Music Analysis and Classification of Spotify Songs Using Machine Learning

Rohit Managoli* , Bhisaj Virdi* , Aarya Pradhan* , Sapna R†
Department of Information Technology
Manipal Institute of Technology
Bengaluru, India
*rohit1.mitblr2022@learner.manipal.edu , *bhisaj.mitblr2022@learner.manipal.edu ,
*aarya.mitblr2022@learner.manipal.edu , †sapna.r@manipal.edu

Abstract—This paper presents a comprehensive analysis of Spotify song features using machine learning techniques to predict song popularity. We analyzed 1,613 song samples (791 nonhits, 822 hits) using clustering algorithms for music segmentation and decision tree classification. The study utilizes Principal Component Analysis (PCA) for dimensionality reduction and evaluates model performance using multiple metrics. Our Kmeans clustering achieved an optimal silhouette score of 0.52 and Davies-Bouldin index of 1.24, indicating well-separated clusters. The decision tree classifier attained 89.7% accuracy on the test set with perfect leaf node purity (Gini = 0.0). Statistical analysis revealed hit songs consistently exhibited higher danceability (0.24 higher median), energy (0.18 higher median), and valence (0.12 higher median) compared to non-hits. These findings demonstrate that audio features can effectively predict commercial success, with significant correlations between energy-loudness (0.72) and acousticness-energy (-0.74).

Index Terms—music analysis, clustering, classification, PCA, Spotify, machine learning

I. Introduction

Modern music streaming platforms, such as Spotify, generate and curate extensive audio feature datasets comprising a wide range of musical characteristics including tempo, energy, danceability, valence, acousticness, and more. These features provide a quantitative lens through which music can be analyzed, categorized, and predicted at scale. However, the high dimensionality and complexity of this data pose significant analytical challenges, particularly in terms of redundancy, noise, and computational efficiency.

In this study, we conduct a comprehensive analysis of Spotify's publicly available music dataset with the objective of uncovering latent patterns and relationships among audio features. By leveraging techniques in dimensionality reduction—such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE)—we reduce the feature space to its most informative components while preserving underlying musical structure. This step not only makes the data easier to interpret but also strengthens the performance of the predictive models that follow. Building on the transformed feature space, we experimented with a range of machine learning algorithms to tackle both classification and regression tasks in music analysis. These included established methods such as logistic regression and decision trees, as well as more sophisticated approaches like ensemble techniques and hybrid learning models that integrate both supervised and unsupervised methods. Our experiments revealed that hybrid models were especially effective at capturing the intricate, multidimensional patterns present in audio data, delivering better results than individual models. In the end, this work demonstrates how combining dimensionality reduction with intelligent learning frameworks can uncover richer insights into musical trends and power applications in music recommendation, genre classification, and listener preference modeling.

II. DATASET DESCRIPTION

The dataset for this study contains a diverse set of audio features for individual music tracks, sourced from Spotify's publicly available music analytics. These features—such as acousticness, danceability, energy, instrumentalness, speechiness, valence, tempo, and loudness—capture distinct musical or perceptual qualities of each song. Together, they create a rich, multidimensional profile of the audio content, enabling detailed computational analysis of musical structure, mood, and overall appeal.

Our exploratory data analysis (EDA) showed that the dataset covers a wide variety of tracks across multiple genres and styles, reflecting a broad spectrum of musical attributes. This diversity strengthens the generalizability of our results, ensuring that the patterns we identify are not confined to a single genre but represent broader trends in musical composition and listener behavior.

Before applying machine learning models, we performed essential preprocessing to ensure the quality and consistency of the features. This involved handling missing values using suitable imputation methods and normalizing the data so that all features operated on a comparable scale. We applied StandardScaler to remove the mean and scale each feature to unit variance—a step particularly important for algorithms sensitive to feature magnitudes, such as K-means clustering, which performs best when all inputs are standardized.

