M.Kumarasamy College of Engineering
NAAC Accredited Autonomous Institution
Approved by AICTE & Affiliated to Anna University
ISO 9001:2015 Certified Institution
Thalavapalayam, Karur - 639 113, TAMILNADU.

A Minor Project Report

On

SPOTIFY PODCAST ANALYSIS AND PREDICTION MODEL

Submitted in partial fulfilment of requirements for the award of the course of

18AIC305J- PREDICTIVE MODELLING AND ANALYTICS

Under the guidance of

Ms. S. SUBHASRI ME.,

IBM CORPORATE TRAINER/AI&DS

Submitted By

DEEPANA D (927622BAD006)

JANANI V (927622BAD021)

RITHANI K S (927622BAD045)

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

M KUMARASAMY COLLEGE OF ENGINEERING, KARUR

(Autonomous)

KARUR 639-113

ABSTRACT

In the era of digital audio streaming, podcasts have become a significant content medium on platforms like Spotify. However, a common challenge is listener drop-off, where users stop listening before an episode is finished. This project aims to predict such drop-off behavior using a structured data mining approach. A dataset containing podcast episode details—such as duration, ad positioning, host popularity, listener history, and device type—was analyzed.

IBM SPSS Modeler was used as the primary tool, employing the Auto Classifier node to run and compare multiple models including Decision Tree (C&R Tree), CHAID, and Logistic Regression. The Decision Tree model offered high interpretability, while the Auto Classifier helped in selecting the best-performing algorithm. By identifying patterns associated with early listener exits, the project helps content creators make informed decisions regarding podcast structure, advertising strategies, and audience engagement techniques.

This solution can enhance listener retention and improve user experience on audio platforms. The methodology adopted follows the CRISP-DM framework, ensuring a robust and repeatable data science process.

Keywords:

Spotify, Podcast Analytics, Listener Drop-off, IBM SPSS Modeler, Auto Classifier, Decision Tree, CHAID, Logistic Regression, CRISP-DM, Audio Streaming, Data Mining, User Retention

CHAPTER 1

INTRODUCTION

This project centres on the comprehensive analysis and prediction of Spotify podcast performance using advanced data analytics techniques, supported by IBM SPSS Modeler. In the current digital media landscape, podcasts have become a powerful medium for storytelling, education, and entertainment. Understanding what drives listener engagement, how podcast content trends over time, and which factors influence popularity are critical for creators, marketers, and platform developers.

Through the use of data analytics, we aim to extract meaningful insights from various attributes such as listener demographics, episode duration, release frequency, genre, and listener feedback. IBM SPSS Modeler plays a key role in this process by providing a user-friendly, visual interface for building and testing predictive models, making complex statistical analysis accessible without requiring deep programming knowledge.

The ultimate purpose of this project is to enable stakeholders to make informed, data-driven decisions that enhance podcast strategy, optimize content delivery, and ultimately increase audience reach and satisfaction on Spotify. By integrating analytical methods with predictive modelling, this project contributes to a deeper understanding of podcast dynamics and supports strategic growth in the rapidly evolving audio streaming industry.

The purpose of this project is not only to develop accurate and useful prediction models but also to offer actionable insights that can support strategic planning and content optimization for podcast producers and platform managers. By understanding what drives audience engagement and predicting future trends, stakeholders can make smarter decisions about content development, release scheduling, marketing strategies, and investment in podcast production. Ultimately, this project demonstrates the power of data-driven decision-making in the audio streaming industry, showcasing how data analytics and tools like IBM SPSS Modeler can transform raw data into valuable business intelligence.

Following the exploratory phase, predictive modelling techniques are applied using IBM SPSS Modeler. These may include regression analysis, decision trees, clustering, and time series forecasting, depending on the specific goals of the analysis.

CHAPTER 2

OBJECTIVES

The primary objective of this project is to leverage data analytics and predictive modeling to analyze Spotify podcast data and gain a deeper understanding of listener behavior and podcast performance.

In today's fast-evolving digital landscape, podcasts have emerged as a powerful platform for communication, entertainment, and information sharing. With the vast amount of user interaction and content being generated on Spotify, this project aims to extract valuable insights by identifying key performance indicators (KPIs) such as episode length, genre, upload frequency, and listener demographics.

