MODELING AND FORECASTING GENRE POPULARITY IN SPOTIFY MUSIC THROUGH AUDIO FEATURE ANALYSIS USING XGBOOST AND ARIMA TIME SERIES MODELING

A Thesis Proposal

Presented to the Faculty of the

Department of Computer, Information Sciences and Mathematics

University of San Carlos

In Partial Fulfillment

of the Requirements for the Degree

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

By

JONAZ JUAN II C. SAYSON

FABIOLA C. VILLANUEVA

CHRISTINE D. BANDALAN, M.Eng

Faculty Adviser

May 2025

ABSTRACT

Forecasting music genre popularity is increasingly important in the age of digital streaming, where data-driven strategies influence both content creation and distribution. This study explores how audio characteristics from Spotify tracks can be used to predict genre popularity trends and identify which musical features contribute most to listener engagement. Using a dataset of approximately 650 high-popularity songs released between 2008 and 2024, audio features were aggregated by genre and year to model broad genre trends. Two models were implemented: ARIMA for forecasting temporal changes in genre popularity, and XGBoost for analyzing the predictive influence of audio features such as energy, tempo, and danceability. The models were validated using walk-forward and time-aware testing strategies to ensure reliability and reduce bias. Results indicate that genre popularity is forecastable within a one- to five-year horizon, and that specific audio features consistently drive popularity across genres. The system successfully predicts future trends and reveals genre-specific characteristics influencing popularity. Aligning creative output with evolving audience preferences, valuable insights are offered by these findings for artists, producers, and music analysts navigating the landscape of the music industry by providing a practical framework for understanding the intersection of musical structure and popularity in a streaming-driven music landscape.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

The study's introduction outlines its purpose, significance, and scope, presenting key concepts and objectives that underpin the research. Establishing the foundation for a predictive system to forecast music genre popularity using Spotify audio feature analysis, leveraging time-series modeling and machine learning techniques for robust insights.

1.1 Rationale of the Study

Driven by the rise of online streaming platforms, the digitalization of music has fundamentally reshaped the consumption and distribution of popular music, with Spotify emerging as a dominant platform (Joshi et al., 2025). According to Gupta (2025, citing Statista Reports), 85% of music listeners rely on Spotify to explore a wide array of music genres, with the platform's user base growing by 14% annually — encompassing both established artists and emerging talents. Spotify's widespread adoption underscores its critical role in listener engagement, music discovery, and the overall dynamics of the music industry. Su & Zhou (2022) demonstrate that audio features — such as tempo, pitch, and timbre — elicit specific emotional responses across diverse genres, highlighting the dynamic interplay between musical composition and listener perception.

As creators, artists deliberately manipulate these audio features to craft balanced artistic expressions with technical precision to evoke emotional experiences (Walfish, 2024). However, inflationary pressures and increasing commercialization have diminished the value of creative work, posing significant challenges for artists navigating a competitive landscape where they are also expected to operate as cultural entrepreneurs (Everts et al., 2021).

As the music industry increasingly demands a professional, entrepreneurial mindset, achieving a sustainable full-time career now requires not only artistic talent but also a deep understanding of what drives commercial success (Barbara, 2023; Pant & Shrestha, 2023). Many artists invest years into honing their craft yet struggle to gain recognition due to limited access to actionable market insights. This research aims to bridge the gap between artistic intuition and market dynamics by providing forecast-driven and feature-informed insights to empower more strategic, creative, and commercially viable decision-making. By leveraging Spotify's rich dataset of audio features and listener trends, this study employs time series analysis to forecast the future popularity of music genres based on historical listener behavior. Additionally, machine learning analysis will be applied to determine which audio features most significantly influence genre popularity. Together, these models offer artists and industry stakeholders data-driven insights to better anticipate shifts in listener preferences and optimize their creative and strategic decisions.

1.2 Statement of the Problem

1.2.1 General Objective

This study aims to develop a system that models and forecasts the popularity of music genres using audio features from Spotify, applying XGBoost for feature impact analysis and ARIMA for time series forecasting.

1.2.2 Specific Objectives

In order to achieve the general objective of modeling genre popularity and identifying influential audio features, the study seeks to answer the following questions.

1. What audio features can be extracted and analyzed from Spotify's historical dataset to model how the popularity of music genres have changed over time?
2. What audio features most significantly influence genre popularity and what are their relative importance in shaping trends?.
3. How can the forecasted outputs be interpreted to provide actionable insights for music producers, songwriters, and artists, enabling them to adapt their creative and technical decisions based on forecasted evolving music directions?

Conducting the study shall address these said objectives and its findings shall provide meaningful insights that will be able to empower and support industry professionals to make informed, strategic decisions in such a fluctuating industry.

1.3 Significance of the Study

By providing forecasting insights into genre trends and analytical methods, this study aims to gain actionable tools to navigate the industry, optimize strategic decisions, and enhance both creative and commercial outcomes. Specifically, this study will benefit the following groups, ordered from general to most directly impacted:

Music Industry Stakeholders. This study shall deliver predictive insights into genre popularity trends, enabling stakeholders to anticipate shifts–such as emerging or declining genres–in listener preferences and adapt strategies accordingly. By identifying which genres are likely to gain traction, it will be able to inform marketing, production and creative investment decisions.

Music Consumers and Listeners. Though listeners indirectly benefit from a more dynamic and responsive music industry, the study's forecasting model will be able to support playlist curation and recommendation algorithms, improving its relevance of music suggested to users on music streaming platforms such as Spotify.

Future Researchers and Academics. Conducting the study will provide a methodological framework for analyzing music consumption patterns, enabling future researchers to develop enhanced models for predicting or forecasting musical trends. Moreover, its findings will be able to contribute to musicology, data science, and cultural studies, which shall foster further research into digital music platforms and listener behavior.

Technology and Data Science Communities. The application and implementation of ARIMA and XGBoost into music data will demonstrate the versatility of time series forecasting and machine learning modeling into cultural domains–such as adapting the study's methodology into film, fashion, etc.–fostering innovation in forecasting analytics and broadening the impact of data-driven approaches.

Educational Institutions and Music Schools. Music programs can incorporate the study's findings into curricula, teaching how audio features influence genre popularity and equipping students with data-driven approaches to modern music creation, industry navigation, and creative demands.

Music Streaming Platforms. This study utilizes a dataset derived directly from Spotify's catalog of popular and non-popular songs. Based on the outputs of genre popularity forecasts and feature influence analysis, music streaming platforms like Spotify could refine their recommendation algorithms and playlist curation strategies to better align with evolving listener preferences, thereby enhancing user engagement and retention. Other platforms such as Apple Music and YouTube Music could similarly adopt these insights to optimize their own personalization systems.

Music Producers and Songwriters. By leveraging insights into audio features that are taken into account of the study, producers and songwriters can drive through popular genres to commercially viable tracks, balancing creativity with market appeal, optimizing their tracks for streaming success.

Independent Musicians and Emerging Artists. Independent artists, often lacking major label support and resources, can leverage the study's genre forecasts to strategically craft music to align with upcoming and emerging trends, enhancing their visibility and market impact.

1.4 Scope and Limitation

The research study focuses on forecasting the future of popular music genres using Spotify's Music Dataset spanning from years 1970 to 2024, that have extracted both popular songs and non-popular songs. The dataset's audio and descriptive features–such as energy, tempo, danceability, loudness, liveness, valence, instrumental-ness, and acoustic-ness–will then be analyzed in the study. An aggregated dataset will be derived which shall identify the most prevalent genres per year, this is determined by which genres had the highest number of songs in Spotify’s top charts annually. The scope of the study will particularly focus on four genres–namely Pop, Rock, Hip-Hop, and Latin–having the largest volume of tracks and consistent data spanning 16 years from 2008 to 2024. Additionally, data before Spotify’s launch in 2008 heavily relies on retrospective cataloging, limiting the reliability of pre-2008 genre trends due to incomplete or inconsistent song metadata.

Time series analysis shall then be conducted using the Autoregressive Integrated Moving Average (ARIMA) model for forecasting genre popularity, and XGBoost will be used to analyze and determine the influence of audio features on popularity. The study will forecast genre trends over a five-year horizon (2025–2029) for genres with sufficient historical data. For genres with fewer data points, the forecast range will be limited at one to two years to maintain model reliability. The accuracy of these forecasts will be evaluated by comparing the predicted values to withheld recent data (e.g., 2022–2023), using quantitative error metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

This research study focuses on forecasting the popularity of music genres, based on Spotify’s dataset from 1970 to 2024, which may not fully capture music consumption patterns across other platforms, such as Apple Music, YouTube Music, or non-streaming sources. The aggregation of one genre per year simplifies genre diversity and may potentially overlook subgenres or hybrid genres that influence broader musical trends. Moreover, the study’s dependence on Spotify’s predefined audio features limits the analysis to quantitative attributes—such as energy, tempo, and valence—without accounting for qualitative factors like lyrical content, cultural context, or artistic branding. The study’s global approach, based on Spotify's popular and non-popular songs, may not reflect regional variations in genre popularity.

CHAPTER 2
REVIEW OF RELATED LITERATURE

By compiling methodologies, techniques, and insights from existing works, this review synthesizes foundational studies and scholarly research relevant to the present study, focusing on the digital transformation of music, user behavior, algorithmic influences, time series analysis and forecast modeling which this thesis addresses.

2.1 Digitalization of Music

According to Editor (2023), the digitalization of music traces back with the introduction of the MP3 format in the late 1990s, which has enabled efficient storage and sharing of audio files, shifting from physical media to digital downloads. Platforms such as Napster, launched in 1999, disrupted traditional music distribution by facilitating a peer-to-peer file sharing service and has revolutionized how music was shared and accessed online, paving the way for the online music economy (Brohan, 2025). Legal digital distribution was then further formalized in 2003 by Apple's iTunes Store by allowing single-track purchases, changing consumer purchasing behavior (Sheridan, 2025). Chhabria (2025) states that by the 2010s, Spotify has become the dominant streaming service as its mode of music consumption offers subscription-based access to vast music libraries, transforming how music is distributed and monetized. Listener data and streaming metrics have become central to industry analysis and decision-making through the ecosystem that the music digitalization has created.

Brusila et al. (2021) states that while digitalization has democratized music access and production, the outcome for this evolution is neither wholly democratic nor undemocratic but rather manifold and context-bound. This has made the relationship between digital technologies and music diverse that has produced multiple outcomes, enabling global music access. Characterized by the transition from analog formats, such as vinyl and cassette tapes, to digital formats like MP3s and streaming files, the digitalization of music has fundamentally reshaped the music industry’s production, distribution, and consumption paradigms (Guo, 2023). Driven by technological advancements in internet infrastructure, cloud computing, and audio compression, this evolution has greatly changed the music business and its access to music (Dolota, 2020). The digital transformation has laid the groundwork for the music streaming platforms, which has shifted from ownership to access-driven models, simultaneously redefining its music access and data collection. The rise of streaming platforms further amplifies these changes, shaping listener behavior and industry dynamics.

2.2 Music Streaming Platforms

Music streaming platforms, including Spotify, Apple Music, Amazon Music, and YouTube Music, have transformed music consumption by offering on-demand access to extensive catalogs via subscription models (Pedersen, 2020). Leveraging music digitalization, these platforms emphasize accessibility, personalization, and data-driven engagement, collecting user interactions to fuel recommendation algorithms (Mokoena & Obagbuwa, 2025). The landscape for the music industry has been strongly shaped and influenced by dominating music streaming platforms.

In a comprehensive analysis on application revenue, streaming statistics, and market trends, Duarte (2025), Curry (2025), and Shaikh (2025) report that music streaming has accounted for approximately 84% of the global music industry revenue and generated around $17.5 billion in revenue as of 2024-2025. Demonstrating consistent growth, 600 million people have subscribed to music streaming platforms globally by expanding over 10% in the past year alone. Among the many services competing in this space, Spotify dominates the market by holding 31.7% share of the global music streaming market. Spotify’s user base grows 600 million monthly active users, with 236 million paid subscribers, making it the largest streaming platform by a significant margin.

See Table 1 as the summary of the market share of leading music streaming platforms as of 2025 and their notable pad subscribers (Duarte, 2025).