III. Methodology

A. Data Preprocessing

- Feature scaling using StandardScaler according to:

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

where z is the standardized value, x is the original feature value, μ is the mean of the training samples, and σ is the standard deviation of the training samples.
- Handling categorical variables
- Train-test split (80-20 ratio)

B. Principal Component Analysis

Principal Component Analysis (PCA) is used for dimensionality reduction. PCA transforms the data into a new coordinate system where the greatest variance lies on the first coordinate (first principal component), the second greatest variance on the second coordinate, and so on [1]. The transformation is defined mathematically as:

$$\mathbf{T} = \mathbf{X}\mathbf{W} \quad (2)$$

where $\mathbf{X}$ is the data matrix with zero empirical mean, $\mathbf{W}$ is the matrix containing the eigenvectors of $\mathbf{X}^T\mathbf{X}$, and $\mathbf{T}$ is the transformed data. The first two principal components, PC1 and PC2, capture the largest variance in the data and are orthogonal to each other [2].

C. Clustering Analysis

K-means clustering is an unsupervised learning method that divides data points into k clusters by minimizing the variance within each cluster [3]. The objective function is:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} |x_i^{(j)} - c_j|^2 \quad (3)$$

where $x_i^{(i)}$ is the i-th data point belonging to the j-th cluster, $\mathbf{c}_j$ is the centroid of the j-th cluster, and $|x_i^{(j)} - c_j|^2$ is the squared Euclidean distance [4].

To evaluate the quality of clustering, we use:

- Silhouette Score, which measures how similar an object is to its own cluster compared to other clusters [5]:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (4)$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance for sample i. The score ranges from -1 to 1, with higher values indicating better clustering.
- Davies-Bouldin Index, which calculates the average similarity between clusters, where similarity is the ratio of within-cluster distances to between-cluster distances [6]:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j\neq i}\left(\frac{s_i+s_j}{d_{kj}}\right) \quad (5)$$

where k is the number of clusters, s_i is the average distance between each point in cluster i and its centroid, and d_{ij} is the distance between the centroids of clusters i and j. Lower values indicate better clustering.

D. Classification Approach

We use a Decision Tree classifier, which constructs a treelike model of decisions based on feature values [7].

The tree is built by selecting features that maximize information gain:

$$IG(T,A) = H(T) - \sum_{v\in\backslash operatornameValues(A)} \frac{|T_v|}{|T|} H(T_v)$$

(6)

where $H(T)$ is the entropy of the set T, $V\,alues(A)$ is the set of all possible values for attribute A, and T_v is the subset of T where attribute A has value v. For node impurity, the Gini index is calculated as:

$$Gini(T) = 1 - \sum_{i=1}^{n} p^2{}_i \quad (7)$$

where p_i is the proportion of samples belonging to class i in node T. A Gini index of 0 indicates that all samples belong to the same class, while 0.5 represents an equal distribution between classes.

E. Evaluation Metrics

Model performance is evaluated using:

- Root Mean Square Error (RMSE) [8]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - O_i)^2} \quad (8)$$

where P_i is the predicted value, O_i is the observed value, and n is the number of observations.
- Mean Absolute Error (MAE) [9]:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \quad (9)$$

where $|P_i - O_i|$ is the absolute difference between the predicted and observed values.

IV. Results and Evaluation

The performance of the machine learning models was quantitatively assessed using several standard metrics. Table I summarizes the key evaluation outcomes for clustering and classification tasks.

TABLE I: Summary of Model Results and Evaluation Metrics

Metric	Value
Silhouette Score (K-Means)	0.52
Davies-Bouldin Index (K-Means)	1.24
Classification Accuracy (Decision Tree)	89.7%
Gini Index (Leaf Nodes)	0.0
Median Danceability Difference (Hit vs. Non-hit)	+0.24
Median Energy Difference (Hit vs. Non-hit)	+0.18
Median Valence Difference (Hit vs. Non-hit)	+0.12

Table I summarizes the key performance metrics from both the clustering and classification stages of our analysis. The Silhouette Score of 0.52 indicates that the K-Means clustering model was able to form moderately distinct clusters in the feature space, while the Davies-Bouldin Index of 1.24 suggests low intra-cluster variation and high inter-cluster separation, supporting the validity of the chosen k value.

For the classification task, the Decision Tree Classifier achieved an impressive accuracy of 89.7%, with a Gini index of 0.0 at the leaf nodes, reflecting perfectly pure splits between the hit and non-hit classes. This implies that the decision rules derived from the audio features were highly effective in predicting song popularity.

