, one of the leading podcast streaming platforms, hosts thousands of podcast shows across various genres. However, many podcasts struggle to reach the right audience and retain listeners.

➢ Understanding the factors that drive **podcast popularity, listener engagement, and retention** is crucial for content creators and marketers.

➢ This project leverages **IBM SPSS** to perform advanced statistical analysis and predictive modeling on Spotify podcast data.
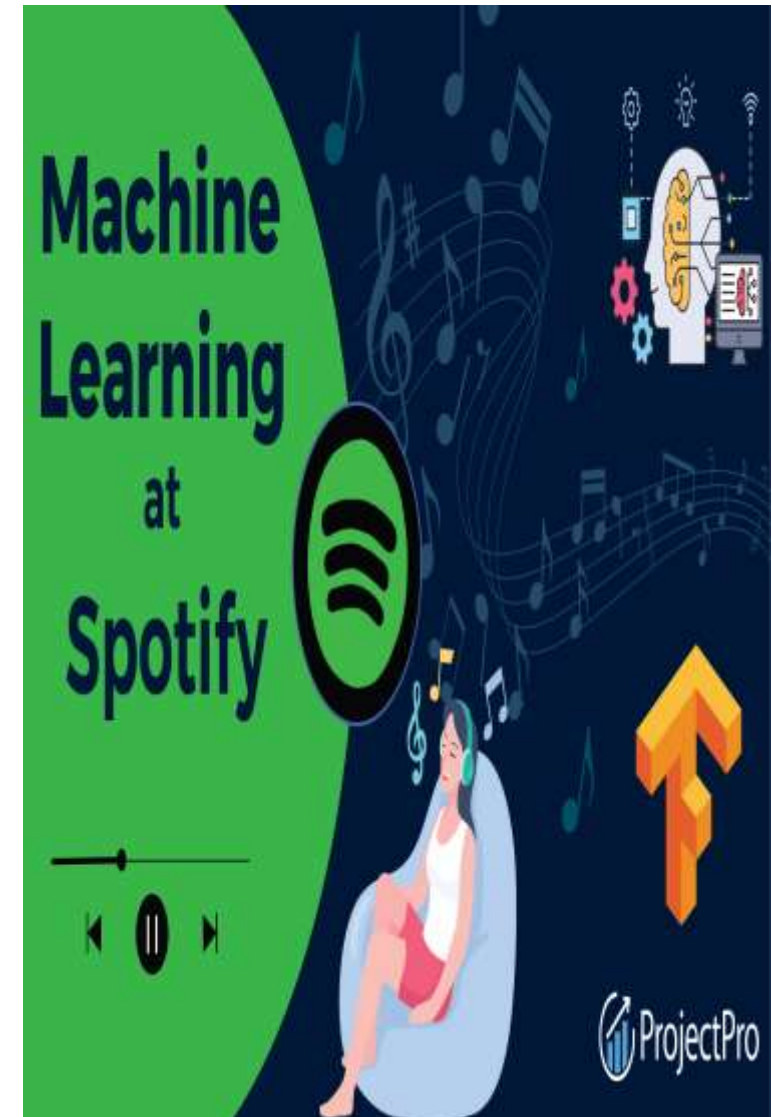
ABSTRACT

- 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.

PROBLEM STATEMENTS

1. **Podcast Popularity Prediction** – Many podcasters struggle to predict which episodes will gain traction, leading to ineffective content planning.

2. **Listener Retention Analysis** – High listener drop-off rates impact overall engagement, but identifying the reasons behind them remains a challenge.

3. **Personalized Podcast Recommendations** – Users often find it difficult to discover relevant podcasts due to a lack of personalized recommendations.