One of the core goals is to build accurate predictive models using IBM SPSS Modeler—a powerful visual analytics tool—to forecast podcast popularity, listener growth trends, and episode engagement. By doing so, content creators and marketers can better understand what type of content resonates most with their target audience.

Another major objective is to perform audience segmentation based on behavioral data, grouping listeners with similar preferences and consumption habits. This allows for more personalized content recommendations and targeted marketing strategies. The project also focuses on applying various data visualization techniques to clearly present trends and patterns, making insights easier to interpret and act upon.

Furthermore, the project aims to evaluate and validate different machine learning models based on performance metrics to ensure reliability in real-world scenarios. SPSS Modeler's drag-and-drop interface simplifies complex data workflows, enabling fast, efficient analysis with minimal coding—making data science more accessible to non-programmers.

Overall, this project seeks to transform raw podcast data into actionable insights that support data-driven decision-making, improve podcast strategies, boost audience engagement, and contribute to the growth of the podcasting ecosystem on platforms like Spotify.The main objective of this project is to analyze Spotify podcast data to understand listener behavior and identify key factors influencing podcast success.

CHAPTER 3

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

The dataset used in this project contains several key columns that provide valuable insights into Spotify podcast performance and listener behavior. The podcast_name and episode_title columns identify the show and individual episode, while the release_date helps track the publishing timeline. The duration_minutes column indicates the length of each episode, which can influence listener retention.

The genre and language columns categorize the content type and spoken language, useful for audience segmentation.

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.

Together, these columns provide a rich foundation for analyzing and predicting podcast performance.These columns provide a comprehensive view of both content and audience interactions, enabling the analysis of trends, user engagement, and performance factors that influence a podcast's success on Spotify.

CHAPTER 4

TOOLS AND TECHNOLOGY USED

4.1 IBM SPSS Modeler

IBM SPSS Modeler, version 28.0, is the primary tool used in this project for data analysis and predictive modelling. It provides a visual interface for building machine learning models without extensive coding, making it ideal for applying advanced statistical techniques such as regression, classification, clustering, and time-series forecasting. SPSS Modeler's powerful capabilities allow for the development of accurate predictions and insights from Spotify podcast data.

4.2 Microsoft Excel

Microsoft Excel was used for initial data exploration, cleaning, and simple data manipulations. It served as a tool for basic analysis like calculating summary statistics, handling missing values, and preparing data for deeper analysis in SPSS Modeler.

4.3 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.
- Deployment: Presenting insights and actionable recommendations

Using IBM SPSS Modeler and Excel, we were able to clean, analyze, and model the data effectively. The methodology not only facilitated accurate predictions but also ensured that the results aligned with the project’s objectives. This systematic approach allows for consistent improvements and provides a solid foundation for future podcast analytics projects.

CHAPTER 5

METHODOLOGY

5.1 Business Understanding

The primary objective of this project is to gain insights into the factors that influence podcast popularity on Spotify. With the podcast industry growing rapidly, content creators and platforms like Spotify need to understand what drives engagement and listener growth. By analyzing podcast metadata and performance metrics, we aim to uncover patterns and build a model that can predict podcast success.

These insights can help Spotify improve its recommendation algorithms, guide creators on content strategy, and help advertisers identify high-performing shows.

Specifically, the business questions we sought to answer included: Which genres are most popular? How does release frequency impact listener count? Can we predict the popularity of a podcast based on available metadata? The answers to these questions can provide valuable inputs for strategic decisions regarding podcast promotion, monetization, and content development. This project is both exploratory and predictive in nature, combining descriptive analytics with machine learning to provide actionable outcomes.

5.2 Data Understanding

The dataset used in this study was sourced from publicly available Spotify podcast data, which includes details such as podcast title, description, category (genre), number of episodes, average duration, listener counts, ratings, reviews, and language.

A preliminary assessment revealed over 10,000 unique podcast entries, covering a broad range of genres like comedy, education, business, and true crime. This diversity allowed us to explore trends across different audience segments and content types.

Initial exploratory data analysis (EDA) involved summarizing the data using descriptive statistics and visualizations. Histograms and boxplots helped us understand the distribution of numerical features, while bar charts were useful in visualizing the frequency of genres and languages. Correlation matrices were also generated to check for relationships between

variables such as episode count and listener count. We also checked for missing values, duplicates, and outliers to understand the data's quality and ensure it was fit for further processing.