Table 1. *Market Share of Leading Music Streaming Platforms (2025)*

Streaming platform	Market share (%)	Paid subscribers (millions)
Spotify	31.7	263
Tencent Music	14.4	121
Apple Music	12.6	93
Amazon	11.1	~80
YouTube Music	9.7	~100
NetEase	6.7	Not specified
Yandex	3.4	Not specified
Deezer	1.3	Not specified
Others	9.7	Not specified

Its dominance is driven by its freemium model, extensive catalog, and AI-driven personalization like curated playlists and Spotify Wrapped, supporting robust data for genre trend forecasting (Khan, 2021; Spotify, 2025). Other platforms–such as Tencent Music (14.4% market share, China-focused), Apple

Music (12.6%, device-integrated), Amazon Music (11.1%), and YouTube Music (9.7%, video ecosystem)--have substantial user bases and regional strengths, Spotify's global scale of reach solidifies its market leadership. With Spotify's unparalleled scale driving trend insights generates a data-rich that captures user behavior like skipping and playlist curation.

2.3 User Behavior and Audio Trends

The rise of the digital music streaming industry and their platforms has not only transformed music access but also reshaped how users interact with audio content, with personalization, convenience, and social dynamics that are now central to the listening experiences. An analysis conducted by Mokoena & Obagbuwa (2025) highlights that algorithmic recommendations on platforms like Spotify and Apple Music heavily influences listener behavior, done by analyzing listening habits, skip rates, and playlist curation to tailor content, providing insights into music popularity trends. For instance, leveraging machine learning to suggest tracks, Spotify's Discover Weekly and Release Radar playlists influence 60% of users to explore new genres weekly, amplifying genre discovery (Khan, 2021). User satisfaction and genre affinity, however, are influenced by a user's skipping behavior. According to Meggetto et al. (2023), user behavior features–such as skip rates and listening patterns–are the most discriminative factors for predicting music skips, with high skip rates often indicating a mismatch between recommended tracks and user preferences; making content and contextual features play a lesser role. As frequent skipping of certain genres may indicate declining popularity, while listening to others is sustained reflects enduring appeal making this behavior particularly relevant for genre popularity analysis. Moreover on how user behavior shapes genre trends, Spotify (2025) on their 2024 Wrapped report highlights pop and hip-hop dominating global listening, 25% and 20% of streams, respectively.

Genre popularity is further influenced by social dynamics and virality, distinguished by a research conducted by Oliveira et al. (2024) finding that virality and success on streaming platforms are fundamentally distinct phenomena though they are closely related: stating that many viral songs do not become long-term hits, and not all hit songs experience viral moments. The research finds that the distinction between the two lies in their diffusion processes, hit songs generally enjoy longer-lasting popularity while viral songs tend to have shorter and more explosive chart trajectories. For instance, K-pop's viral surges on Spotify are fueled by fan-driven campaigns, in contrast to the steady popularity of lo-fi hip-hop as a favored background music. With playlist curation and shared listening sessions fostering community engagement, patterns as such highlight the role of user-driven amplification in genre trends.

Duarte (2025) states a reinforcement on genre preference through social validation wherein approximately 70% of users share their Spotify Wrapped summaries. Highlighting the complex and multifaceted nature of music popularity in the digital age, the relationship between viral trends that are often amplified by social media does not hold universally though it can forecast streaming success for some songs. Beyond music itself, the evolving consumption patterns have been influenced by the rise of non-music audio, such as podcasts. Indicating users' growing preference for diverse audio experiences, Spotify reports a 20% increase in podcast streaming in 2024 (Curry, 2025). While music remains the primary focus though, user behavior data offers a rich foundation for analyzing genre popularity dynamics in the digital age, paving the way for understanding the psychological and social interplay between listeners and music that shapes consumption patterns.

2.4 The Dynamic Interplay Between Music and Listeners

The reciprocal relationship between music and listeners drives real-time engagement, shaping emotional responses, preferences, and social interactions beyond the aggregate consumption patterns explored in user behavior trends. This discussion shall examine how listeners and music co-adapt during the listening experience, offering insights critical for forecasting genre popularity and refining recommendation systems. By exploring psychological, social, and methodological dimensions of this interplay, providing context for modeling listener-driven genre trends using Spotify data.

Music profoundly influences listener emotions and preferences, creating dynamic engagement. Bhattacharya et al. (2023) introduces the Dynamic Music Engagement Model which conceptualizes music engagement as a context-sensitive and evolving process that supports mood regulation and mental health, advocating for context-aware music systems that adapt to social settings, enhancing listener experiences and generating behavioral data. To analyze how music engagement supports mental health, mood regulation, and community interaction, the literature's methodology combines large-scale surveys with computational modeling, revealing how audio features like valence and tempo evoke emotional responses, driving genre preferences. Further supporting participatory features such as live-streamed concerts and social listening platforms, this approach emphasizes the need for music information systems to be context-aware and adaptive. This interplay is further enriched by social and contextual factors, such that 40% of users participating in shared listening sessions are from Spotify's collaborative playlists and live streamed concerts that ultimately fosters community, amplifying genre popularity. (Duarte, 2025).

Methodologically, this field was also researched by Schiavio et al. (2021), offering frameworks to capture the interplay's complexity. The literature models musical engagement as an adaptive process using dynamical systems theory. Wherein its nonlinear time-series analysis employs techniques such as recurrence plots that make it identify stable listener interaction patterns that reflect genre preference shifts. This listener-driven interplay shapes consumption patterns, influencing the strategies of producers and artists, who adapt to platform demands and genre trends in the competitive digital landscape of music.

2.5 Producers, Artists and the Competitive Landscape of Music

Streaming platforms such as Spotify have fundamentally transformed music production by placing data-driven decision-making at the forefront of the creative process. To monitor real-time streaming metrics–such as play counts, skip rates, and listener demographics–producers and artists now routinely utilize analytic tools like Spotify for Artists (Sajdak, 2024; Slomka & Sajdak, 2024). Creators are prompted to tailor their work to genres and audio features that align with high-engagement trends–such pop, hip-hop, or tracks with elevated energy and danceability–informing insightful production choices. Production strategies increasingly reflect the streaming data that underpins genre popularity forecasting models with many producers optimizing their output for algorithmic playlist inclusion and sustained listener engagement as a result.

A producer or an artist positioning themselves in today's music industry extends well beyond the act of releasing music; multifaceted strategies encompassing visibility, algorithmic recognition, and audience engagement within an increasingly competitive environment has become a necessity. Streaming platforms have democratized access to audiences but have simultaneously intensified competition, particularly for emerging artists. According to Everts et al. (2021), early-career musicians must navigate a complex ecosystem where creative output has become one component of success. Wherein engagement in continuous self-promotion, digital marketing, and audience cultivation has become a must to be able sustain their careers. With emerging artists relying heavily on streaming channels for exposure, algorithmic playlists and user-generated content curation dominates streaming consumption. Pant & Shrestha (2023) underscored the ability to predict and influence commercial success increasingly hinges on understanding and leveraging streaming data, playlist placements, and listener engagement metrics. Consequently, to maximize algorithmic favorability and user retention, artists must optimize metadata, pitch strategically for playlists, and deploy marketing tactics.

Requiring consistent content output and active fan engagement across multiple platforms, independent and emerging artists face intense competition in today's algorithmic-driven music industry (Walfish, 2024). New artists often face difficulty in gaining visibility without investing in social media campaigns, influencer partnerships, and playlist pitching as Spotify's algorithm favors familiarity and favors repeat listening. Tailoring their music to algorithmic preferences by producing shorter tracks with early hooks to reduce skip rates and improve playlist compatibility (Barbara, 2023). Despite the vibrancy and creative potential of the local music scene, artists and industry stakeholders in the Philippines continue to face significant structural and economic barriers–such as limited representation in policy making, inconsistent income, and challenges related to digital disruption and global competition (Logronio, 2024; Gutierrez, 2024). Due to local infrastructure and economic constraints, Filipino artists encounter additional hurdles, compelled to leverage social media virality and niche genre targeting for exposure. Continuous adaptation to evolving platform algorithms and audience behaviors is required to sustain a career and the rise of AI-generated music and revenue consolidation further complicates the landscape, challenging artists to maintain authentic connections with listeners. Success now demands a sophisticated balance of artistic integrity, data-driven marketing, and persistent audience engagement to thrive in the competitive global music market.

2.6 The Significance of Audio Features in Music

Central to modeling and forecasting music genre popularity on platforms like Spotify is through understanding the significance of audio features. Nijkamp (2020) investigates the relationship between Spotify's audio features–such as key, tempo, and energy–and song popularity measured by stream counts. The study highlights the complexity of predicting success in cultural markets like music and underscores the importance of incorporating a broad range of attributes into predictive models, though the study finds only explanatory power of these features. Providing aligned insights into identifying which audio features most significantly influence genre popularity and their relative importance into shaping trends. Furthermore, recent advances in feature extraction enhance genre classification and popularity forecasting. Zhang, J. (2021) uses convolutional neural networks with gating and attention mechanisms to automate hierarchical feature extraction, improving classification accuracy; having the approach reduce human bias in feature selection, proposing a deep learning framework that converts raw audio into spectrograms. A research reviewed by Jitendra and Radhika (2020) on classical techniques like time-domain (e.g., zero-crossing rate), frequency-domain (e.g., spectral centroid), and perceptual features (e.g., MFCCs), emphasizing their role in reducing dimensionality and enhancing model performance for genre classification.

Castellon et al. (2021) capturing temporal and spectral patterns for music information retrieval tasks by introducing an audio language modeling framework that learns compact representations through unsupervised training, enabling robust feature representations for temporal trends. A research conducted by Saragih (2023) analyzes nearly 27,000 songs, correlating features like danceability and acousticness with Billboard chart success, reinforcing feature importance and guiding creators toward high-energy tracks for playlist-driven markets. These studies underscore the practical impact of your thesis in forecasting genre popularity. Building on these feature-driven insights, the next section explores music genres and trends shaping Spotify's streaming landscape. Leveraging these feature-driven insights, the following section examines music genres and trends shaping Spotify's streaming landscape.

2.7 Music Genres and Trends

Understanding music genre classification and trends is pivotal for forecasting popularity on Spotify, Shariat and Zhang (2023) demonstrated that combining feature selection with ensemble learning, such as XGBoost, enhances genre classification accuracy by prioritizing relevant audio features like danceability and valence, ensuring robust predictive performance by focusing on Spotify's data-driven metrics. Advancements in deep learning further captures complex audio patterns as Zhang (2022) employs convolutional neural networks (CNNs) to extract hierarchical features from spectrograms, achieving high accuracy across diverse genres, leveraging Spotify's comprehensive audio features to model nuanced genre characteristics. Similarly, Tang et al. (2023) validate the robustness of deep learning in tracking genre dynamics over time, providing reliable feature representations for temporal trends.

Large-scale streaming data analyses offer insights into genre trends and listener behavior, Jiang et al. (2024) analyzed streaming patterns, revealing how genre-based consumption reflects cultural and market shifts, capturing temporal fluctuations in genre prevalence. Zhuang et al. (2020) conducts a study wherein transformer-based classifiers that model long-range audio dependencies, enable finer-grained genre discrimination critical for accurate popularity modeling.

Comparative studies further strengthen your methodological choices. Qi et al. (2022) evaluate machine learning models, highlighting trade-offs between interpretability and accuracy in genre classification. Murauer and Specht (2018) confirm XGBoost's effectiveness in handling high-dimensional audio datasets, which is highly relevant to this thesis study's feature importance analysis. Collectively, these studies underscore the power of quantified audio features and advanced machine learning to capture genre trends, offering actionable insights for producers and platforms to adapt to evolving musical tastes. These technical foundations raise ethical questions, which the following section on ethical considerations in music analytics explores.

2.7.1 Extreme Gradient Boosting Model

Chen and Guestrin's (2016) seminal work on XGBoost introduced a scalable, efficient, and highly accurate tree boosting system that has since become a leading method for structured data modeling across various domains, including music information retrieval and genre classification. XGBoost's core strengths–ability to handle nonlinear relationships, manage missing data, and rank feature importance–have made it particularly valuable for music genre classification tasks, as demonstrated in recent studies (Gan et al., 2024; Murauer & Specht, 2018). Notably, XGBoost's feature discovery capability allows researched literature to identify which audio features (e.g., tempo, acousticness, duration) are most influential in distinguishing between genres, offering actionable insights for both academic research and industry applications.

Providing a robust framework for both classification and feature impact analysis, XGBoost proves its effectiveness in music genre classification and its ability to rank the importance of audio features significantly influencing genre trends and popularity. Furthermore, XGBoost's scalability and interpretability ensure efficient computation and accessibility to music producers, songwriters, and industry stakeholders seeking to adapt to evolving musical landscapes.