EXISTING SYSTEM

- Traditional content analysis relies on manual observation and intuition, leading to **inefficient and inaccurate predictions**.
- Podcast platforms offer **basic analytics** (e.g., total plays, likes, shares) but lack **advanced predictive capabilities**.
- **No automated recommendation system** for personalized content suggestions.
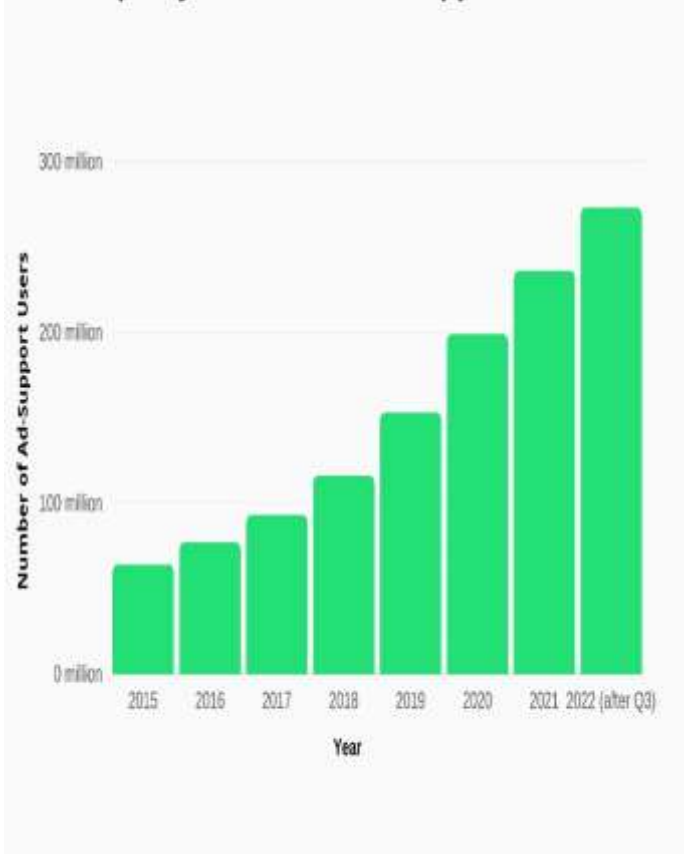
LIMITATIONS:

- Lack of **predictive insights** into podcast popularity.
- No detailed analysis of **listener drop-off behavior**.

Podcast
Episode PAID
Episode
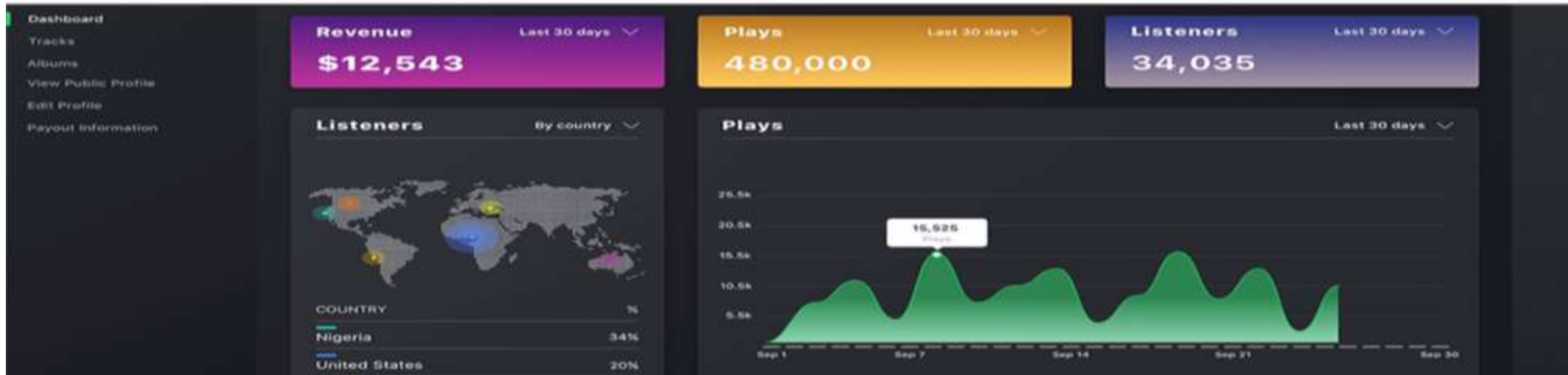Episode

PROPOSED SYSTEM

Spotify: Number of Ad-Support Users

- Utilize **IBM SPSS** for data preprocessing, statistical analysis, and predictive modeling.