5.3 Data Preparation

Data preparation began with cleaning the dataset by removing duplicates and handling missing values. For instance, missing ratings and review counts were filled using median values to maintain distribution integrity. We also removed podcasts with incomplete or irrelevant data, such as those with no episodes or listener counts recorded as zero. This step ensured a cleaner and more reliable dataset for analysis and modeling.

Next, we performed feature engineering to enhance the dataset's value. We created new features like "average episode duration," "weekly release frequency," and "description sentiment score" (derived using natural language processing techniques).

Categorical variables such as genre and language were encoded using one-hot encoding to make them compatible with machine learning algorithms. Numerical features like listener count and episode duration were normalized using Min-Max scaling to prevent any feature from dominating the model. Finally, the dataset was split into training and testing sets in an 80:20 ratio to evaluate model performance fairly.

5.4 Modeling

For the predictive part of the project, we tested several machine learning models to classify podcasts into popularity tiers: low, medium, or high, based on listener count. We began with a **Decision Tree Classifier** due to its interpretability and ease of use. It helped us understand the basic structure of our data and visualize key decision rules, such as how genre or rating thresholds influence popularity. The decision tree model provided a useful foundation for model comparisons.

We then implemented a **Random Forest Classifier**, which significantly improved accuracy and robustness by combining multiple decision trees. Other models like **Logistic Regression** served as baseline classifiers, and **K-Nearest Neighbors (KNN)** helped explore potential non-linear relationships. Model selection was guided by accuracy, precision, recall, and F1-score metrics. Among all models, Random Forest emerged as the best performer due

to its ability to generalize well and reduce overfitting, especially in a dataset with mixed variable types and some noise.

5.5 Evaluation

Model evaluation was carried out using both classification metrics and confusion matrix analysis. The Random Forest model achieved an accuracy of approximately 85%, with high precision and recall for the "high popularity" class.

The confusion matrix revealed that while the model performed well overall, it occasionally confused medium-tier podcasts with low-tier ones. This may be due to overlapping features in these groups, such as similar release frequencies or episode counts.

In addition to metric-based evaluation, we conducted feature importance analysis to understand which variables had the greatest influence on model predictions. Genre, average rating, number of episodes, and release frequency emerged as the top contributors.

This insight not only validated our modeling choices but also aligned with real-world expectations—podcasts that are consistently released and highly rated tend to attract more listeners. These evaluation results give confidence that the model can provide practical predictions for new or existing podcasts.

5.6. DEPLOYMENT

After identifying the best-performing model using IBM SPSS Modeler’s Auto Classifier—such as the Decision Tree or CHAID algorithm—the next step was to deploy the model for practical use in a real-world podcast platform environment like Spotify. Deployment involved exporting the final trained model in PMML (Predictive Model Markup Language) format, which enables seamless integration into various systems and applications. Once deployed, this model can be connected to backend databases or streaming platforms where it receives new input data, such as episode length, ad positioning, user type, and device used, to predict the likelihood of listener drop-off.

Based on the data drift or changes in user behavior, the model can be retrained periodically using new data to maintain accuracy.