2.8 Algorithmic Influences

Spotify has profoundly shaped music consumption and production patterns by increasing its dominance as an algorithmically driven recommendation system on music streaming platforms. Hesmondhalgh et al. (2023) highlights how these personalized algorithms mediate music discovery by analyzing user behaviors, though their opaque nature leaves both listeners and creators uncertain about the criteria that govern content selection and promotion. Algorithmic bias is attributed to this transparency often at the expense of independent or niche creators, where popular artists and mainstream genres disproportionately benefit from greater exposure (Born et al., 2024). Hence, creative innovation within the industry is potentially constrained by such biases, narrowing the diversity of music consumed and produced. Conversely, demonstrated by Hansen et al. (2021), recommendation algorithms can be designed to promote diversity by introducing users to music outside of their established music preferences, elevating less popular tracks. This shift toward diversity not only enhances user satisfaction but also correlates with improved long-term business metrics such as user conversion and retention. However, despite these potential benefits, the inherent biases within Spotify's hybrid recommender system-which combines collaborative and content-based filtering-persist as a limitation of the streaming data used in the literature. Observed genre popularity trends may partly result from the platform's recommendation dynamics rather than purely organic listener preferences by data reflected on these algorithmic influences. Critically interpreting the dataset and deriving insights from it should include the recognition of these limitations.

Data may overemphasize the popularity of already dominant genres by the presence of algorithmic bias, implying that some genres or artists may be systematically underrepresented. This necessitates a modeling approach capable of capturing both the temporal evolution of genre popularity and the complex, potentially nonlinear relationships between audio features and streaming success, while acknowledging the filtered nature of the data. The dual objectives-to forecast temporal trends in genre popularity and to identify the most influential audio features shaping these trends-require methods that can handle time-dependent patterns and complex feature interactions within biased and incomplete data. Providing a robust framework, informed by these considerations, for extracting meaningful insights despite the inherent constraints imposed by algorithmic mediation in the dataset, the selection of models address these said challenges.

2.9 Methods of Prediction and Forecasting

Predicting music genre popularity requires an understanding of both temporal patterns and the audio-based features that influence user engagement. To address these dual aspects, this study adopts a hybrid forecasting methodology that includes time-series analysis for trend projection and supervised machine learning for feature importance evaluation. Interiano et al. (2018) conducted one of the most comprehensive studies to date on the predictability of success in contemporary music. Their research showed that certain audio features such as danceability, energy, and valence consistently correlate with the likelihood of a song becoming a hit. The study employed various statistical learning models and demonstrated that forecasting success in music is a feasible and quantifiable task when historical streaming data and feature-rich datasets are used. Drawing from their methodological precedent and empirical error ranges, this research integrates similar metrics—such as MAE and RMSE—to evaluate forecasting performance, using thresholds inspired by their findings.

To meet the study's dual objectives—(1) forecasting genre popularity over time, and (2) identifying which audio features most influence that popularity—two distinct yet complementary models were selected. ARIMA (AutoRegressive Integrated Moving Average) was chosen for its capability to model temporal dependencies in univariate time series data. Meanwhile, XGBoost (Extreme Gradient Boosting) was selected for its proven effectiveness in handling nonlinear relationships in high-dimensional, structured datasets, especially in ranking feature importance. These modeling choices are not arbitrary; they are informed by the nature of Spotify's data and grounded in well-established literature on music informatics and predictive modeling. By combining these approaches, this study not only aims to predict future genre trends but also to provide interpretability through ranked audio features. The selected models are suited to both retrospective trend analysis and forward-looking predictions, aligning with the research goal of offering actionable insights to artists, producers, and industry stakeholders.

2.10 The Autoregressive Integrated Moving Average Statistical Model

The ARIMA model is a classical statistical approach for time series forecasting, designed to capture trends, cycles, and seasonality within univariate datasets. It combines three components: autoregression (AR), integration (I), and moving average (MA), making it flexible for modeling various types of time-dependent patterns. In the context of this study, ARIMA is used to analyze historical trends in genre popularity, where the primary variable of interest is the median popularity score of genres across years. According to Hyndman and Athanasopoulos (2021), ARIMA is particularly effective for time series with consistent patterns and relatively fewer data points—common in annualized or genre-level aggregates. The model is trained using walk-forward validation, simulating real-world forecasting scenarios where new data is continuously appended, and forecasts are generated for future time steps. The model's performance is assessed using MAE and RMSE, with acceptable thresholds defined based on Spotify's 0–100 popularity metric and empirical benchmarks from prior research. In this study, ARIMA helps address the temporal forecasting objective by providing interpretable, seasonally-aware projections of genre popularity, serving as a baseline and benchmark against which more complex models may be compared.

2.11 Data Validation, Visualizations and Ethical Considerations in Music Analytics

Robust validation and ethical scrutiny are vital to ensuring the reliability and relevance of forecasting models in music analytics. Following the experimental and build methodologies outlined by Zelkowitz and Wallace (1998), this study emphasizes not only statistical performance but also the interpretability and reproducibility of results. For the ARIMA and XGBoost models, validation involves multiple steps. Quantitative validation uses walk-forward testing for ARIMA and time-aware train-test splits for XGBoost. MAE, RMSE, and R^2 are used as core evaluation metrics. Acceptable performance thresholds are drawn from Interiano et al. (2018), with RMSE ≤ 8 and MAE ≤ 6 for ARIMA, and MAE ≤ 5 and $R^2 \geq 0.7$ for XGBoost. If models do not meet these criteria, they are retrained using alternative hyperparameters or revised feature sets. Additionally, a final holdout set (the most recent 10% of data) is used to test real-world generalization.

Visual inspection complements quantitative methods, with forecast plots and feature importance charts reviewed to ensure logical alignment with known trends (e.g., hip-hop showing high energy and tempo). The use of open-source libraries—such as statsmodels (Seabold & Perktold, 2010) for ARIMA and XGBoost (Chen & Guestrin, 2016) for tree-based modeling—ensures reproducibility and transparency. Ethical considerations also inform this study's limitations. Recommendation systems and user-generated data may reflect algorithmic bias, platform effects, and social amplification rather than pure musical preference. This critical lens acknowledges that genre popularity trends may be shaped as much by algorithmic visibility as by organic listener behavior, prompting a cautious interpretation of model outputs. The inclusion of qualitative validation (e.g., expert judgment, alignment with known music trends) further enhances the robustness of the forecasting framework. Together, these validation protocols and ethical safeguards aim to ensure the study's conclusions are both statistically sound and contextually grounded.

CHAPTER 3
TECHNICAL BACKGROUND

This chapter contains the definition of technical terms used in prior chapters or to be used in the upcoming chapters. The said terminologies shall range from terminologies in the field of Computer Science and the field of Music.

3.1 Music Streaming Platforms

Digital services that allow users to access, play, and interact with vast libraries of music content over the internet. The platforms deliver audio content via real-time data streaming, often incorporating adaptive bitrate streaming, compression, and caching technologies to optimize playback quality. They typically offer personalized recommendations based on user behaviour and preferences, and support various models such as subscription-based or ad-supported access.

3.2 Spotify Web API

An interface provided by Spotify that allows developers to access metadata, track information, user playlists, and audio features of songs. The dataset was extracted from Spotify through Spotify Web API for trend analysis and music popularity prediction.

3.3 Spotify Music Dataset

Is a dataset collected from Spotifys' API using two python scripts to extract popular and non-popular songs and their associated audio and descriptive features. Descriptive features of a song include information about the song such as the artist ame, album name and release date. Audio Features include key, valence, danceability and energy which are results of spotify’s audio analysis.

3.4 Feature Importance

A technique used in machine learning to determine which input variables (features) have the greatest influence on a model’s predictions. In music analytics, feature importance analysis helps identify which audio attributes—such as danceability, tempo, or energy—are the most critical in predicting song popularity.

3.5 Time Series Analysis

Is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time.

3.6 Autoregressive Integrated Moving Average (ARIMA)

A statistical model used for time series forecasting. It combines autoregression (AR), differencing (I) to remove trends, and moving average (MA) components to capture temporal structure in data. ARIMA is applied in this study to predict future changes in genre popularity.

3.7 Extreme Gradient Boosting (XGBoost)

A powerful, scalable, and efficient gradient-boosting machine learning algorithm. It builds an ensemble of decision trees and uses regularization to improve model generalization. In this thesis, XGBoost is used to determine which audio features most strongly influence genre popularity.

3.8 Machine Learning (ML)

A subset of artificial intelligence (AI) that enables systems to learn and make predictions from data without being explicitly programmed. Machine learning techniques are widely used in music trend analysis to predict song popularity based on historical data and audio features.

3.9 Supervised Learning

A machine learning approach where models are trained on labeled data, meaning the input data is associated with known output values. In the context of music prediction, supervised learning algorithms can be used to classify songs as "popular" or "unpopular" based on their audio features.

3.10 Time Series Analysis

A method of analyzing data points collected or recorded at specific time intervals. It is used to detect trends, seasonal patterns, and structural changes. In this thesis, it supports forecasting genre popularity over multiple years.

3.11 Stationarity

A statistical property of a time series where its mean, variance, and autocorrelation structure do not change over time. Stationarity is a prerequisite for ARIMA modeling and is tested using methods like the Augmented Dickey-Fuller (ADF) test.

3.12 Walk-Forward Validation

A method of validating time series models by training on an expanding window of past data and testing on subsequent time points. It mirrors real-world forecasting conditions and is used for validating ARIMA and XGBoost in this study.

3.13 Mean Absolute Error (MAE)

A metric for evaluating model accuracy. It calculates the average absolute difference between predicted and actual values. A lower MAE indicates higher model precision.

3.14 Root Mean Squared Error (RMSE)

A commonly used evaluation metric that measures the square root of the average squared differences between predicted and actual values. RMSE penalizes larger errors more heavily than MAE.

3.15 R-Squared (R^2)

A statistical measure of how well the model's predictions approximate actual outcomes. An R^2 value closer to 1.0 indicates that the model explains most of the variance in the dependent variable.

CHAPTER 4
DESIGN AND METHODOLOGY

The study's methodology includes the conceptual framework, analysis and design, development model, schedule for system design, development, and testing, and the roles and tasks assigned to the research team.

4.1 Conceptual Framework

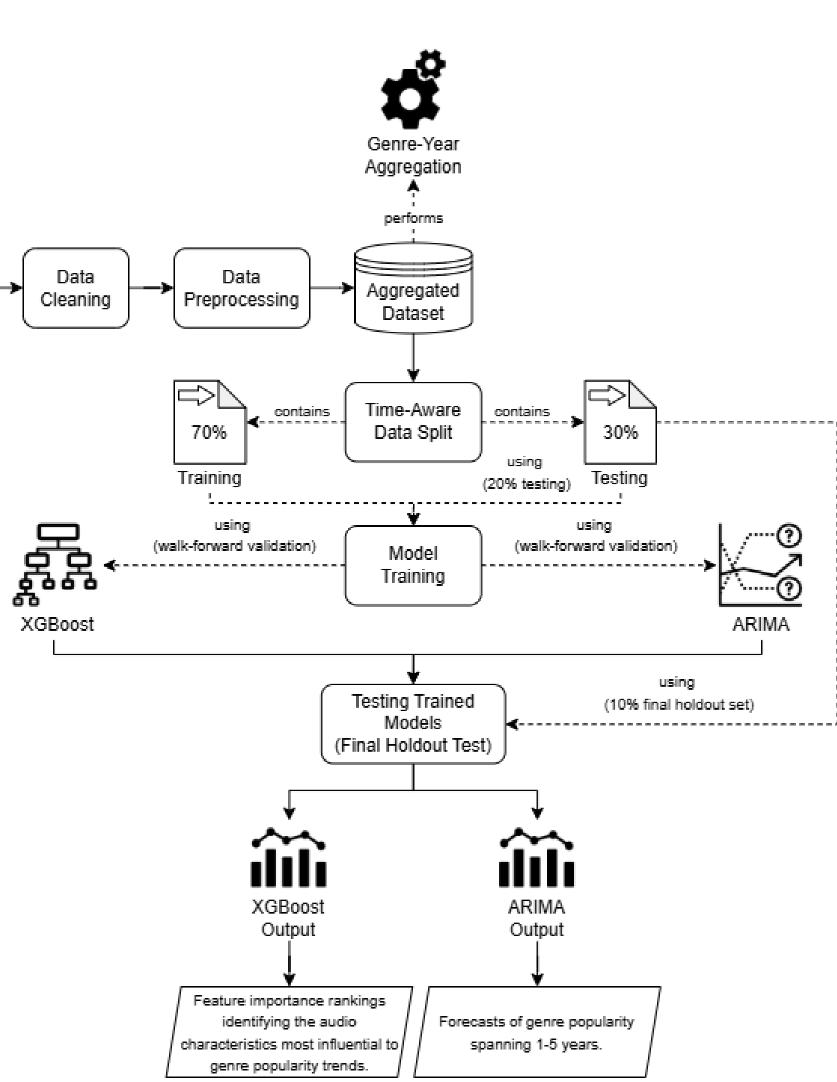


Figure 1. Conceptual Framework

As shown in Figure 1, the diagram illustrates the proposed modeling and forecasting system. The dataset—comprising audio characteristics and descriptive attributes of approximately 650 high-popularity songs (popularity ≥ 68)—will be sourced from Kaggle, representing Spotify tracks across genres selected based on consistent representation from 2008 to 2024 (typically 6–13 usable years). Only songs from 2008 onward will be included to ensure consistency with Spotify's operational timeline and to avoid sparse or unreliable data prior to that period.

