1 DATASET

1.1 Data Collection Procedure

To create our dataset, we performed a two-step procedure. Initially, we launched an online survey to identify willing participants and gather their Spotify IDs and personal attributes. Subsequently, utilizing these Spotify IDs and the Spotify API, we systematically retrieved their publicly available playlists along with detailed information about the songs and artists within these playlists.

The Survey. We recruited participants for our research via an online survey. Initially, participants were requested to furnish their distinctive Spotify IDs, a handle for accessing their Spotify data, particularly their playlists. Subsequently, we sought additional self-provided information from participants to serve as attributes for our analyses. In total, we gathered 16 attributes:

- **demographics**: gender, age, country, relationship, if they live alone, economic status, and occupation;
- **habits**: sport, smoking, and alcohol habits, and if they are Premium Spotify subscribers;
- **personality**: we utilized the 10 short personality questions [6] to retrieve OCEAN personality traits [9].

Table **??** reports a short description of the attributes. The survey was conducted in English and had an approximate completion time of 4 minutes. To ensure data quality, we incorporated attention-check questions and cross-referenced information with the Spotify API, enabling us to identify and filter out inconsistent or unreliable responses. The survey was active from May to September 2022, and distributed primarily through social networking platforms, notably Reddit, Facebook, and Telegram. Our survey was strategically shared within popular Spotify and music-oriented groups, including */r/Spotify/*, *r/Music*, and *Facebook Spotify Music*. By participating in the survey, users explicitly permitted us to utilize their data for this study. We also offered various points of contact, enabling users to request the removal of their data from our dataset. As an additional safeguard, we meticulously anonymized their data, eliminating any Personally Identifiable Information (PII) they might have shared with us. For privacy considerations, we have decided not to publicly release our dataset, not even in anonymized form, as the risk of re-identification remains a concern.[1]

Survey results and validation. We received a total of 1081 responses from individuals across 76 different countries, spanning an age range of 13 to 55 years.[2] Subsequently, we excluded respondents who did not pass the attention test or had no publicly accessible playlists on Spotify, resulting in 739 users constituting our final dataset. Table **??** reports the distributions of our targets at user and playlist levels. Similar to previous works [2, 3, 8], we grouped age in bins of interest (e.g., underage), and the results of personality traits, which range from 0 to 100, into three categories, differentiating low, middle, or high scores (boundaries = [33.3, 66.6, 100]). Our distributions align to global Spotify statistics[7]. For instance, 58% of users are Male (our is 68%), and the majority of users (62%) are in the age range 18-35 (we also have the majority in the range 18-30). The majority of our participants are European (40%) and North American (33%), similar to the global distribution (34% and 24%, respectively) [1]. Despite being far smaller than the overall amount of Spotify users, such a number still allows us to draw statistically significant results. Indeed, we are above the minimum sample size of 384 required by setting a confidence level of 95%, a margin of error of 5%, a population proportion of 50%, and a population size of 500 million [5].

Remark. It is crucial to emphasize that our primary objective is to establish the existence of a connection between users' playlists and personal attributes. However, this relationship may exhibit variations in a broader and more representative context. Nonetheless, all our analyses are rigorously validated through statistical tests, ensuring the robustness and reliability of our findings.

Retrieving data from Spotify. Leveraging the official Spotify API[3], we systematically gathered the public playlists of our participants by means of their Spotify ID. Each API call yields a JSON file per playlist, containing details about the playlist owner (e.g., Spotify ID and nickname) and the playlist's constituent tracks (comprising track names, artists, and the dates when tracks were added to the playlist). In the following sections, we discuss in deep our data.

1.2 Features

To create our playlist dataset, we harnessed multiple Spotify APIs to collect information about the playlist, their constituent songs, and the artists who composed them. Then, we consolidated this information to craft a unified representation consisting of 111 features for each playlist, serving as the foundation for our experiments. In the following sections, we provide a succinct overview of the diverse feature categories, which encompass *Songs*, *Artists*, *Genres*, and *Miscellaneous* (Misc) attributes. For a comprehensive list of these features, please refer to our repository.

1.2.1 Songs Features. For each song in a playlist, we accessed the `tracks` and `audio-features` APIs to retrieve songs' information. Specifically, we collected the popularity, whether it contains explicit content, the release year, duration, and several audio features calculated by Spotify algorithms, namely, danceability, energy, loudness, speechiness, acousticness, instrumentalness, liveness, valence, and tempo. For each playlist, we aggregated the songs's information in a single entry by calculating the mean, standard deviation, minimum, and maximum of each value, except for the feature *explicit*, where we calculated the occurrence percentage. For an exhaustive explanation of the song features, we recommend referring to the official Spotify API documentation.