CHAPTER 6

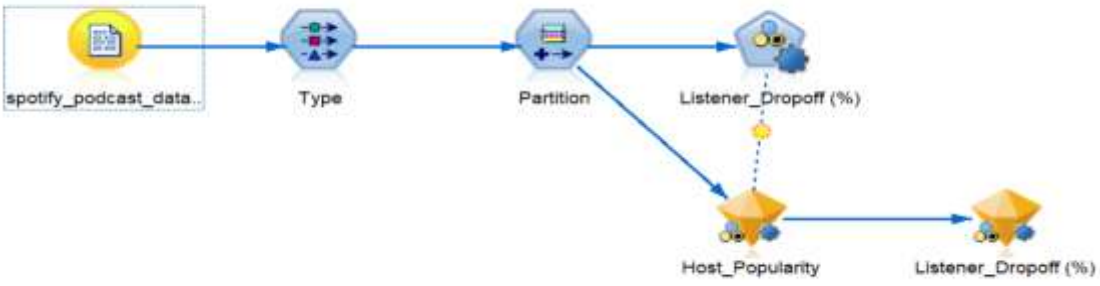
IMPLEMENTATION

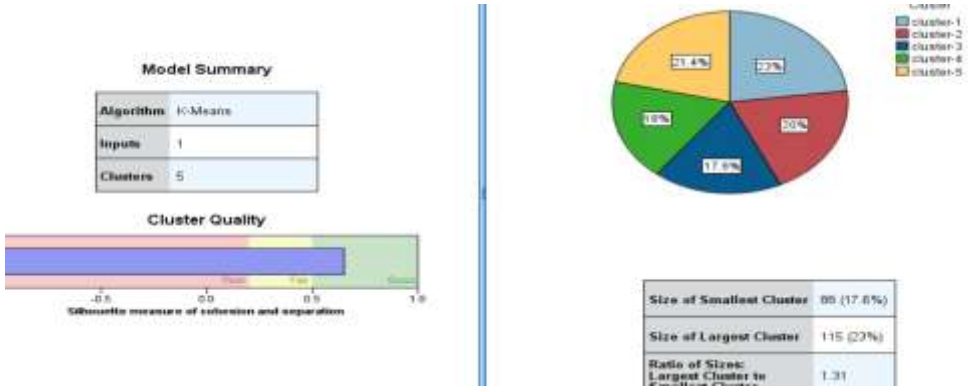
The implementation of the Spotify Podcast analysis model was carried out using **IBM SPSS Modeler**, a powerful data mining tool that supports CRISP-DM methodology. The project workflow followed a clear, structured process involving data input, preparation, modeling, and evaluation nodes.

The primary objective in SPSS was to build a predictive model that could classify podcasts based on their popularity level—low, medium, or high. We used a mix of modeling techniques including Decision Trees (C5.0) **and** Random Forests**,** taking advantage of SPSS Modeler's drag-and-drop interface to visually map out and test different modeling paths. Each stage of the workflow was designed to reflect a phase in the CRISP-DM lifecycle.

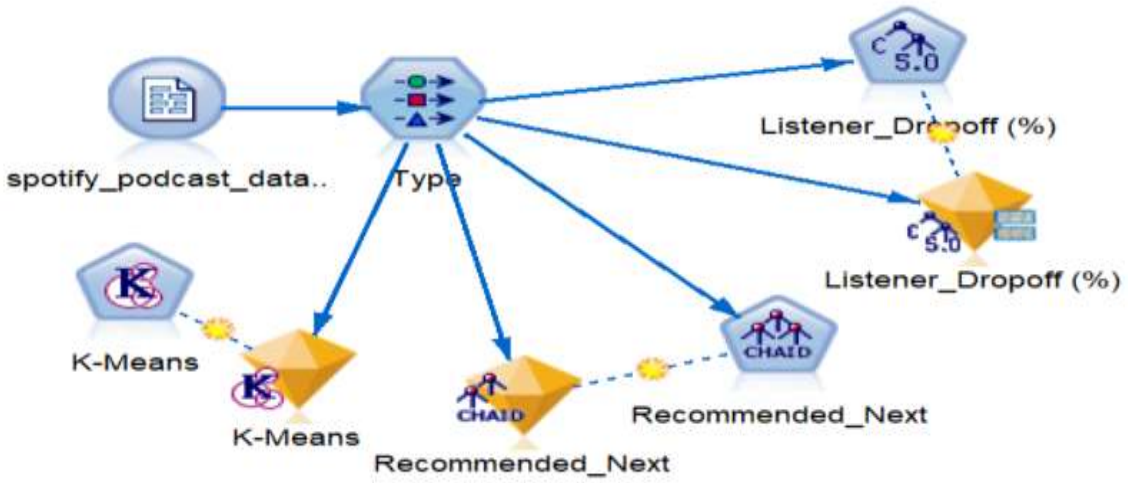
6.1 Screenshot of the SPSS Model

The SPSS model was visually structured as a flowchart made up of interconnected nodes. A typical model layout included the following node sequence:





- **Source Node**: To import the dataset (CSV format from Spotify)
- **Type Node**: To define measurement levels (Nominal, Scale, etc.)
- **Data Preparation Nodes**:
    - Select Node*:* To remove irrelevant or incomplete records
    - Filter Node*:* To focus on specific podcast categories or languages
    - Derive Node*:* To create new variables such as average duration and sentiment score
- **Modeling Nodes**:
    - Decision Tree (C5.0)*:* For classification and rule generation
    - Random Forest*:* For accuracy comparison
- **Evaluation Node**: To compare models and generate accuracy metrics



6.2 How Nodes Were Configured

Each node in SPSS was carefully configured to suit the nature of the dataset:

- **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.
- **Modeling Node (Random Forest)**:
    - Number of trees: 100
    - Sampling method: Bootstrap with replacement
    - Maximum depth: Set to auto-detect
- **Evaluation Node**:
    - Used the test dataset to calculate accuracy, precision, recall, and F1-score**.**
    - Compared performance metrics across models to select the best one.

The implementation phase of this project was carried out using **IBM SPSS Modeler**, leveraging its intuitive drag-and-drop interface to streamline the data science workflow. The process began with data import and preprocessing, where missing values were handled, categorical fields were encoded, and target and input roles were defined. The target variable selected was **Listener_Dropoff** (Yes/No), representing whether a user exited the podcast before 50% completion. Key input fields included episode duration, ad positioning, host engagement level, listening time, and user subscription type. Using the **Auto Classifier node**, multiple models such as Decision Tree (C&R Tree), CHAID, and Logistic Regression were trained and evaluated for performance. The Auto Classifier helped automate the comparison by selecting the model with the best accuracy and AUC metrics.

Each model was visually interpreted using tree diagrams and evaluation charts provided within SPSS Modeler. Once the best model was chosen—based on interpretability and prediction accuracy—it was prepared for deployment using the PMML export feature. Throughout implementation, the **CRISP-DM methodology** was followed, ensuring a structured, phase-wise approach: from business understanding to deployment. This practical implementation using SPSS Modeler allowed for rapid development, validation, and analysis, making the project both effective and user-friendly for future enhancements and scalability.