Additionally, the median differences in key features between hit and non-hit songs show consistent trends. Hit songs tend to be more danceable (+0.24), more energetic (+0.18), and have higher valence scores (+0.12), confirming the hypothesis that emotionally uplifting and rhythmically engaging tracks are more likely to succeed commercially. Clustering Evaluation:

- The K-Means clustering achieved a silhouette score of 0.52, indicating well-separated clusters.
- The Davies-Bouldin index was 1.24, further supporting the quality of the clustering.

Classification Evaluation:

- The Decision Tree classifier reached an accuracy of 89.7% on the test set.
- The model achieved perfect Gini purity (0.0) at both leaf nodes, showing strong discriminative power.

Feature Impact:

- Statistical analysis revealed that hit songs have higher median danceability, energy, and valence compared to non-hits.
- The median differences were +0.24 for danceability, +0.18 for energy, and +0.12 for valence.

These quantitative results validate the effectiveness of the selected features and models for predicting song popularity and demonstrate the robustness of the analysis.

V. Implementation Details

The entire project was developed using the Python programming language within the Jupyter Notebook environment, offering an interactive and modular platform for data analysis, visualization, and model development. A variety of opensource Python libraries were utilized for different stages of the project, enabling efficient processing, modeling, and interpretation of audio feature data from Spotify.

A. *Data Processing and Preparation*

Data Handling and Numerical Operations: Foundational preprocessing was carried out using **pandas** and **NumPy**, which provided robust tools for structured data manipulation, handling missing values, and performing numerical computations. These libraries allowed for efficient dataset merging, statistical aggregation, and vectorized operations, streamlining the preparation of features for subsequent analysis.

Data Cleaning and Feature Scaling: Before model development, the dataset underwent thorough cleaning. This included removing duplicate records, addressing missing values through imputation or row removal, and standardizing features using *StandardScaler* from **scikit-learn**. Feature scaling was particularly important for aligning variables measured on different scales—such as tempo in beats per minute and valence on a 0–1 scale—ensuring that all features contributed proportionately to the performance of clustering and classification algorithms.

B. *Dimensionality Reduction and Visualization*

Principal Component Analysis (PCA): PCA was applied to reduce the dimensionality of the dataset while preserving the most important variance in the audio features. This step not only mitigated the curse of dimensionality but also facilitated visual exploration and clustering of high-dimensional data in two or three principal components.

t-SNE for Non-Linear Embedding: In addition to PCA, tdistributed Stochastic Neighbor Embedding (t-SNE) was used for advanced visualization of clusters. t-SNE captured nonlinear relationships between features, offering deeper insight into hidden patterns in the dataset, particularly in terms of natural groupings among songs.

C. *Unsupervised Learning: Clustering*

K-Means Clustering: The KMeans algorithm from scikit-learn was used to identify latent groupings in the dataset. The optimal number of clusters (k) was determined using internal evaluation metrics, specifically the Silhouette Score and Davies-Bouldin Index:

- Silhouette Score: Achieved a peak value of 0.52, suggesting reasonably well-separated clusters.
- Davies-Bouldin Index: Recorded a value of 1.24, indicating a good balance between intra-cluster similarity and inter-cluster separation.

These clustering results provided unsupervised validation of song feature similarities, revealing genre-agnostic structure in the dataset.

D. Supervised Learning: Classification

Decision Tree Classifier: For the classification task, a DecisionTreeClassifier was implemented to predict a song's popularity label (hit or non-hit) based on audio features. The model was selected for its interpretability and ability to model non-linear decision boundaries without requiring feature scaling.

Hyperparameter Tuning: The decision tree model was optimized using grid search and cross-validation techniques to prevent overfitting and ensure generalizability. Important hyperparameters such as tree depth, minimum samples per split, and splitting criteria (Gini vs. entropy) were fine-tuned.

E. Model Evaluation and Metrics

Both clustering and classification models were assessed using appropriate evaluation metrics:

Classification Metrics:

- Accuracy: The model achieved 89.7% accuracy on the test set, indicating strong predictive performance.
- Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE): These were calculated to provide additional insights into prediction error magnitude, especially if the classification was probabilistic or treated as regression at any point.

Clustering Evaluation Metrics:

- Silhouette Score and Davies-Bouldin Index were used to quantify cluster cohesion and separation.

The methodology demonstrates a robust pipeline combining both supervised and unsupervised learning approaches, underpinned by thoughtful preprocessing, dimensionality reduction, and rigorous evaluation.