1.2.2 Artists Features. Within each song, there is often a collaboration among one or more artists. To comprehensively capture the characteristics of these contributing artists, we employed the `artists` API. This API enabled us to extract essential information, including an artist's popularity and the number of followers. When aggregating the artists' information for each playlist, we computed various metrics such as the total count of artists (both overall and unique), the proportion of artists with lower popularity (popularity < 20, both overall and unique), the percentage of songs created by a

[1] However, we are willing to share the data with reviewers to ensure reproducibility, and the code underpinning our experiments will be made available upon acceptance.
[2] Participants under the age of 13 were excluded to adhere to European regulations, as further detailed in the Ethical Consideration section.
[3] https://developer.spotify.com/

single artist, the prevalence of artists making multiple appearances, the artist diversity (measured through the Simpson Index [4]), and insights into artists' popularity and follower counts on Spotify. Analogous to our approach for song features, we aggregated the numerical attributes associated with the contributors to all tracks by computing mean, standard deviation, min, and max values.

1.2.3 Genres Features. Calling the `artists` API, we gained access to the genres attributed to each artist. We associated each song with the genres corresponding to the artists involved in its creation. In our data aggregation process, we calculated the proportion of songs within the playlist that fell under specific genres, encompassing 30 popular genres (e.g., rock, pop, indie, metal). Additionally, we incorporated a feature for local genres, denoting those linked to geographical locations, and a category for other genres.

1.2.4 Misc Features. For each playlist, we systematically compiled the total number of songs, the count of followers, the diversity in terms of the albums from which the songs originated, and a record of the years when the songs were added to the playlist.

1.3 Final Dataset

The final playlist dataset comprehends 10,286 public playlists made by 739 users. On average, each user has 13.9 playlists (STD = 35.2), and each playlist has 38.9 songs (STD = 34.3). Each playlist is linked with the attributes of the user who created it (e.g., age, gender), and comprehends the aggregated information of all its songs and relative artists, for a total of 111 features. In total, we extracted information from 221,008 unique songs and 55,074 unique artists.

2 MODEL SELECTION

For each model we perform a complete model selection strategy to select the hyperparameters' values in a correct and fair way, and ensure the Each model is selected through a grid-search validation using the validation set. We now describe models' hyperparameters:

- Logistic Regression (LR): (1) C = $\{0.01, 0.1, 1, 10\}$; (2) fit intercept = $\{true, false\}$; (3) class weight = $\{None, balanced\}$.
- Decision Tree (DT): (1) criterion = $\{gini, entropy\}$; (2) max depth = $\{None, 3, 5, 10\}$; (3) class weight = $\{None, balanced\}$.
- Random Forest(RF): (1) criterion = $\{gini, entropy\}$; (2) max depth = $\{None, 3, 5, 10\}$; (3) class weight = $\{None, balanced\}$; (4) n estimators = $\{16, 32, 64, 128\}$.
- kNN: (1) n neighbours = $\{3, 5, 7, 10\}$; (2) n estimators = $\{uniform, distance\}$.
- MLP: (1) hidden layer size = $\{(|\mathbf{p}|,), (|\mathbf{p}|, \frac{|\mathbf{p}|}{2}), (\frac{|\mathbf{p}|}{2},)\}$; (2) activation = $\{relu, tanh\}$; (3) solver = $\{adam\}$; (4) learning rate = $\{adaptive\}$; (5) learning rate init = $\{0.01, 0.001, 0.0001\}$; (6) alpha = $\{0.01, 0.001, 0.0001\}$.

REFERENCES

[1] Maleha Afzal. 2023. Top 20 Spotify Users by Country. https://finance.yahoo.com/news/top-20-spotify-users-country-190010495.html. Accessed: October 2023.

[2] Sara Bunian, Alessandro Canossa, Randy Colvin, and Magy Seif El-Nasr. 2017. Modeling individual differences in game behavior using HMM. In *Artif. Intell. Interact. Digit. Entert. Conf.*

[3] Terence Chen, Roksana Boreli, Mohamed-Ali Kaafar, and Arik Friedman. 2014. On the effectiveness of obfuscation techniques in online social networks. In *Privacy Enhancing Technologies: 14th International Symposium, PETS 2014, Amsterdam, The Netherlands, July 16-18, 2014. Proceedings 14.* Springer, 42–62.

[4] CJ Keylock. 2005. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. *Oikos* 109, 1 (2005), 203–207.

[5] J. Kotrlik and C. Higgins. 2001. Organizational research: Determining appropriate sample size in survey research. *Inf. Tech. Learn. Perf. J.* (2001).

[6] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality* 41, 1 (2007), 203–212.

[7] Kate Sukhanova. 2023. Spotify Statistics 2023: User Demographics, Growth Rate, and Revenue Breakdown. https://techreport.com/statistics/spotify. Accessed: October 2023.

[8] Pier Paolo Tricomi, Lisa Facciolo, Giovanni Apruzzese, and Mauro Conti. 2023. Attribute inference attacks in online multiplayer video games: A case study on Dota2. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy.* 27–38.

[9] Jerry S Wiggins. 1996. *The five-factor model of personality: Theoretical perspectives.* Guilford Press.