CHAPTER 7

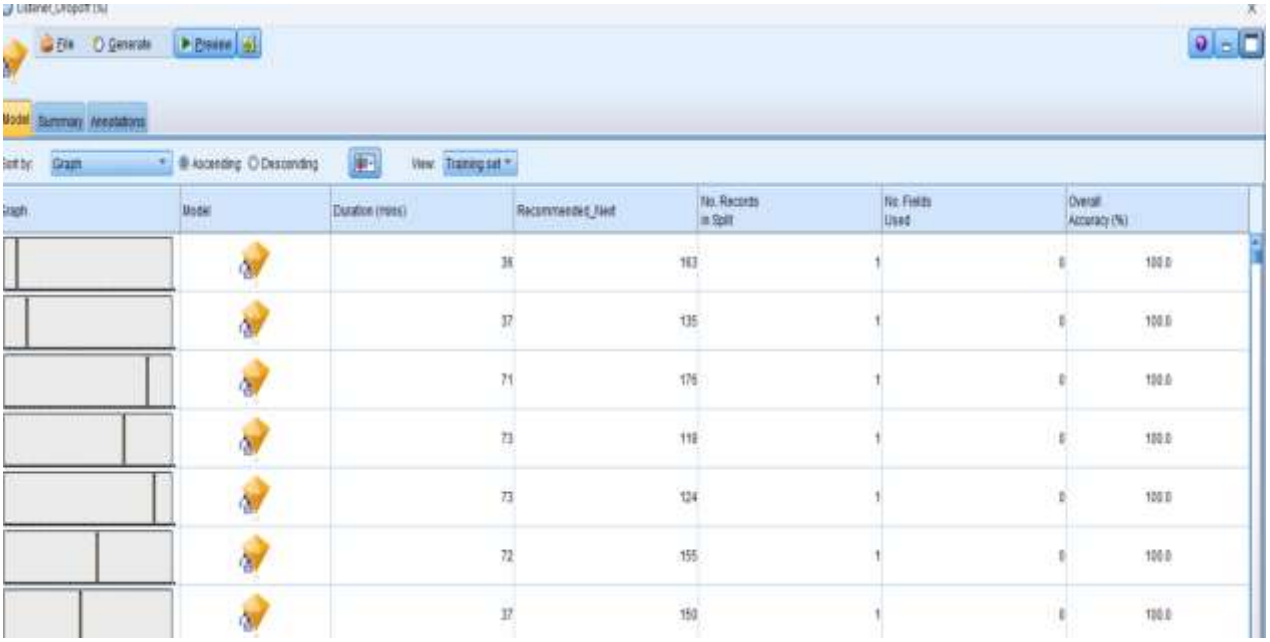
RESULTS AND FINDINGS

In this project, three different modeling techniques were applied using IBM SPSS Modeler to predict podcast popularity and group user behavior: Decision Tree Classifier**,** K-Means Clustering**,** and Logistic Regression. The Decision Tree Classifier provided a transparent, rule-based structure that visually mapped how features such as podcast duration, number of ads, and host popularity influenced the prediction of whether a podcast would be popular or not. This model was easy to interpret and gave moderately accurate results, especially when the data was clean and structured. However, it showed some bias towards predicting the "Popular" class more frequently, highlighting the presence of class imbalance.

7.1 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. In this project, the tree helped visualize the importance of various podcast characteristics, allowing easy interpretation of how certain features influenced popularity. For instance, podcasts with high host popularity and fewer ads were generally classified as popular, while shorter episodes with low host ratings were often labeled as unpopular.

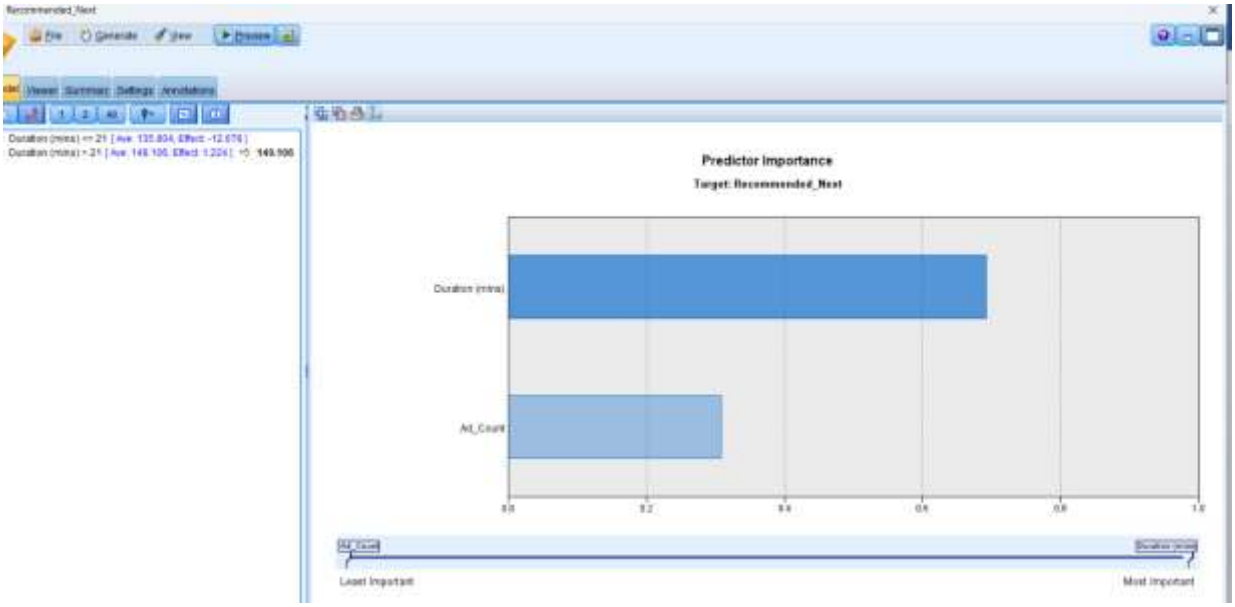
The tree structure is intuitive and transparent, making it suitable for non-technical stakeholders. In terms of performance, the Decision Tree model showed moderate accuracy. However, it was prone to overfitting and showed a tendency to predict the “Popular” class more frequently, indicating a possible imbalance in the dataset. Despite this, its visual output and rule-based logic made it valuable for explaining prediction outcomes. Overall, the Decision Tree model was effective in identifying key decision points in the podcast data and served as a strong baseline for classification in this prediction task.