The extracted raw dataset will first undergo data cleaning, which includes the removal of null values and duplicates to ensure integrity. Following this, the data will pass through a preprocessing pipeline that incorporates a structured Exploratory Data Analysis (EDA) phase.

The exploratory data analysis (EDA) phase was conducted to examine the structure, integrity, and variability of the dataset. This step helped identify potential outliers and confirmed the presence of consistent patterns across key features. Based on both statistical representation and cultural relevance, the study focuses on four primary genres: Pop, Hip-Hop, Latin, and Electronic. These genres were selected not only for their sufficient presence in the dataset but also for their significant influence on global music consumption. Pop and Hip-Hop dominate mainstream streaming platforms (IFPI, 2023; Morris & Powers, 2015), Latin has gained substantial international traction through globally successful artists (Mora et al., 2021; Nielsen Music, 2023), and Electronic music is integral to digitally curated and algorithm-driven listening environments (Fikentscher, 2000; Prey, 2018).

Following refinement, the dataset will be aggregated by genre and release year, computing the median values of selected audio features—including energy, tempo, danceability, loudness, liveness, valence, speechiness, instrumentalness, and acousticness—as well as the median popularity for each genre-year pair. These features are selected for their established relevance in genre differentiation and listener engagement (Interiano et al., 2018).

The resulting dataset will support two primary modeling strategies:

1. Time Series Forecasting Using ARIMA

ARIMA models will be used to forecast historical trends in genre popularity for Pop, Hip-Hop, Latin, and Electronic, based on the aggregated dataset. The modeling begins with an initial manual tuning phase, where values for (p, d, q) are estimated through stationarity testing (Augmented Dickey-Fuller test), and inspection of autocorrelation (ACF) and partial autocorrelation (PACF) plots.

After manual tuning, automated hyperparameter optimization will be applied using a combination of random search, grid search, or Bayesian optimization—depending on the stability of the time series for each genre. This hybrid tuning approach allows for interpretability in early parameter selection, while maximizing predictive performance during refinement (ACM, 2024; MDPI, 2024).

To simulate real-world forecasting and ensure robustness, ARIMA will be validated using walk-forward validation with the following setup:

- In-sample window: 3 years (used for model training). This window is large enough to capture short-term genre trends while preventing overfitting, consistent with recommendations for non-seasonal, annual-level data (Hyndman & Athanasopoulos, 2021).
- Out-of-sample window: 2 years (used for testing). A 2-year horizon balances practical forecast length with the need to test generalization. Prior research shows that forecasting accuracy tends to decline sharply beyond this range in cultural data (Interiano et al., 2018).
- Step size: 1 year. Each validation window rolls forward by 1 year, maintaining temporal continuity and ensuring that future data is never used to predict the past—a core

requirement for avoiding look-ahead bias in time series forecasting (Bontempi et al., 2013).

- Minimum iterations per genre: 10. This ensures the model is exposed to diverse temporal contexts and mitigates the risk of overfitting to any single period, as recommended in sequential model validation practices (Tashman, 2000).

Performance will be evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) at each iteration. Final model reliability will be assessed on a holdout set comprising the most recent years (2023–2024).

2. Regression-Based Modeling Using XGBoost

XGBoost will be used to model the relationship between aggregated audio features and genre popularity, as well as to rank the relative importance of each feature. The tuning process follows a hybrid approach:

- Manual tuning (baseline): Initial values for parameters such as max_depth, learning_rate, and n_estimators will be chosen based on dataset size, feature dimensionality, and observed variance during EDA.
- Automated tuning (refinement): The model will then undergo optimization using methods such as random search, grid search, or Bayesian optimization, consistent with practices endorsed by recent studies on hybrid tuning strategies (SAGE, 2024; SAS, 2024; SSRN & JETIR, 2024).

Validation will be conducted using a time-aware cross-validation approach to ensure that temporal order is preserved. Each training fold will precede the test fold chronologically to avoid look-ahead bias and simulate real-world forecasting scenarios.

Model performance will be evaluated using:

- Root Mean Squared Error (RMSE): Measures overall prediction error, with higher penalty for large deviations.
- Mean Absolute Error (MAE): Captures average magnitude of prediction errors, regardless of direction.
- R^2 (coefficient of determination): Indicates how well the model explains the variance in the target variable.

Improvement will be tracked by comparing pre- and post-tuning performance to demonstrate the impact of parameter optimization.

To ensure robust generalization and prevent look-ahead bias, both models will undergo walk-forward validation using a 3-year in-sample window and a 2-year out-of-sample window, with a step size of 1 year to preserve temporal order (Hyndman & Athanasopoulos, 2021).

- ARIMA and XGBoost models will be trained using a walk-forward validation approach applied to data from 2008 to 2019 (~70%) and tested on subsequent 2-year intervals up to 2022 (~20%). This strategy allows the models to learn evolving patterns over time and validate on future-aligned data windows, consistent with real-world forecasting protocols (Hyndman & Athanasopoulos, 2021).A minimum of 10 walk-forward iterations per genre will be performed to capture dynamic temporal patterns.
- Final model performance will be evaluated on a holdout set (2023–2024, ~10%).

Together, the ARIMA and XGBoost models provide both predictive and explanatory power. ARIMA captures temporal evolution in genre popularity, while

XGBoost reveals which audio features most strongly influence those trends. The combined insights offer a data-driven framework for understanding genre dynamics and can assist music analysts, producers, and researchers in making strategic decisions.

This framework acknowledges limitations such as the exclusion of artist-level metadata, social media influence, and playlist curation—which, while influential, are beyond the scope of the dataset used.

4.2 Research Instrument or Sources of Data

The study will utilize data from the audio streaming platform Spotify, accessed via a public dataset hosted on Kaggle. The dataset, comprising approximately 650 high-popularity songs (popularity score ≥ 68), was compiled using Python scripts that extracted song data via the Spotify Web API. Each entry includes audio characteristics—such as energy, tempo, danceability, loudness, liveness, valence, speechiness, instrumentalness, acousticness—as well as descriptive attributes like song title, artist name, album release year, genre categories, and subgenre classifications.

4.3 Research Procedures

This section outlines the research procedures undertaken to prepare the dataset for analysis and modeling. The objective is to utilize a Kaggle-hosted dataset sourced from Spotify to analyze and forecast music genre popularity and identify influential audio characteristics contributing to genre trends. The research procedures are structured into two primary stages: Exploratory Data Analysis (EDA) and Data Treatment and Aggregation.

4.3.1 Exploratory Data Analysis

To establish a comprehensive understanding of the dataset's structure and variability, an exploratory data analysis (EDA) was conducted. The objective was to assess data integrity, identify potential outliers, and evaluate how genre-specific characteristics manifest in the dataset from 2008 to 2024.

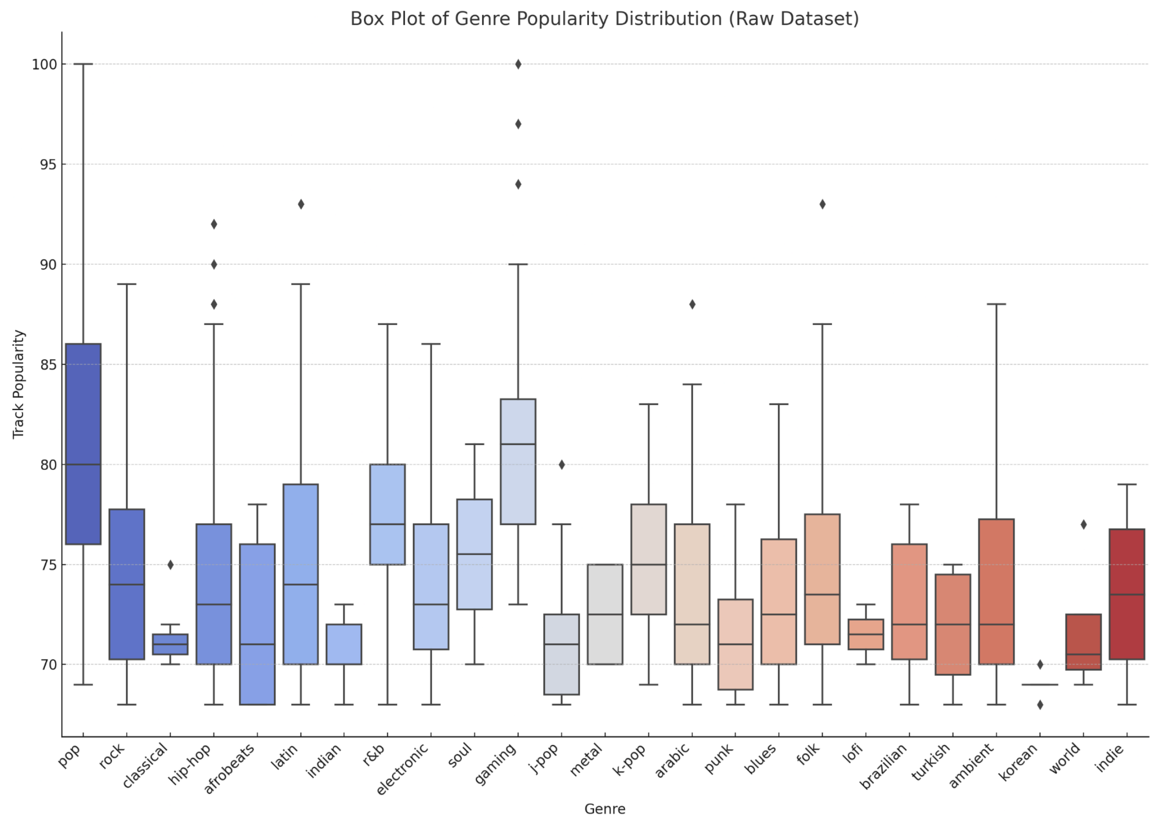


Figure 2. Box Plot (Before Refinement)

Before initiating data refinement, a box plot was generated to visualize the distribution of track popularity scores across all genres present in the dataset. The visualization reveals substantial variability in popularity scores, with genres such as Pop, Rock, and Hip-Hop exhibiting the widest range of values. These genres contain several extreme values, with some tracks reaching popularity scores as high as 100, indicating the presence of disproportionately popular tracks that could significantly skew aggregated metrics (Mora, García, & García, 2021; Nielsen Music, 2023).

In contrast, genres such as Electronic, Indie, and Classical displayed more concentrated distributions, with fewer extreme values and narrower interquartile ranges, suggesting more consistent popularity scores. The concentration of scores in these genres indicates a relatively stable level of popularity without the same level of high-impact outliers seen in more mainstream genres like Pop and Hip-Hop.