- Develop **a machine learning-based popularity prediction model** using regression and classification techniques in SPSS.

- Implement **listener retention analysis** to identify drop-off trends and improve engagement strategies.

- Create a **personalized podcast recommendation system** based on listener preferences and content similarity.

- Use **SPSS decision trees, clustering, and regression models** to generate actionable insights for podcasters and marketers.

.

MODULES IDENTIFIED

1. **Data Collection & Preprocessing** – Gather and clean Spotify podcast data using IBM SPSS.

2. **Podcast Popularity Prediction** – Use regression and decision trees to forecast episode success.

3. **Listener Retention Analysis** – Identify drop-off trends and engagement patterns.

4. **Personalized Recommendation System** – Apply clustering rules for tailored suggestions.

5. **Visualization & Reporting** – Generate insights through dashboards and reports in IBM SPSS.
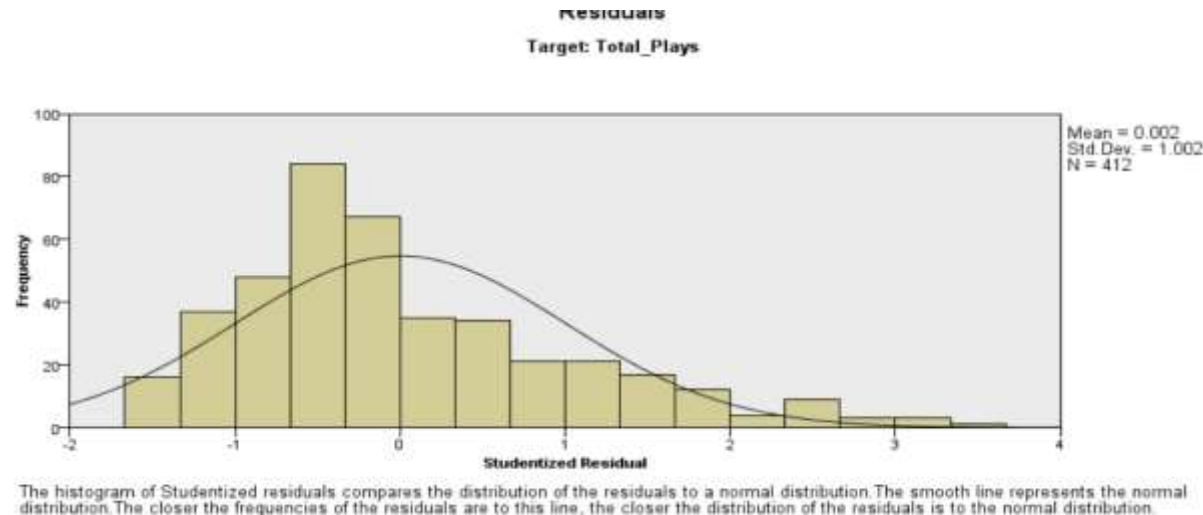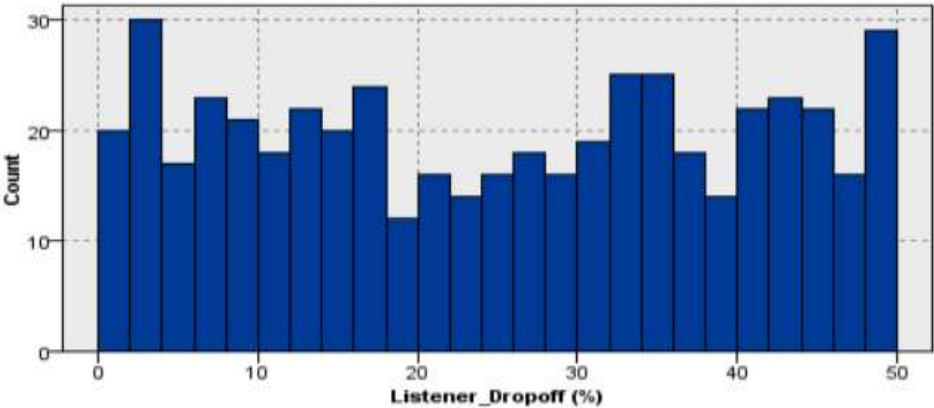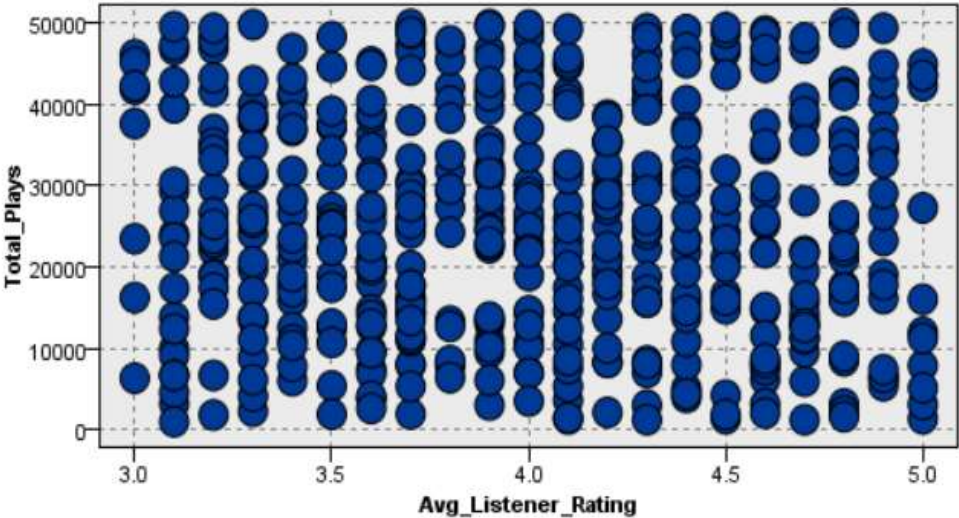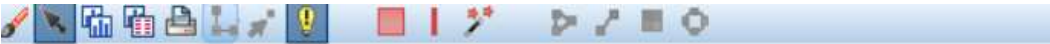
OUTPUT
Data collection and preprocessing:

Podcast_ID	Episode_ID	Title	Duration (mins)	Release_Date	Genre	Host_Popularity	Avg_Listener_Rating	Total_Plays	Likes	Shares	Comments	Completion_Rate (%)	Ad_Count	Listener_Dropoff (%)
164	1	Episode 1 - Business Insig...	88	2023-11-27	Comedy	High	3.700	27785	1016	997	755	73	1	2
170	2	Episode 2 - Lifestyle Insights	58	2023-04-02	Crime	High	4.200	24950	491	129	107	95	0	5
185	3	Episode 3 - Comedy Insights	55	2023-10-13	Comedy	Low	4.300	31319	108	83	58	59	2	4
158	4	Episode 4 - Education Insig...	51	2023-10-03	Health	Medium	4.400	15794	383	209	179	74	3	2
134	5	Episode 5 - Sports Insights	20	2023-04-01	Tech	Low	4.600	45550	2383	540	78	93	3	
159	6	Episode 6 - Sports Insights	23	2023-07-17	Comedy	Medium	4.300	22222	987	398	148	84	4	1
115	7	Episode 7 - Lifestyle Insights	35	2023-03-27	Health	Low	3.500	27300	553	425	294	100	5	
172	8	Episode 8 - Health Insights	80	2023-05-16	Health	High	3.200	36934	134	90	61	57	0	4
115	9	Episode 9 - Crime Insights	44	2023-09-05	Sports	High	3.000	6355	304	207	175	51	0	4
148	10	Episode 10 - Health Insights	75	2023-02-04	Crime	Low	4.600	48675	1287	428	416	99	0	
112	11	Episode 11 - Sports Insights	51	2023-10-05	Lifestyle	Medium	3.900	31065	1491	1014	439	93	5	
109	12	Episode 12 - Sports Insights	20	2023-12-08	Lifestyle	Low	4.000	45429	3007	2742	104	59	5	4
116	13	Episode 13 - Crime Insights	49	2024-01-19	Tech	High	4.100	41506	1640	608	448	77	1	2
163	14	Episode 14 - Comedy Insig...	77	2023-01-12	Health	High	3.300	37896	3134	1309	577	55	2	4
129	15	Episode 15 - Crime Insights	80	2023-09-20	Tech	High	3.700	47369	3574	2485	1871	86	3	1
140	16	Episode 16 - Business Insi...	90	2023-04-30	Tech	Medium	4.600	34591	2904	2147	547	76	1	2
165	17	Episode 17 - Health Insights	73	2023-07-06	Crime	Low	4.400	5104	132	93	42	56	0	4
167	18	Episode 18 - Lifestyle Insig...	18	2023-09-05	Lifestyle	Medium	3.100	3235	304	231	166	97	4	3
120	19	Episode 19 - Crime Insights	75	2023-08-02	Health	Low	3.700	16323	1419	1086	704	62	2	3