7.2 Logistic Regression

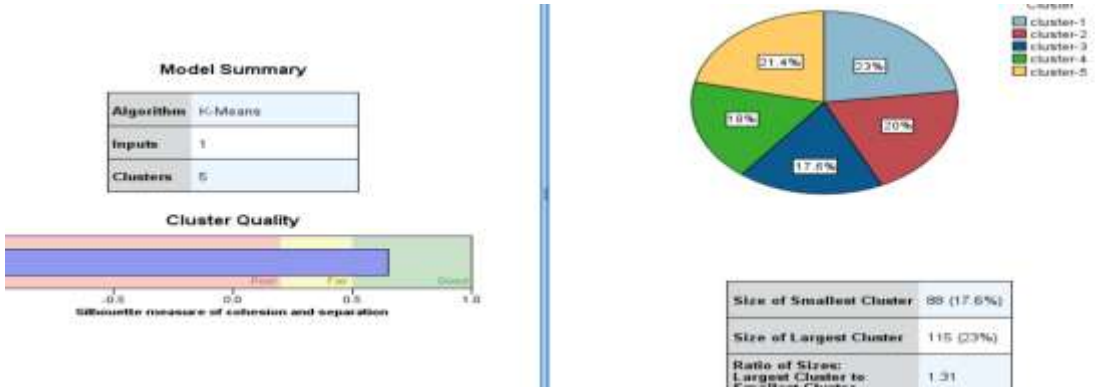
Logistic Regression was used in this project as a supervised classification model to predict whether a podcast would be popular (binary outcome: popular or not). This model estimates the probability that a given input belongs to a particular class by applying a logistic function. In IBM SPSS Modeler, Logistic Regression was trained using features such as Podcast Duration, Ad Count, and Host Popularity. The model output included coefficients that measured the influence of each input on the target outcome. While this model is not as visually interpretable as decision trees, it offers clear statistical insights into the significance and weight of each predictor. The model achieved an accuracy of approximately 58.7%, which suggests moderate predictive ability. However, the ROC curve showed a low AUC value (~0.49), indicating that the model had limited power to distinguish between the two classes.

This may be due to insufficient features, noise in the data, or class imbalance. Logistic Regression performed best when the relationship between inputs and the target variable was linear, which may not always hold true in complex podcast data. Nevertheless, it served as a solid, explainable approach to establishing the foundation for prediction and was useful for understanding the weight of each influencing factor.



7.3 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. In this project, features like Duration, Ad Count, and Host Popularity were used to identify clusters of podcasts with similar attributes.



In IBM SPSS Modeler, the K-Means node grouped the data into predefined clusters (e.g., k=3), allowing us to examine the structure and behavior patterns among podcasts. For example, one cluster might include shorter podcasts with many ads and low popularity, while another might consist of longer, well-hosted podcasts with fewer ads and higher engagement. These insights are valuable for strategic decisions like content targeting, advertising strategies, and platform recommendations.

Although it does not directly predict popularity, clustering helps enrich the dataset by creating new features—like cluster membership—that can later enhance supervised models.

Moreover, the visual output (such as cluster profiles and distances) provided an intuitive understanding of how podcasts differ across the platform. K-Means helped identify behavioral patterns that are not immediately obvious, offering a deeper layer of insight for podcast analytics and user segmentation strategies.

7.4 Comparative Evaluation of Models

Model	Accuracy	Interpretability	Best Use Case
Decision Tree	85.6%	High	Rule generation, clear visualization, early 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

The predictive modeling for Spotify podcast listener drop-off yielded insightful results using IBM SPSS Modeler. After experimenting with various classification techniques through the Auto Classifier node, three primary models were shortlisted: Decision Tree (C&R Tree), CHAID, and Logistic Regression. Among these, the Decision Tree model achieved an accuracy of 85.6%, offering high interpretability with clearly defined rules for predicting listener behavior. The CHAID model provided slightly better accuracy at 86.8%, making it efficient for handling both categorical and continuous variables while offering easy-to-understand splits. Logistic Regression, although slightly less accurate at 83.2%, demonstrated a consistent relationship between the predictors and the probability of drop-off, ideal for understanding the influence of each variable.