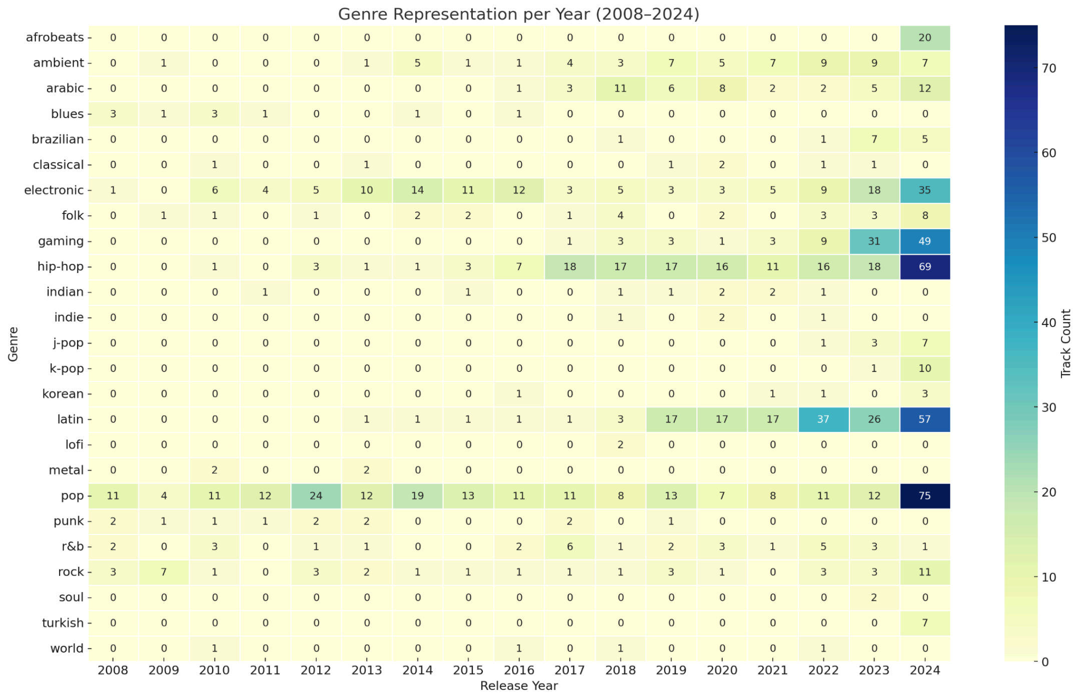


Figure 3. Heat Map

Here is a generated heatmap of genre representation per year (2008–2024). It shows how many tracks each genre has in each year. Based on the genre-year distribution heatmap from 2008 to 2024, the selected genres—Pop, Hip-Hop, Latin, and Electronic—exhibit consistent representation throughout the dataset, regularly meeting or exceeding the minimum threshold of 10 tracks per year necessary for statistical reliability (Hyndman & Athanasopoulos, 2021). This ensures robust temporal trend analysis and reduces the risk of bias due to data sparsity. In contrast, genres such as Indian, Indie, and Rock display sporadic or insufficient representation across multiple years, often falling below the modeling threshold, making them less suitable for time series forecasting or machine learning–based feature analysis.

Beyond their statistical reliability, the inclusion of these four genres is supported by their cultural and commercial relevance in the streaming era. According to the IFPI Global Music Report (2023), Pop continues to dominate global streaming consumption, serving as the mainstream baseline for international chart trends and playlist curation (Morris & Powers, 2015). Hip-Hop has grown into a global cultural movement, influencing fashion, language, and digital culture, and is especially prominent among younger demographics worldwide (Bradley, 2017). Latin music has experienced exponential global growth, with artists like Bad Bunny and J Balvin consistently topping international charts—reflecting its rise from a regional genre to a globally consumed one (Mora, García, & García, 2021; Nielsen Music, 2023). Meanwhile, Electronic music reflects the digitization of music production and consumption, thriving on platforms where algorithmic curation favors genres with measurable audio features such as tempo and loudness (Fikentscher, 2000; Prey, 2018).

Thus, the selection of Pop, Hip-Hop, Latin, and Electronic is empirically justified not only by their data availability and statistical viability, but also by their documented cultural and commercial influence in the global music landscape.

To assess the impact of outlier removal, a second box plot was generated to visualize the distribution of track popularity scores across the selected genres—Pop, Hip-Hop, Latin, and Electronic—after data refinement. As shown in Figure 4, Pop and Hip-Hop continued to exhibit the widest range of popularity scores, though the number of extreme outliers was significantly reduced.

Outliers were identified using the Interquartile Range (IQR) method, where the threshold was calculated by applying a 1.3 multiplier to the IQR—a more conservative approach than the conventional 1.5 multiplier. This adjustment was based on the recommendation of Rousseeuw & Hubert (2018) for datasets with known variability, such as streaming popularity data. The resulting upper threshold of 89.4 served to exclude tracks with exceptionally high popularity scores, thereby stabilizing the distribution without causing excessive data loss.

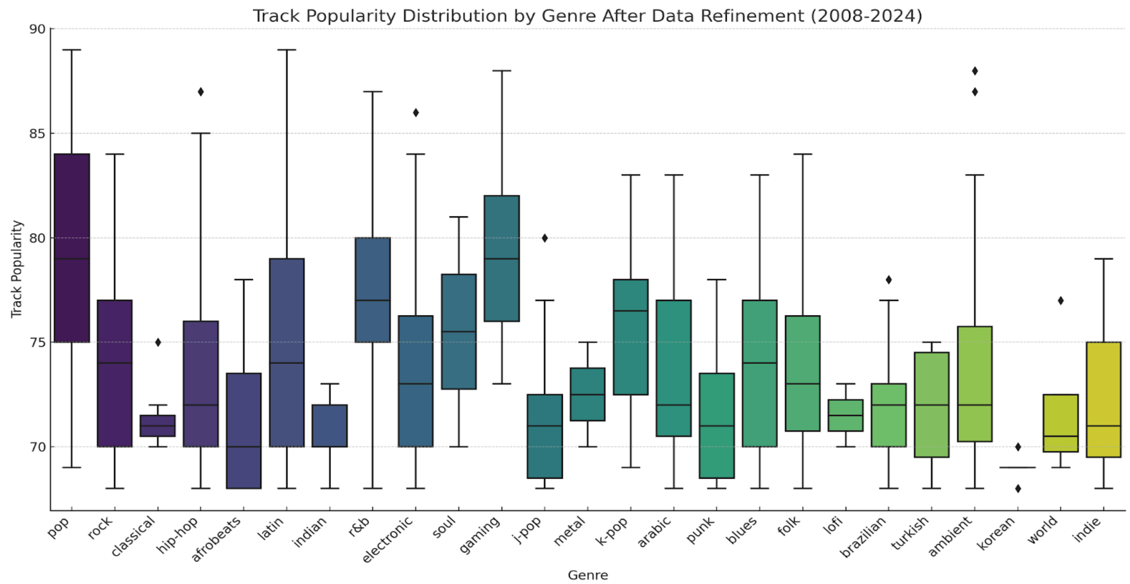


Figure 4. Box Plot (After Refinement)

This refinement step was particularly important for genres like Pop and Hip-Hop, which contained several tracks with scores approaching 100. Left unfiltered, these outliers could disproportionately influence aggregate statistics and obscure broader genre trends. By removing these extreme values, the dataset becomes better suited for median-based aggregation and forecasting, ensuring that central tendencies reflect typical rather than exceptional patterns.

4.3.2 Treatment of Data

The dataset underwent a structured data preprocessing pipeline to ensure quality and readiness for analysis and modeling. The goal of this stage was to prepare a cleaned, consistent, and analytically robust dataset suitable for temporal forecasting and regression-based modeling.

Preprocessing includes:

- Removing null values and duplicates to maintain data integrity.
- Filtering the dataset to include only tracks released from 2008 to 2024, aligning with Spotify's operational timeline and ensuring consistent metadata coverage.
- Track popularity outliers were identified and removed using the Interquartile Range (IQR) method with a 1.3 multiplier, yielding an upper threshold of 89.4. This reduced multiplier was chosen based on recommendations by Rousseeuw & Hubert (2018), who suggest such adjustments in datasets with known variability to balance between effective outlier removal and data retention. Tracks exceeding the 89.4 threshold—often seen in genres like Pop and Hip-Hop—were excluded to prevent these exceptionally popular tracks from skewing median calculations. This step is crucial for stabilizing data distributions and ensuring robust aggregation (Wilcox, 2017; Huber & Ronchetti, 2009).
- Extracting the release year from the track's release date for use in annual trend aggregation.

A genre representation analysis was performed by calculating the number of songs per genre per year within the filtered time frame. Based on the results, the study focuses on Pop, Hip-Hop, Latin, and Electronic, as these genres consistently met the minimum threshold of 10 songs per genre-year, supporting statistical reliability for both ARIMA forecasting and XGBoost regression (Hyndman & Athanasopoulos, 2021).

Genres such as Rock, Indian, and Indie were excluded due to insufficient representation, with average yearly track counts falling below the threshold. This exclusion is acknowledged as a dataset limitation, and the potential for supplementary external data collection is proposed for future research. Nonetheless, subgenre classifications within the retained genres were preserved to enable exploratory analysis of intra-genre variation (e.g., distinguishing between mainstream and niche subtypes such as Afro-Latin or underground electronic).

Following cleaning and filtering, the dataset was aggregated by genre-year pairs. For each pair, the median values of key audio features—energy, tempo, danceability, loudness, liveness, valence, speechiness, instrumentalness, and acousticness—as well as median popularity, were calculated. The median was chosen over the mean due to its resilience against outliers, a choice supported by Wilcox (2017) and Huber & Ronchetti (2009). This approach ensures that central tendencies accurately reflect typical genre characteristics without being distorted by extreme values.

Features such as tonality, musical key, and rhythmic structure were excluded based on findings from Interiano et al. (2018) and other music information retrieval studies, which suggest that these features have relatively low predictive power for modeling broad popularity trends. The selected features provide a more meaningful basis for both temporal modeling and feature importance analysis.

4.4 Analysis and Design

Figure 5 illustrates the structured architecture of the predictive system developed in this study, where the flow of actions is distributed across distinct roles: the Researcher initiates preprocessing, ARIMA performs forecasting, and XGBoost handles regression-based feature analysis. ARIMA was chosen over SARIMA and Prophet for its simplicity, interpretability, and suitability for univariate data with limited seasonal structure. Similarly, XGBoost was selected over Random Forest and SVR due to its robustness to multicollinearity, built-in regularization, and superior scalability across feature interactions (Chen &

Guestrin, 2016). Deep learning models such as LSTM were excluded due to dataset size limitations and interpretability constraints.

The system follows a linear, step-by-step data flow consistent with the structured development approach. The data is sourced from Kaggle and contains audio and descriptive features of tracks available on Spotify. Only tracks released between 2008 and 2024 are included, reflecting Spotify's operational data timeline and ensuring the reliability of popularity metrics (Hyndman & Athanasopoulos, 2021). After extraction, the dataset undergoes a preprocessing phase, during which null values and duplicates are removed to ensure data integrity. The cleaned dataset is then grouped by genre and release year, allowing for the creation of an aggregated dataset. For each genre-year pair, the median values of selected audio features—including energy, tempo, danceability, loudness, liveness, valence, speechiness, instrumentalness, and acoustic-ness—are computed, along with the median popularity. These features were selected based on their established relevance to listener engagement and genre classification in prior research (Schedl et al., 2014; Su & Zhou, 2022).

These median aggregations form the foundation for modeling genre behavior over time and understanding how musical traits correlate with listener engagement. This approach prioritizes genre-level trend modeling, though it may obscure the diversity of subgenres (e.g., indie vs. synth-pop). As such, the study focuses on identifying broad patterns rather than capturing fine-grained subgenre shifts.