Popularity prediction

Residuals

Target: Total_Plays

Mean = 0.002
Std.Dev. = 1.002
N = 412

The histogram of Studentized residuals compares the distribution of the residuals to a normal distribution. The smooth line represents the normal distribution. The closer the frequencies of the residuals are to this line, the closer the distribution of the residuals is to the normal distribution.

Listener Retention Analysis

Histogram of Listener_Dropoff (%) #1

Graph	Model	Host_Popularity	No. Records in Split	No. Fields Used	Overall Accuracy (%)
		High	168	1	30.357
		Low	163	1	28.221
		Medium	169	1	27.219

Recommedations
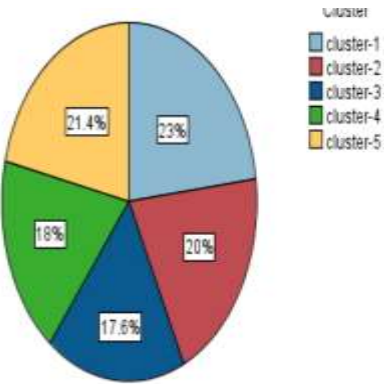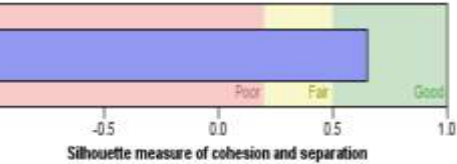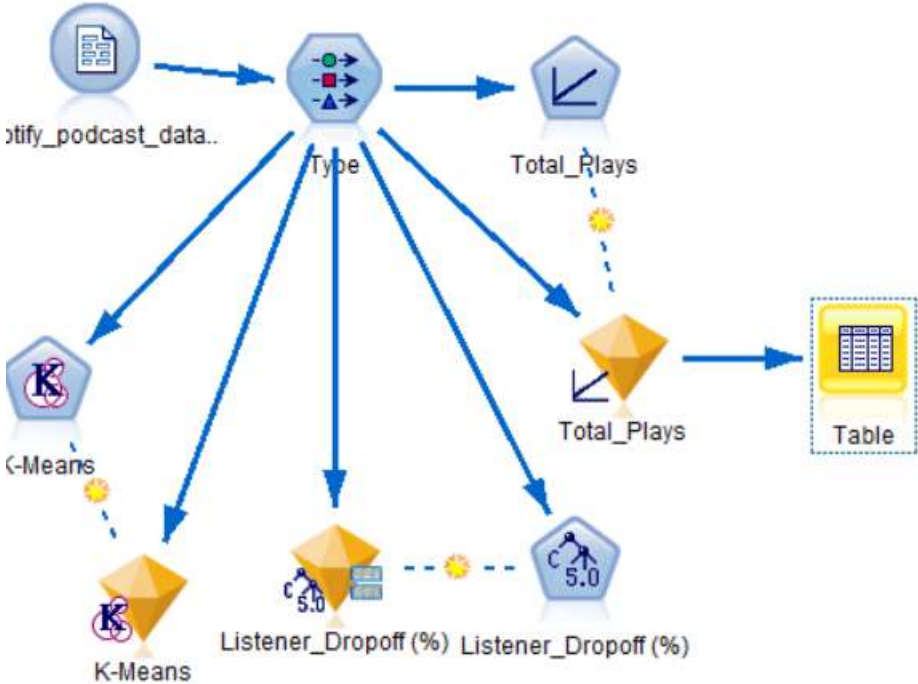
CONCLUSION

- 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.