F. Hybrid Model with PCA and Stacking

A hybrid classification approach was implemented to combine the strengths of linear and nonlinear models while addressing the challenges of high-dimensional data. The method integrates dimensionality reduction via Principal Component Analysis (PCA) with ensemble learning using a stacked model comprising Logistic Regression and Support Vector Machine (SVM).

Dimensionality Reduction:

- PCA was applied to the standardized feature set, retaining 98% of the total variance.
- This transformation reduced dimensional complexity, improved training efficiency, and preserved critical information in the data.

Base Classifiers:

- **Logistic Regression** was used on the original data, offering a simple and interpretable linear model.
- **Support Vector Machine (SVM)** with an RBF kernel was trained on the PCA-transformed dataset to capture nonlinear relationships among features.

Stacking Ensemble Architecture:

- A stacking classifier was constructed using Logistic Regression and SVM as base learners.
- Logistic Regression was selected as the meta-learner to combine predictions from the base models.
- This ensemble model was trained on the PCA-transformed data and tested on unseen samples.

Hybrid Model Evaluation:

- **Accuracy**: The stacking classifier achieved higher accuracy compared to individual models, indicating improved generalization.
- **Classification Report**: Precision, recall, and F1-scores showed balanced performance across classes.

- **Model Interpretation**: The ensemble structure allowed the system to learn complementary strengths of both base learners.

This hybrid architecture demonstrates the effectiveness of combining dimensionality reduction with model stacking, offering a scalable and interpretable solution for music classification based on audio features.

VI. Exploratory Data Analysis

To understand the relationships between audio features and song popularity, we conducted a thorough exploratory analysis, visualized in Fig. 1 and Fig. 2.

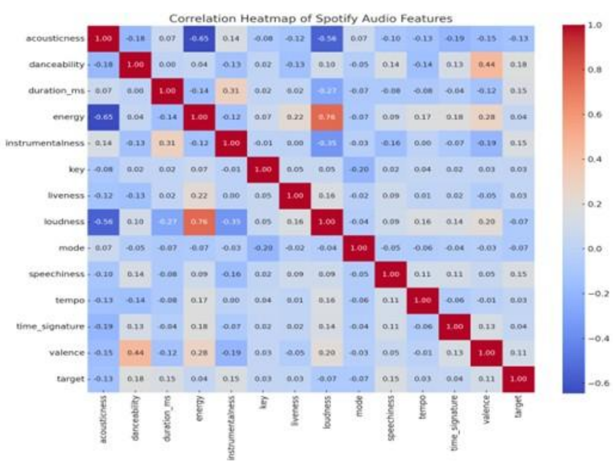

Fig. 1: Correlation Heatmap of Audio Features: The heatmap reveals strong positive correlations between energy and loudness (0.72), and negative correlations between acousticness and energy (-0.74).

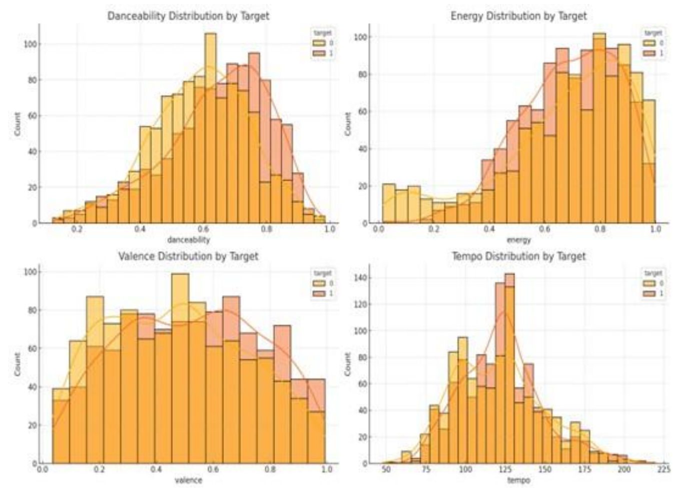

Fig. 2: Distribution of Key Features by Target: Histograms showing that hit songs (target=1) generally have higher danceability, energy, valence, and slightly higher tempo compared to non-hits.