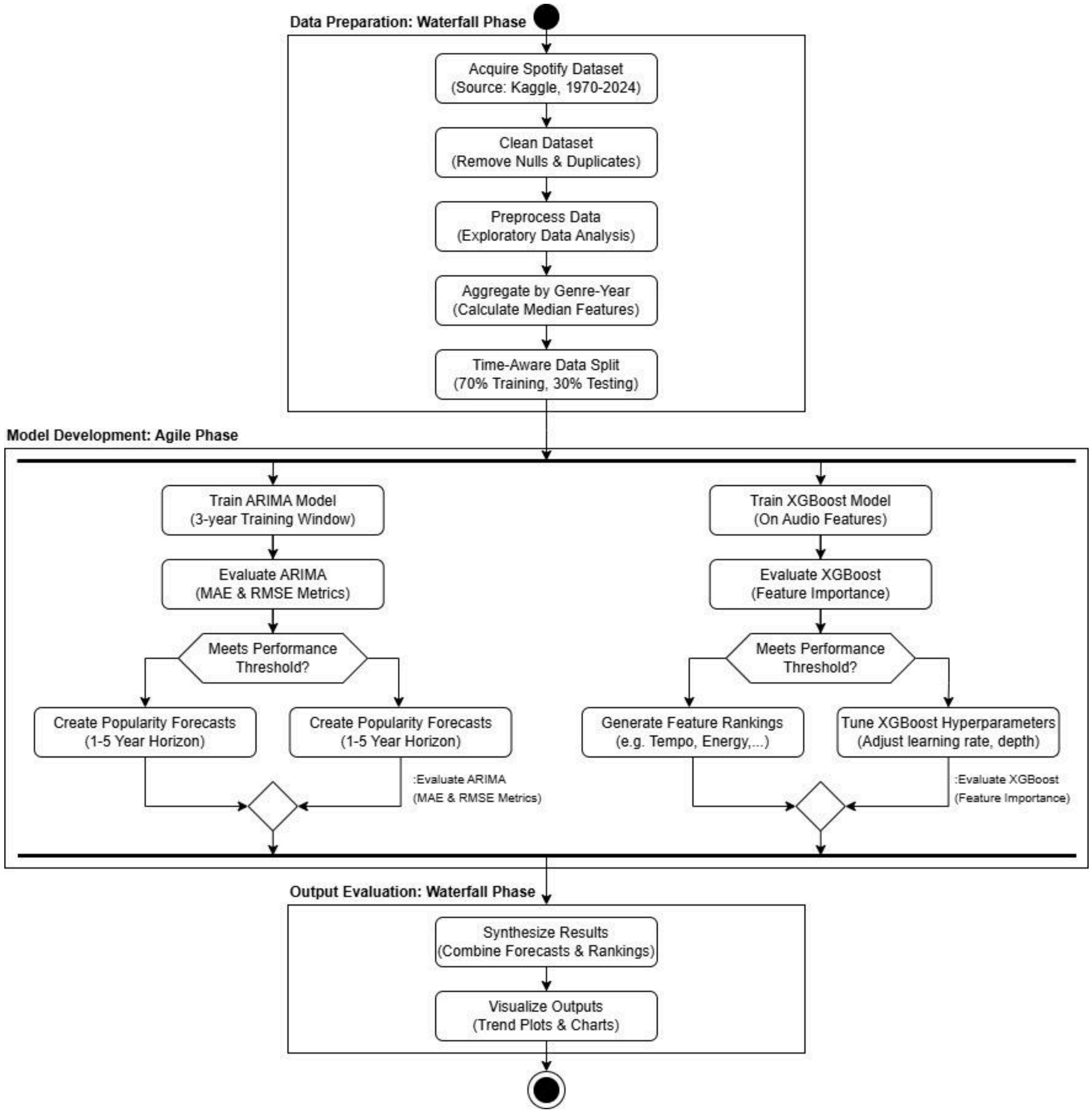


Figure 5. Activity Diagram

The activity diagram illustrated by Figure 5, provides a detailed visual representation of the methodological workflow for the research outline. Structured around a hybrid Waterfall-Agile development model, the diagram is divided into three phases–Data Preparation: Waterfall Phase, Model Development: Agile Phase, and Output Evaluation: Waterfall Phase. Data Preparation includes sequential steps such as acquiring Spotify data from Kaggle which spans from 1970 to 2024m cleaning it by removing nulls and duplicates, preprocessing with exploratory data analysis that includes box plots and heat maps, aggregating by genre and year to calculate median audio features, and performing a time-aware data split of 70% training and 30% testing. This ensures a robust dataset for analysis. This phase shall then transition into the Agile Phase via a fork bar, which initiates parallel processes, critical for the study's systematic yet flexible approach.

The development of the model features two iterative paths managed by thick synchronization bars: the ARIMA for time-series forecasting and the XGBoost for feature importance analysis. ARIMA's model shall be trained with a 3-year window using walk-forward validation, indicating 10 iterations per genre and shall then be evaluated with MAE and RMSE metrics. Either a 1-5 year popularity forecast is generated if the thresholds $(RMSE \leq 8, MAE \leq 6)$ are met, otherwise, it shall look back to tune its hyperparameters (p, d, q orders). Similarly, XGBoost shall be trained on audio features, evaluates feature importance, and either produces rankings or tunes hyperparameters with a loop back for re-evaluation. Converging at a joint bar, these iterative loops reflect Agile sprints–combining forecasts and rankings into trend plots and bar charts in the final Waterfall phase. This structure highlights the diagram's overall role in supporting the study's analysis and design through demonstrating a rigorous, adaptable methodology tailored to forecasting genre popularity on Spotify.

4.5 Verification, Validation, and Testing

This section outlines the strategies that will be used to verify, validate, and test the system developed in this study. The project will utilize both quantitative and qualitative evaluation approaches to ensure the accuracy, reliability, and relevance of the modeling outputs.

4.5.1 Verification

- Verification will ensure that the system is developed correctly according to its intended design specifications and methodological plan. The following procedures will be applied:
- Data preprocessing steps—including cleaning, filtering, and aggregation—will be verified through exploratory data analysis (EDA) and by cross-checking summary statistics such as count, mean, and standard deviation per genre-year pair.
- For ARIMA modeling, stationarity will be verified using the Augmented Dickey-Fuller (ADF) Test, with differencing and transformation applied as needed. Autocorrelation and partial autocorrelation plots (ACF and PACF) will be used to support order selection. Model diagnostics (e.g., residual analysis) will ensure the model satisfies ARIMA assumptions (Hyndman & Athanasopoulos, 2021).

For XGBoost modeling, input features will be examined for multicollinearity and missing values, and model configuration will be verified through hyperparameter tuning using grid or randomized search conducted within a time-aware validation framework that preserves the chronological structure of the data. This ensures that tuning respects temporal dependencies and avoids look-ahead bias. Feature importance plots will be used to confirm interpretability and relevance of the predictors (Chen & Guestrin, 2016).

4.5.2 Validation

- Validation will ensure that the developed models meet the research objectives of forecasting genre popularity and understanding the influence of audio features.
- For ARIMA, forecasts will be validated using walk-forward validation. The model will be trained on an expanding window of

prior years (e.g., 2008–2019) and tested on subsequent years (e.g., 2020–2024). Evaluation metrics will include Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) (Hyndman & Athanasopoulos, 2021).

- For XGBoost, validation will also use a time-aware train-test split, reflecting chronological order. Model accuracy will be measured using R^2 and MAE. Feature importance rankings will be reviewed to assess their alignment with known genre characteristics (e.g., Hip-Hop songs typically showing high energy and tempo).

Qualitative validation will include expert judgment through manual inspection of forecast plots and feature importance outputs to determine whether model behavior reflects real-world music trends and logical expectations.

MAE measures the average magnitude of errors between predicted and actual values, without considering their direction. It's a straightforward indicator of forecast accuracy.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \hat{y}_i \right|$$

RMSE gives more weight to larger errors by squaring the differences before averaging. It is sensitive to outliers and emphasizes significant deviations in predictions.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

R^2 indicates how well the model explains the variability of the target variable. A value closer to 1 means better predictive performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

$$y_i = actual\ value\ at\ i^{th}\ iteration$$

$$y^i = predicted\ value\ at\ i^{th}\ observation$$

$$\bar{y} = mean\ of\ all\ actual\ values$$

$$n = total\ number\ of\ observations$$

An acceptable RMSE for ARIMA will be ≤ 8 and MAE ≤ 6, based on the 0–100 scale of Spotify's popularity metric and error ranges reported in (Interiano et al., 2018). For XGBoost, acceptable performance is defined as MAE ≤ 5 and $R^2 \geq 0.7$.

If these thresholds are exceeded during validation, model retraining will be conducted using alternative hyperparameter configurations or revised feature sets. Additionally, a final holdout set (e.g., the most recent 10% of data) will be reserved for post-validation testing to evaluate model generalization and real-world robustness.

4.5.3 Testing

Testing will ensure that all system components function correctly without errors or inconsistencies.

- Model fitting procedures for both ARIMA and XGBoost will be tested using different data subsets to check reproducibility and consistency across multiple runs.
- The Python libraries statsmodels (for ARIMA modeling) and XGBoost (for feature-based regression) will be reviewed and

validated to ensure proper parameter configuration, correct model fitting behavior, and accurate output generation according to their intended forecasting or predictive functions (Seabold & Perktold, 2010; Chen & Guestrin, 2016).

- Additional testing will be performed to ensure that no runtime errors, exceptions, or incorrect parameter initializations occur during model training and forecasting.
- Visualization outputs (e.g., forecast plots, feature importance charts) will be manually inspected for consistency, completeness, and readability.

4.5.4 Scientific Methodologies to be Applied

Following the research guidelines for Computer Science (Zelkowitz & Wallace, 1998; PSITE, 2014; Wieringa, 2014), the following methodologies will be applied:

- The Experimental Method will be applied by testing different ARIMA and XGBoost configurations, and by evaluating model accuracy using defined metrics (MAE, RMSE, R^2).
- The Build Method will be applied through the construction of a predictive system with data preprocessing pipelines, modeling modules, and visualization components.
- The Model Method will be partially applied by treating ARIMA and XGBoost as abstract representations of genre popularity behavior over time, enabling analysis of temporal and feature-based trends.

4.6 Development Model

A Hybrid development model shall be used for the execution and completion of the machine learning model's development process. The combination of both development methodologies of Waterfall and Agile fits the linear and iterative nature of the study's framework. This section aims to balance structure and flexibility of the study wherein its requirements are partially known and subject to evolve or to be discovered. This development model shall then be able to have strict control over certain phases while also benefiting from the flexibility of Hybrid's adaptability.

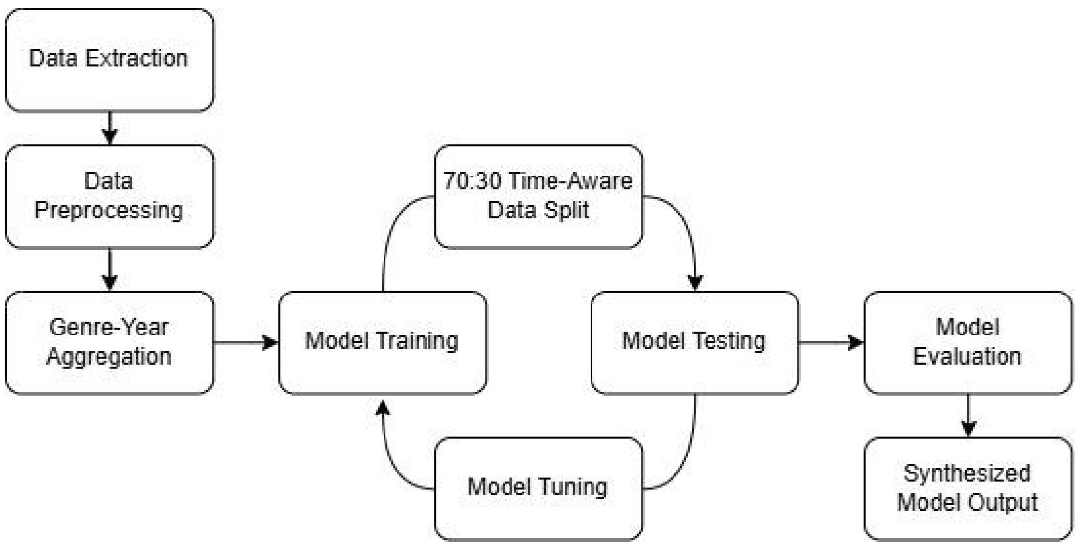


Figure 6. Hybrid Development Model

Figure 3 illustrates the early phases of the project–such as requirements gathering and planning–that follow a Waterfall-like sequential approach to ensure the preparation of the data cleaning and aggregation. Later phases then adopt an Agile-like practice of iterative development in sprints to allow adaptability and flexibility to changes. To ensure a seamless flow from the diagrams complete structure, a complete description for the phases integrated into the model are as follows.

The project is initiated with a Waterfall-like model wherein requirements, scope, and timelines for data preparation and aggregation are defined, ensuring a structured foundation before data processing. The sequential data preparation then starts with retrieving the raw Spotify Music Dataset from Kaggle and preparing it for analysis by removing null values and duplicates as its data cleaning phase. The genre-year aggregation is then processed, completing the first phase of the Waterfall-like model.

The iterative model development phase begins with the model training and testing, implementing a time-aware data split division of the aggregated dataset, wherein 70% is dedicated for training and 30% is for testing. Based on training feedback per iteration, hyperparameters are optimized iteratively to modify and/or improve performance and refining through iterative cycles. The models shall then be assessed through model evaluation by synthesizing results in the Agile-like review process. The final output shall then be produced which involves feature significance rankings and genre popularity forecasts, concluding the end of the development iteration phase.