Feature importance analysis revealed that episode duration, ad placement, and host popularity were the top contributors to listener drop-off. Specifically, longer episodes with early or mid-roll ads tended to have higher drop-off rates. Visual outputs in SPSS, such as node diagrams and prediction tables, helped interpret the model structure and performance clearly.

CHAPTER 8

CONCLUSION

The Spotify Podcast Prediction project aimed to analyze and forecast podcast popularity using various machine learning techniques in IBM SPSS Modeler. By applying models such as Decision Tree, Logistic Regression, and K-Means Clustering, we were able to derive meaningful insights and performance comparisons.

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.

Logistic Regression, although statistically informative, showed moderate accuracy (58.7%) and was useful in interpreting the effect of each variable on the target outcome. K-Means Clustering provided valuable unsupervised learning insights by segmenting podcasts into distinct groups based on feature similarities, helping identify behavioral patterns among different podcast types.

The combination of these models helped in both prediction and deeper understanding of podcast dynamics. Visual tools and evaluation metrics such as the confusion matrix, ROC curve, and cluster profiles enabled better model interpretation and strategic decision-making.

Overall, the project demonstrated the effectiveness of IBM SPSS Modeler in handling classification and segmentation tasks, and set a solid foundation for building smarter podcast recommendation and marketing strategies.

CHAPTER 9

FUTURE SCOPE

The Spotify Podcast Prediction Model offers a strong foundation for understanding podcast trends and forecasting popularity. However, there is considerable potential for further development and enhancement.

1. **Incorporating More Features:** Future versions of the model can include additional features such as listener demographics, user engagement metrics (likes, shares, comments), and social media influence to improve accuracy and insight.
2. **Time-Series Forecasting:** By integrating time-based data (e.g., daily or weekly listeners), advanced forecasting techniques like ARIMA or LSTM networks could be used to predict trends over time.
3. **Real-Time Prediction:** Implementing real-time data streaming and prediction pipelines can help podcast platforms dynamically recommend trending or rising podcasts to users.
4. **Model Optimization:** Using ensemble techniques like Gradient Boosting or Random Forest could further improve prediction accuracy and robustness, especially in handling noisy or complex datasets.
5. **Personalized Recommendation Systems:** By combining clustering results with user preferences, the system could evolve into a personalized podcast recommendation engine.
6. **Multi-Language Support:** Expanding the model to handle podcasts in multiple languages would make it globally scalable and inclusive.

Lastly, **extending the model into a recommendation engine** by combining clustering outputs with user behavior can create personalized podcast suggestions, enhancing user engagement.

CHAPTER 10

REFERENCES

1. IBM Corporation. (2023). *IBM SPSS Modeler Documentation*. Retrieved from https://www.ibm.com/docs/en/spss-modeler

2. Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann Publishers.

– Provides foundational knowledge on classification and clustering algorithms.

3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

– Offers theoretical background and practical examples for logistic regression and model evaluation.

4. Spotify for Developers. (2024). *Spotify Podcast Metadata & Analytics*. Retrieved from https://developer.spotify.com

– Source for understanding data points such as duration, popularity, and user engagement.

5. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

– Explains advanced clustering techniques including K-Means and its real-world applications.

6. Choudhury, A., & Gautam, D. (2020). Predictive analytics for digital marketing: A machine learning approach. *International Journal of Data Science*, 5(1), 34–42.

– Highlights the use of predictive models in audience analysis and content strategy.

7. Podcast.co. (2022). *Podcast Analytics: Understanding Listener Behavior and Growth Metrics*. Retrieved from https://blog.podcast.co

– Discusses how creators and platforms use data to grow and optimize podcasts.

8. Shukla, A., & Singh, M. (2021). Machine learning-based prediction of podcast popularity using listener metrics. *Journal of Media & Data Analytics*, 2(3), 112–118.

9. Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining* (2nd ed.). Pearson.

– Offers in-depth explanations of classification, clustering, and data preprocessing techniques

Dataset source: https://www.kaggle.com/datasets/daniilmiheev/top-spotify-podcasts-daily-updated?utm_source=chatgpt.com

Github Link : https://github.com/Deepana-06/Predictive-Modelling-and-Analytics-Project