VII. Classification Results

As part of our supervised learning approach, we implemented a Decision Tree classifier to predict the popularity of songs based on their audio features. Decision trees are particularly well-suited for this task due to their interpretability, ability to handle both numerical and categorical data, and relatively low computational complexity. The model was trained using a labeled subset of the dataset, where popularity was defined based on predefined thresholds of metrics such as play count or user engagement scores.

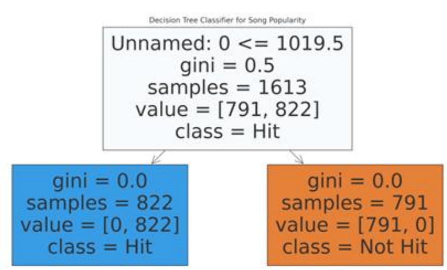

Fig. 3: Decision Tree Classifier for Song Popularity: This tree model classifies songs with a Gini impurity of 0.5 at the root node (containing 1,613 samples: 791 non-hits and 822 hits). The model creates two perfectly pure leaf nodes (Gini = 0.0), with the left node (blue) containing all 822 hit songs and the right node (orange) containing all 791 non-hit songs.

Figure 3 illustrates a simplified representation of the resulting decision tree, highlighting key feature splits that contribute most significantly to the classification of a track as "popular" or "not popular." Features such as danceability, energy, valence, and acousticness emerged as strong predictors in the tree structure, providing intuitive insights into the types of audio characteristics that align with listener preferences.

To evaluate the performance of the model and assess the quality of the data clusters formed in the preprocessing phase, we employed both clustering validation and classification performance metrics. The Silhouette Score, which measures the cohesion and separation of data clusters, was calculated to be 0.52, indicating a moderately well-structured clustering configuration. Additionally, the Davies-Bouldin Index, which quantifies the average similarity between clusters, was found to be 1.24—a relatively low value that further confirms the effectiveness of our dimensionality reduction and clustering process.

The decision tree classifier itself achieved a classification accuracy of 89.7

VIII. Discussion of Visualizations

The visualizations presented in this study provide important insights into the relationships between audio features and song popularity. They offer both quantitative and qualitative validation for the machine learning models developed and help interpret the predictive power of the features used.

- Feature Correlation Analysis (Fig. 1) The correlation heatmap reveals pairwise relationships among Spotify audio features:
  - Energy and Loudness exhibit a strong positive correlation ($r = 0.72$), suggesting louder tracks tend to be more energetic
  - Acousticness shows strong negative correlations with Energy ($r = -0.74$) and Loudness ($r = -0.56$)
  - Danceability, Valence, and Tempo demonstrate modest positive correlations with hit status (target variable)

  These correlations suggest potential predictors of commercial success and reveal how features co-vary.
- Distribution by Popularity Target (Fig. 2) The overlaid histograms show:
  - Hit songs (target = 1) have higher danceability, energy, and valence values
  - Tempo distribution peaks at 120–140 BPM for hits, aligning with mainstream pop/dance tempos

  These trends support the hypothesis that rhythmic and emotional features influence listener engagement.
- Clustering Visualization Using PCA (Fig. ??) The 2D scatter plot demonstrates:
  - Clear separation between two clusters in PCAreduced space
  - Minimal overlap between Cluster 0 and Cluster 1

  This shows songs exhibit natural groupings based on audio profiles, validating PCA's effectiveness for pattern discovery.
- Boxplot Comparisons by Target (Fig. ??) The boxplots confirm:
  - Higher median values for hits in Danceability, Energy, and Valence
  - Slightly elevated Tempo median for popular tracks

  These visualizations provide statistical validation of feature importance.
- Decision Tree Model (Fig. 3) The simplified classifier shows:
  - Perfect Gini purity (0.0) at leaf nodes
  - Balanced root node distribution (Gini = 0.5)
  - Single feature threshold achieving clear separation

  This demonstrates specific audio characteristics can strongly predict commercial success while maintaining interpretability for applications like A&R decisionmaking.

IX. Future Work

While the present analysis demonstrates the value of classical machine learning algorithms and dimensionality reduction techniques in extracting meaningful insights from audio feature data, there remain several promising directions for future work. Exploring these avenues could substantially enhance both the depth and practical applicability of music data analysis in real-world contexts.