By effectively integrating the strengths of both the Waterfall and Agile methodologies to address certain requirements of the machine learning model development for this study, the Hybrid development model establishes the structure's foundation through sequential data preparation and aggregation. Ensuring data integrity and readiness for analysis. Ensuring that the structure supports the necessary flexibility to adapt to evolving requirements and to optimize model performance through continuous feedback, enables the iterative process for model training, tuning, and evaluation. This balanced approach enables responsive refinement in later stages, ultimately supporting the delivery of robust, data-driven outputs for feature significance rankings and genre popularity forecasts.

4.7 Development Approach

This section outlines the systematic development approach employed in this study, utilizing a Bottom-Up methodology to ensure a logical and data-driven flow throughout the research process. Figure 4 illustrates this methodology, beginning with data acquisition and progressing through data processing, model implementation, and synthesis of results.

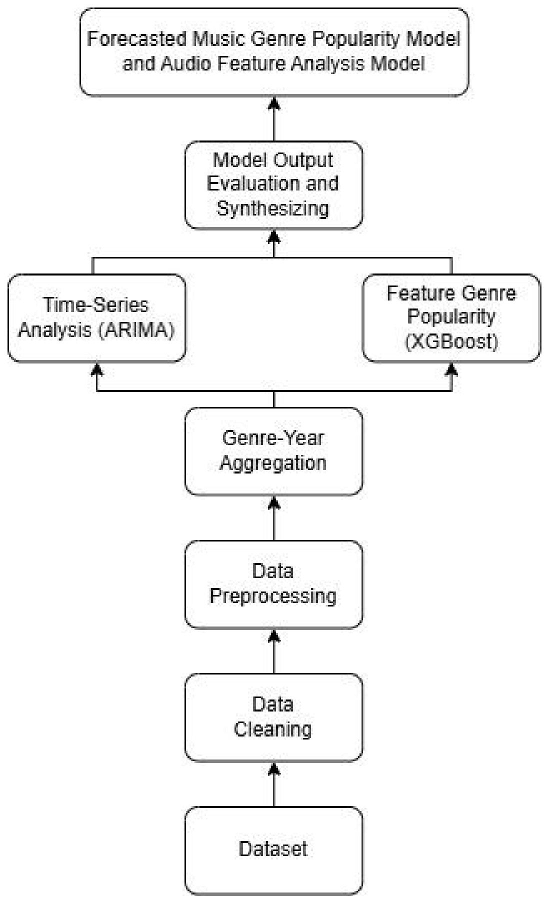


Figure 7. Development Approach

The bottom-up development approach model highlights how Spotify's audio features, as granular data, are refined through exploratory data analysis to lay the groundwork for an effective predictive model. This process is followed by the application of advanced analytical tools like XGBoost and ARIMA to construct a robust forecasting model for genre popularity trends. By leveraging data-driven exploration of Spotify's predefined audio features and incorporating component testing during the development of machine learning models, this method ensures precise forecasting and insightful audio feature analysis. Ultimately, this approach fosters the creation of targeted data points and validated techniques, paving the way for a more comprehensive and reliable model.

4.7.1 Data Acquisition

The dataset, sourced from Kaggle, serves as the foundational input for the analysis, comprising audio feature data spanning 2008–2024. The raw data undergoes rigorous cleaning and processing to ensure quality and consistency. Subsequently, the dataset is aggregated by genre and year, with the median values of key audio features (e.g., tempo, energy, danceability) calculated to provide a robust analytical foundation. This aggregated dataset serves as the input for both the ARIMA and XGBoost models.

4.7.2 Model Implementation

1. Time Series Analysis using ARIMA:
    - The ARIMA model is applied to forecast future genre popularity based on historical data patterns, capturing temporal dynamics within each genre.
    - The model is trained using a 3-year in-sample window and tested on a 2-year out-of-sample window, iterating with a 1-year step size to maintain chronological order.
    - This iterative walk-forward process ensures at least 10 validation iterations per genre, reducing the risk of look-ahead bias and enhancing generalization.
2. Regression Analysis using XGBoost:
    - XGBoost is employed to identify the audio features that most significantly impact genre popularity.

- By training the model on aggregated audio features, the analysis reveals the relative importance of features such as tempo, energy, and acousticness, providing feature-driven insights.

4.7.3 Synthesis and Integration

The integration of ARIMA and XGBoost provides both predictive and explanatory power, combining temporal forecasting with feature-driven analysis. The synthesized results will be presented through visual plots, illustrating both forecasted popularity trends and feature importance rankings. This dual approach enables a comprehensive perspective on genre dynamics, supporting data-driven decision-making in the music industry.

4.8 Software Development Tools

This section details the tools, software and library packages used in the research study.

Table 2. *Software Development Tools*

Software	Version	Uses
Visual Studio Code	1.99	Lightweight IDE for coding in Python and managing the project structure.
Jupyter Notebook	7.4.0	Interactive development environment for writing and testing code. Ideal for data exploration and documentation.
Git	2.49.0	A version control system that tracks changes

		in your code and datasets, making it easy to revert to earlier versions if needed.
GitHub	3.16.0	A cloud-based hosting service for Git repositories. Used to store, manage, share, and collaborate on your thesis project online.
Python	3.13.3	Main programming language used for data preprocessing, modeling, and visualization.
Pandas	2.2.3	Used for handling and manipulating datasets, especially for cleaning and aggregating Spotify audio features.
NumPy	2.2.4	Supports numerical operations, arrays, and mathematical functions. Used heavily in data analysis.
Matplotlib	3.10.1	For creating static visualizations such as genre trend line charts.
Seaborn	0.13.2	Enhances Matplotlib with more visually appealing statistical graphics.
XGBoost	2.0.3	Used to model the relationship between Spotify audio features and genre popularity and to identify feature importance rankings for each genre.
Statsmodels (ARIMA)	0.14.0	ARIMA is employed to model the temporal evolution of genre popularity over time, capturing underlying trends, seasonality, and noise components.

4.9 Project Management

To ensure the successful completion of the research project, effective planning and coordination shall be set in place. Each subsection provides detailed insights into the management and support for the timely and efficient execution of the research project.

4.9.1 Schedule and timeline

The following tables shall be schedules and timelines for completing the research proposal paper, which shall outline specific deliverables and their respective deadlines.

Table 3. *Gantt Chart of Activities, Second Semester, A.Y. 2024-2025*

DETAILS		TIMELINE																			
Deliverables		JAN				FEB				MAR				APR				MAY			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Research Topic Ideation	Theme																				
	Research Title																				
Research Definition and Planning	Research Objectives																				
	CH 2 Review of Related Literature																				
	CH 3 Technical Background																				
	CH 4 Design and Methodology																				
Research Development	Appendices and Curriculum Vitae																				
	CH 1 Introduction																				
	Abstract																				
Research Revision	Proposal Document Draft																				
	Initial Research Revision																				
Research Proposal Defense	Research Paper Submission																				
	Research Proposal Defense																				
	Research Paper Revision																				
	Submission of Final Revised Paper																				

Spanning the second semester of the 2024-2025 academic year, Table 2 illustrates the key deliverables and timeline of activities for the research proposal. This chart maps the duration for each deliverable, from research topic ideation to the research proposal defense and the dissemination plan following the thesis proposal and submission of the final revised paper. It details the timeline for submitting the research to Scopus-indexed journals and academic conferences, targeting acceptance by at least one journal and one conference, providing a clear overview of each achieved deadline. The timeline serves as a foundation for the second phase of the research study, set for the first semester of the 2025-2026 academic year, which will focus on developing the research output.

Table 4. *Gantt Chart of Activities, Summer 2025*

DETAILS		**TIMELINE**															
Deliverables		**MAY**				**JUN**				**JUL**				**AUG**			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Data Preparation	Gathering and Verifying Data		■														
System Development	Model Development		■	■	■	■											
System Testing and Verification	Model Testing					■	■	■									
	Accuracy Testing					■	■	■									
	Fixing Errors and Bugs							■	■	■							
Research Documentation	CH 5 Results and Discussions									■	■	■					
	CCH 6 Conclusion									■	■	■					
Research Journals	Searching for Conferences & Journal				■	■	■	■	■	■	■						
	Submission of Research Paper										■	■	■	■			
	Editing of Accepted Paper										■	■	■	■			
	Final Submission													■	■		
	Paper Presentation														■	■	■

Following the thesis proposal and submission of the final revised paper, Table 3 outlines the second half of the research study consisting of planned activities for the second phase of the research study, spanning May to August of the summer 2025. The timeline begins with data preparation in May, focusing on data collection and verification. System development follows, with model development occurring from May to June. This leads into system testing and verification, encompassing model and accuracy testing throughout June to early July, alongside debugging and error correction. The process concludes with research documentation, drafting Chapter 5 (Results and Discussions) and finalizing Chapter 6 (Conclusion) in July. This structured timeline ensures a seamless progression from technical development to comprehensive reporting.

Table 5. *Gantt Chart of Activities, First Semester, A.Y. 2025-2026*

DETAILS		**TIMELINE**																			
Deliverables		**AUG**				**SEP**				**OCT**				**NOV**				**DEC**			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Research Conferences	Application to Research Conference																				
	Preparation of Conference Requirements																				
	Submission of Conference Requirements																				
	Intended Conference Presentation																				
Research Documentation	Final Document Revisions																				
	Submission of Research Document to the Department																				

Illustrated in Table 4 is the intended planning that includes the application, submission and presentation to research conferences from August to early September. By the end of the semester, the entire research document shall then undergo final revisions before final submission to the department around December, fulfilling the requirements in completion for Thesis II for the first semester of the academic year 2025-2026.

4.9.2 Responsibilities

This research project is conducted by two researchers. Responsibilities have been divided based on the core modules and activities required for the completion of the study to ensure a collaborative and efficient workflow. The assignment of tasks is summarized below:

Table 6. *Table of Responsibilities*

Member	Assigned to	Description
Jonaz Juan II C. Sayson	Researcher	Gather data from Kaggle
	Analyst	Clean and pre-process dataset
		Train Models
		Analyze and plot outputs
Fabiola Villanueva	Researcher	Gather data from Kaggle
	Analyst	Clean and pre-process dataset
		Train Models
		Analyze and plot outputs

4.9.3 Budget and Cost Management

This research primarily utilizes publicly available datasets and open-source libraries, resulting in minimal direct costs. The following table summarizes the estimated expenses incurred throughout the course of the project:

Table 7. *Table of Expenses*

Items	Cost (₱)
Laptops (Lenovo Gaming Ideapad & Dell Inspiron)	71,995.00
Scopus-Indexed Journal Submission and Conference Presentation	35,000.00
Labor Cost – Research Work *(45 weeks active × 10 hrs a week) × ₱61.11 hourly rate × 2 researchers*	41,175.00
Printing, Transportation & Cloud storage	5,000.00
Total	153,170.00

This budget reflects a reasonable and practical allocation of resources required to conduct a quantitative, machine learning–based research thesis. The hardware expense supports the computational demands of statistical modeling. Miscellaneous costs cover research logistics. The paper submission cost reflects the intent to contribute findings to academic discourse through conference presentations or Scopus-indexed journal submissions. The labor component acknowledges the personal effort and time investment by the researchers throughout the semester-long study.