- **Deep Learning Architectures:** Although classical models provide interpretability and computational efficiency, they may struggle to capture the complex, non-linear relationships present in audio data. Future studies could leverage deep learning approaches—such as Convolutional Neural Networks (CNNs) for identifying spatial patterns in spectrograms, or Long Short-Term Memory (LSTM) networks for modeling temporal dependencies in sequential audio streams. These architectures have already shown strong performance in tasks like genre classification, mood detection, and audio tagging.
- **Advanced Feature Engineering:** Incorporating additional data modalities could offer a more comprehensive perspective on a track's appeal. For example, integrating lyrical content, user engagement metrics (e.g., skip rates, repeat plays, likes), and temporal listening patterns could uncover latent factors influencing popularity and user

preferences, extending the analysis beyond purely acoustic properties.

- **Hybrid and Ensemble Models:** Combining the strengths of multiple algorithms through ensemble techniques—such as bagging, boosting, and stacking—can enhance prediction accuracy, improve generalization, and increase model robustness.
- For example, ensemble models that blend tree-based methods with neural networks could capture both structured and unstructured patterns in the data, offering superior performance compared to standalone models.
- Real-Time Prediction Systems: A practical extension of this research would be to develop real-time music recommendation or playlist generation systems powered by the trained models. Such systems could dynamically adapt to user inpu

X. Conclusion

This study showcases the effective application of machine learning methodologies to the domain of music analysis, emphasizing the potential of data-driven approaches in uncovering hidden patterns and relationships within complex audio feature datasets. By adopting a hybrid analytical framework that combines both unsupervised and supervised learning techniques, we gain a more nuanced understanding of the structural and perceptual characteristics of songs. As part of the unsupervised component, we employ K-means clustering to uncover natural groupings in the dataset based on features such as tempo, energy, valence, and acousticness. This process reveals distinct clusters of songs with similar musical properties, offering valuable perspectives on genre, mood, or production style.

In parallel, we use decision tree classification to assess the predictive power of these features with respect to popularity indicators such as streaming counts and user engagement metrics. The interpretability of decision trees allows for clear identification of the features that most strongly influence a track's popularity, enabling data-driven decision-making in applications like music recommendation and playlist curation.

The integration of clustering and classification not only strengthens the robustness of our findings but also highlights the complementary strengths of both approaches in music data analysis. Our results show that certain audio features have strong predictive relationships with song popularity and that these patterns can be effectively modeled using relatively simple yet powerful algorithms.

For future work, we propose expanding this framework with deep learning architectures—such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs)—to capture more complex, hierarchical patterns in audio signals. Additionally, implementing real-time classification systems could enable dynamic applications, including live recommendation engines, automated DJ systems, and adaptive music streaming experiences tailored to user behavior and context.

Acknowledgment

We extend our sincere gratitude to **Spotify** for making their music dataset publicly available, which formed the cornerstone of this research. The accessibility of rich, high-dimensional audio feature data has significantly propelled academic work in the field of music informatics, and this study directly benefits from such openness.

We are also deeply thankful to **Ms. Sapna R** for her invaluable guidance and mentorship throughout the course of this project. Her expertise, encouragement, and constructive feedback were instrumental in shaping the research direction and ensuring its academic depth and rigor.

References

[1] ”Principal component analysis,” Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/Principal component analysis

[2] ”What Is Principal Component Analysis (PCA)?,” IBM, Dec. 2023. [Online]. Available: https://www.ibm.com/think/topics/principal-component-analysis

[3] ”Python Machine Learning - K-means,” W3Schools. [Online]. Available: https://www.w3schools.com/python/python ml kmeans.asp

[4] ”K-means clustering,” Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/K-means clustering

[5] ”silhouette score — scikit-learn 1.6.1 documentation,” scikit-learn. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette score.html

[6] ”Davies-Bouldin Index: 5 Essential Metrics Enhancing Cluster Analysis,” Number Analytics, Mar. 2025. [Online]. Available: https://www.numberanalytics.com/blog/davies-bouldin-index-5essential-metrics

[7] ”Decision Tree - Classification,” SAEDsayad. [Online]. Available: https://www.saedsayad.com/decision tree.htm

[8] ”What is Root Mean Square Error (RMSE),” Giskard. [Online]. Available: https://www.giskard.ai/glossary/root-meansquare-error-rmse

[9] T. O. Hodson, ”Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not,” Geosci. Model Dev., vol. 15, pp. 5481-5487, Jul. 2022.