BIBLIOGRAPHY

Journal Articles

Joshi, S. Gupta, N., & Mahajan, R. (2025). Innovative tunes: Utilizing Machine Learning for predicting Spotify music popularity based on audio features. *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, 279–295. https://doi.org/10.1007/978-3-031-77075-3_23

Su, J., & Zhou, P. (2022). Machine learning–based modeling and prediction of the intrinsic relationship between human emotion and music. *ACM Transactions on Applied Perception*, 19(3), 1–12. https://doi.org/10.1145/3534966

Everts, R., Hitters, E., & Berkers, P. (2021). The working life of musicians: Mapping the work activities and values of early-career pop musicians in the Dutch music industry. *Creative Industries Journal*, 15(1), 97–117. https://doi.org/10.1080/17510694.2021.1899499

Brusila, J., Cloonan, M., & Ramstedt, K. (2021). Music, Digitalization, and democracy. *Popular Music and Society*, 45(1), 1–12. https://doi.org/10.1080/03007766.2021.1984018

Guo, X. (2023). The Evolution of the Music Industry in the Digital Age: From Records to Streaming . *Journal of Sociology and Ethnology*, 5(10). https://doi.org/10.23977/jsoce.2023.051002

Dolata, U. (2020). The digital transformation of the music industry. The second decade: From download to streaming. *Universität Stuttgart, Institut für Sozialwissenschaften, Abteilung für Organisations- und Innovationssoziologie, Stuttgart*. https://doi.org/https://hdl.handle.net/10419/225509

Pedersen, R. R. (2020). Datafication and the push for ubiquitous listening in music streaming. *MedieKultur: Journal of Media and Communication Research*, *36*(69), 071–089. https://doi.org/10.7146/mediekultur.v36i69.121216

Mokoena, N., & Obagbuwa, I. C. (2025). An analysis of artificial intelligence automation in digital music streaming platforms for Improving Consumer Subscription Responses: A Review. *Frontiers in Artificial Intelligence*, *7*. https://doi.org/10.3389/frai.2024.1515716

Oliveira, G. P., Couto da Silva, A. P., & Moro, M. M. (2024). What makes a viral song? unraveling music virality factors. *ACM Web Science Conference*, 181–190. https://doi.org/10.1145/3614419.3644011

Bhattacharya, A., Backonja, U., Le, A., Antony, R., Si, Y., & Lee, J. H. (2023). Understanding the influence of music on people's Mental Health Through Dynamic Music engagement model. *Lecture Notes in Computer Science*, 91–108. https://doi.org/https://doi.org/10.1007/978-3-031-28035-1_8

Schiavio, A., Maes, P.-J., & van der Schyff, D. (2021). The Dynamics of Musical Participation. *Musicae Scientiae*, *26*(3), 604–626. https://doi.org/10.1177/1029864920988319

Hansen, C., Mehrotra, R., Hansen, C., Brost, B., Maystre, L., & Lalmas, M. (2021). Shifting Consumption towards Diverse Content on Music Streaming Platforms. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 238–246. https://doi.org/10.1145/3437963.3441775

Logronio, J. D. (2024). Exogenous factors affecting the Manila music industry: A proposed marketing strategy model. *International Journal For Multidisciplinary Research*, 6(3). https://doi.org/10.36948/ijfmr.2024.v06i03.21979

Nijkamp, R. (2020). Explaining song popularity by audio features from Spotify data. *University of Twente*. https://essay.utwente.nl/75422/1/NIJKAMP_BA_IBA.pdf

Zhang, J. (2021). Music feature extraction and classification algorithm based on Deep Learning. *Scientific Programming*, 2021, 1−9. https://doi.org/10.1155/2021/1651560

Jitendra, M. S., & Radhika, Y. (2020). A Review: Music Feature Extraction from an Audio Signal. *International Journal of Advanced Trends in Computer Science and Engineering,* 9(2), 973–980. https://doi.org/10.30534/ijatcse/2020/11922020

Castellon, R., Donahue, C., & Liang, P. (2021). Codified audio language modeling learns useful representations for music information retrieval. *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*. https://doi.org/10.5281/zenodo.5624604

Saragih, H. S. (2023). Predicting song popularity based on Spotify's audio features: insights from the Indonesian streaming users. *Journal of Management Analytics*, 10(4), 693−709. https://doi.org/10.1080/23270012.2023.2239824

Jiang, J., Ponnada, A., Li, A., Lacker, B., & Way, S. F. (2024). A genre-based analysis of new music streaming at scale. *ACM Web Science Conference*, 191–201. https://doi.org/10.1145/3614419.3644002

Murauer, B., & Specht, G. (2018). Detecting music genre using extreme gradient boosting. *Companion of the The Web Conference 2018 on The Web Conference 2018*, 1923–1927. https://doi.org/10.1145/3184558.3191822

Qi, Z., Rahouti, M., Jasim, M. A., & Siasi, N. (2022). Music genre classification and feature comparison using ML. *2022 7th International Conference on*

Machine Learning Technologies (ICMLT), 42–50. https://doi.org/10.1145/3529399.3529407

Shariat, R., & Zhang, J. (2023). An empirical study on the effectiveness of feature selection and ensemble learning techniques for music genre classification. *Proceedings of the 18th International Audio Mostly Conference*, 51–58. https://doi.org/10.1145/3616195.3616217

Tang, H., You, F., & Liu, Y. (2023). Deep learning based music genre classification algorithm. *Proceedings of the 2023 3rd International Conference on Bioinformatics and Intelligent Computing*, 126–130. https://doi.org/10.1145/3592686.3592710

Zhang, W. (2022). Music genre classification based on Deep Learning. *Mobile Information Systems*, *2022*, 1–11. https://doi.org/10.1155/2022/2376888

Zhuang, Y., Chen, Y., & Zheng, J. (2020). Music genre classification with Transformer classifier. *Proceedings of the 2020 4th International Conference on Digital Signal Processing*, 155–159. https://doi.org/10.1145/3408127.3408137

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Gan, R., Huang, T., Shao, J., & Wang, F. (2024). Music genre classification based on VMD-IWOA-XGBOOST. *Mathematics*, *12*(10), 1549. https://doi.org/10.3390/math12101549

Hesmondhalgh, D., Campos Valverde, R., Kaye, D., & Li, Z. (2023). The impact of algorithmically driven recommendation systems on music consumption and production: a literature review. *UK Centre for Data Ethics and Innovation Reports*. https://www.researchgate.net/profile/David-Hesmondhalgh/publication/368

818373_The_Impact_of_Algorithmically_Driven_Recommendation_Systems_on_Music_Consumption_and_Production_A_Literature_Review/links/63fb3f815749505945448676/The-Impact-of-Algorithmically-Driven-Recommendation-Systems-on-Music-Consumption-and-Production-A-Literature-Review.pdf

Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. Frontiers in Psychology, 9, Article 904. https://doi.org/10.3389/fpsyg.2018.00904

Zelkowitz, M. V., & Wallace, D. R. (1998). Experimental models for validating computer technology. *IEEE Computer*, 31(5), 23–31. http://dx.doi.org/10.1016/j.jss.2008.06.040

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. Proceedings of the Python in Science Conference, 92–96. https://doi.org/https://doi.org/10.25080/Majora-92bf1922-011

Web Pages

Gupta, S. (2025, January 27). *Spotify trends, statistical facts and predict future move in 2025*. Xicom Blog | Software, Mobile App Development. https://www.xicom.biz/blog/spotify-statistical-facts/

Walfish, G. (2024, February 22). *The 6 most common challenges faced by musicians: Xposure Music*. Xposure Music Blog. https://info.xposuremusic.com/article/the-6-most-common-challenges-faced-by-independent-musicians

Barbara, S. (2023, July 5). *Unlocking success: How to make it in the music industry*. Li2Net Green. https://www.greenlinnet.com/what-does-an-artist-need-to-achieve-success/

Pant, T., & Shrestha. (2023, February 2). *Is it possible to predict music's commercial success?* American Marketing Association. https://www.ama.org/2023/02/02/is-it-possible-to-predict-musics-commercial-success/

Editor, E. (2023, September 14). *The impact of MP3 on the music industry: MP3 JUICE*. https://www.entrepo.co.za/impact-mp3-music-industry-mp3-juice/

Brohan, M. (2025, March 27). *Music ecommerce pioneer Napster acquired by Infinite Reality*. Digital Commerce 360. https://www.digitalcommerce360.com/2025/03/27/napster-acquired-by-infinite-reality-music-commerce/

Sheridan, N. (2025). *iTunes Marketing Mix 2025: A Case Study*. Latterly.org. https://www.latterly.org/itunes-marketing-mix/

Chhabria, S. (2025, January 17). *'Revolutionizing Rhythm': How Digital Platforms Are Reshaping The Music Industry In 2025*. Loudest.

https://loudest.in/authored-article/revolutionizing-rhythm-how-digital-platforms-are-reshaping-the-music-industry-in-2025-18002.html

Duarte, F. (2025, April 24). Music Streaming Services Stats (2025). Exploding Topics. https://explodingtopics.com/blog/music-streaming-stats

Curry, D. (2025, April 11). Music Streaming App Revenue and usage statistics (2025). Business of Apps. https://www.businessofapps.com/data/music-streaming-market/

Shaikh, E. (2025, March 25). Music Streaming Statistics 2025: Market Analysis & Trends. SimpleBeen. https://simplebeen.com/music-streaming-statistics/

Khan, S. (2021, February 16). Spotify vs. Apple Music vs. YouTube Music: Which is the Best Music Streaming Service for you?. Tech Times. https://www.techtimes.com/articles/257103/20210216/spotify-vs-apple-music-youtube-best-streaming-service.htm

Spotify. (2025, February 4). Spotify Reports Fourth Quarter 2024 Earnings. https://newsroom.spotify.com/2025-02-04/spotify-reports-fourth-quarter-2024-earnings/

Sajdak, S. (2024, August 6). *How streaming changed the music industry? Current impact, trends, and predictions*. Miquido. https://www.miquido.com/blog/how-streaming-changed-the-music-industry/

Słomka, D., & Sajdak, S. (2024, September 16). Experts Perspective: The Future of Music Streaming. Miquido. https://www.miquido.com/blog/the-future-of-music-streaming/

Gutierrez, V. A. (2024, May 2). Most Filipinos in the music industry earn less than P20,000. https://www.stii.dost.gov.ph/1867-most-filipinos-in-the-music-industry-earn-less-than-p20-000

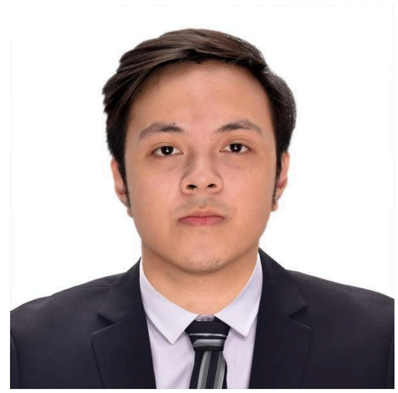
Born, G., Diaz, F., Ferraro, A., & Ferreira, G. (2024, January 7). *WP4B: Interdisciplinary interventions in the design of Music Recommender Systems*. UCL Anthropology. https://www.ucl.ac.uk/anthropology/research/music-artificial-intelligence-musai/projects/wp4b-interdisciplinary-interventions-design

Books

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice*. OTexts. https://otexts.com/fpp3/

APPENDICES

CURRICULUM VITAE



CONTACT INFORMATION

Full Name: Jonaz Juan II C. Sayson

Current Address: Guizo, Mandaue City

Telephone number: 09166245728

Email address: jonaz.juan@hotmail.com

PERSONAL INFORMATION

Age: 24

Citizenship/Nationality: Filipino

Gender: Male

Marital Status: Single

Language proficiency: English, Cebuano, Tagalog

Computer PMS and Software skills: Proficient in C, React.js, Next.js, JavaScript, MantineUI, Python, Git, TailwindCSS, Github

EDUCATIONAL BACKGROUND

Educational Level

Tertiary Level (2022 – present)

University of San Carlos

Bachelor of Science in Computer Science

STRENGTHS/TRAITS & SKILLS

Resilience

Proactivity

Adaptability

Detail-Oriented

Action-Oriented Learner

CAREER OBJECTIVES

I aim to leverage my skills and passion for technology to develop innovative, tech-driven solutions that address real-world problems. By combining creative problem-solving with technical expertise, I aspire to contribute meaningfully to projects that improve efficiency, enhance user experiences, and drive positive change across industries.

Last updated on April 30, 2025

CURRICULUM VITAE

CONTACT INFORMATION

Full Name: Fabiola C. Villanueva

Current Address: Pitalo, San Fernando, Cebu, 6018

Telephone number: 0917 322 2103

Email address: fabiolavillanueva3312@gmail.com

PERSONAL INFORMATION

Age: 22

Citizenship/Nationality: Filipino

Gender: Female

Marital Status: Single

Language proficiency: English, Tagalog, and Cebuano

Computer PMS and Software skills: Proficient in C, React.js, Next.js, JavaScript, Tailwind-DaisyUI, Python, Git

EDUCATIONAL BACKGROUND

Educational Level

Tertiary Level (2022 – present)

University of San Carlos

Bachelor of Science in Computer Science

STRENGTHS/TRAITS & SKILLS

Resilience

Proactivity

Adaptability

Detail-Oriented

Action-Oriented Learner

CAREER OBJECTIVES

To leverage my passion to educate and inspire future generations through my expertise in computer science while ultimately building a career driven by continuous exploration and innovation in technology.

Last updated on April 30, 2025