Using Machine Learning to predict the impact of music on moods and emotions

1. Introduction

1.1. Literature review

Mental health is a critical component of the well-being of individuals that can have deep-rooted consequences on societies, economies and businesses. While the root-causes of prevalent mental health issues, such as depression, stress or anxiety are still poorly understood, such issues are becoming increasingly widespread in today's societies: Depressive disorders are the most common, with current research estimating that up to 1 in 3 women and 1 in 5 men will experience depression at some point in their lives (Dattani et al, 2023). Such alarming figures were further exacerbated by the COVID-19 pandemic: Research conducted during that period estimates that over 50% of adults experienced psychological distress during the lockdowns, with a further 20-30% experiencing PTSD, depression or anxiety (Nochaiwong et al, 2021). Young individuals were disproportionately affected, with one CDC survey reporting over 37% of high school students feeling that "their mental health was not good most or all of the time" (Schaeffer, 2022). Beyond the obvious impact on personal well-being, recent studies have suggested that poor mental health can prove extremely costly for businesses and the economy as a whole: Greenberg et al (2003) puts the total cost in the US alone at 52.9 billion dollars in 2000 (increasing by 21% vs. 1990), with workplace costs bore by businesses making-up over 62% of the figure. As a result, tackling and improving mental health represents a significant opportunity with a tangible economic and societal impact.

Interestingly, some studies have highlighted a link between well-being and music: Gustavson et al (2021) and Burns et al (2002) both suggest that exposure to certain types of music, songs and sounds could help regulate emotions and have a positive impact on overall mental health state and stress. Furthermore, research from McCraty et al (1998) and Labbé et al (2007) found that specific genres of music had different impacts on an individual's mental state: Genres such as grunge and heavy metal were associated with higher tension, stress and hostility while classical and designer music led to better relaxation and mental clarity. While such studies offer promising perspectives, it is worth noting that their sample sizes were relatively small (144 subjects for McCraty et al and 56 for Labbé et al). Gustavson et al (2021) further underlines that point, highlighting that despite the multiple studies conducted and encouraging findings, there is a clear lack of largerscale research on the topic, with most papers using smaller samples.

1.2. Purpose

Given the growing prevalence and potential economic cost associated with mental health issues, there is a clear opportunity for both businesses and public stakeholders to conduct further research on the topic and identify potential root causes and remedies. For this paper, we want to specifically focus on the research conducted by Gustavson et al (2021) and Burns et al (2002), and attempt to understand how music and listening habits can help us predict the mental state of individuals and take appropriate mitigation actions should potential risks or issues be detected.

Additionally, there is a significant opportunity for largerscale studies and machine learning applications, given the concerns on sample sizes highlighted by Gustavson et al (2021). Our approach will therefore be to use relatively large surveys and dataset that includes information about individuals' listening habits and mental health status. We will use this dataset to extract specific features related to music (Genres, BPM, danceability, characteristics, etc.) and general user profile (age, mental state, etc.) to attempt to predict different mental health issues - anxiety, depression, insomnia and OCD – and grade their severity and risk potential.

Such an analysis would provide an effective way to leverage the user profile and listening data that Spotify and other platforms collect on a regular basis, and use it to tackle mental issues early on. Streaming platform could then use a gentle nudge system based on the predictions, either directing the user to mental health helplines and resources (similar to what Google does when prompted with "suicide" or related search words) or offer music recommendation tailored to improving the mood and mental state of the individual. Such a framework would then provide an effective solution to identify potential risks and take preventive actions early-on, so that users can promptly seek help.

2. Dataset overview

Our analysis is centered around an existing survey uploaded to Kaggle that collects valuable information about the listening history, habits and mental health status for 736 participants. As this data only includes listening habits at genre-level (i.e. favourite genre and listening frequency across genres, the details of the songs the users listened to is not included), we combined the Kaggle data with a set of representative playlists for each genre from Spotify to get more granular song-level features. Important to note that we assume our playlists are representative of each genre, i.e. they should match the users' listening habits on aggregate: which we ensure by choosing popular genre focused playlists We then used the Spotify API to extract the respective metrics from the songs in the playlists and collected their lyrics using a mix of Azapi (see reference 5.2.4) and manual copy and paste. The raw datasets and respective variables are as follows:

2.1. Data sources

Kaggle Music & Mental Health Survey Results (See reference 5.2.1): This is a Google survey conducted by University of Washington over a period of 2 months in 2022 for 736 users. It collects information about the profile, listening habits and mental state of respondents with the following variables:

- Timestamp (date): Time of survey submission.
- Age (integer): Age of the user.
- **Primary streaming service (cat.):** Primary music streaming platform (e.g. Apple Music, Spotify, etc.).
- Hours per day (integer): Number of hours per day respondents spent listening to music.
- While working (binary): Whether respondents listened to music while working.
- Instrumentalist (binary): Whether respondents play an instrument
- **Composer (binary):** Whether respondents compose music.
- Fav gender (cat.): Favourite music genre.
- **Exploratory:** Whether users explored new genres and artists
- Foreign Languages (binary): Whether users listened to music in a foreign language.
- **BPM (number):** Average BPM (Beats Per Minute) of favourite genre.
- Frequency (cat).: Includes 16 children columns associated with each genre as follows: Classical, Country, EDM, Folk, Gospel, Hip hop, Jazz, K-pop, Latin, Lofi, Metal, Pop, R&B, Rap, Rock, Video game music. Each of these columns specify how frequently users listened to each genre with a categorical label (never, rarely, sometimes, very frequently)
- Anxiety, Depression, Insomnia, OCD (integer): Self-reported severity of each mental health issue on a scale of 0 to 10.
- **Music effects (cat.):** Whether music improved the mental health of respondents (i.e. Improve, no effects, worsen)

While the survey data mainly aggregates listening behavior on a per Genre level, presumably the Genre itself is just an intermediate between the impact of a song metrics and the mental health state. To gain further insights into individually preferred song characteristics, we aim to approximate this information by analysing song features in more detail. For this we first extract song metrics (provided by Spotify). Furthermore, we suspect that song lyrics could be an important information carrier and aim to analyse these further. To understand which songs to analyse per user, we choose a representative playlist per genre.

Representative Playlists per Genre (Spotify): We manually selected one playlist per genre from Spotify to get a list of representative songs. If available, we chose playlists titled 'Top X Genre' or alternatively, we opted for 'Genre X Essentials' playlists. This is our foundation for further song-level analysis. Selecting top playlists as such ensures that our song selection is a "close-enough" representation of the potential set of songs each user could listen to depending on their preferences.

Spotify API for song metrics (see reference 5.2.3): Based on the song names we retrieved a set of 21 features for each song included in the playlists above using the Spotify API, out of which we retained the 9 most relevant: danceability, energy, loudness, mode, speechiness, acousticness, instrumentalness, liveness and valence. These features helped us to get a better understanding of the moods and feelings associated with each song, providing valuable additional information to our analysis.

Azapi API for song lyrics (see reference 5.2.4): Our last remaining missing feature is lyrics, as these are not provided in the Spotify API. We leveraged Azapi (See reference 5.2.4) which collects lyrics from AZlyrics.com, based on the song and artist's name. We retrieve the full song lyrics for each selected song. As some songs were not available on AZlyrics, or not collectable due to language restriction, we manually collected these lyrics and included them in the dataset. To provide meaningful insights, these lyrics need to be analysed.

2.2. Feature extraction

Song lyrics sentiment analysis Approach #1: NLP The key feature missing from our combined dataset is the sentiments from the lyrics themselves. While the song metrics "valence" provides some degree of measure on the song's undertone by providing a score between 0 and 1 rating the song's positiveness, it is not clear whether this is calculated on the basis of lyrics or musical qualities (e.g. key, danceability, BPM, etc.) of the song. We therefore decided to include an additional variable in the dataset describing the emotion conveyed by the lyrics. To extract this, we first opted to employ a T-5 model. It is a generative text-to-text model developed by Google, extensively pretrained for complex NLP tasks, and specifically applicable for transfer learning. Given its ability for multilabel generation it is a fit to conduct an emotion analysis on the songs' lyrics. As we want to finetune the model for our task, we need to develop a task identifier prefix (which will be "classify emotion: "), that sets the model focus and find a labelled dataset for supervised training.

We started by collecting a labelled training dataset for our model: We chose the Kaggle dataset highlighted in section 5.2.5, which provides over 40,000 tweets and 13 associated feelings (sadness, enthusiasm, worry, etc.). As we could not find any labelled training data (in English) with song lyrics or poems specifically, we hoped that this dataset is able to provide a good basis for model training.

We embedded the model in a Pytorch lightning module to facilitate training and testing using the tweets and their associated sentiment. For this we conduct a train-test split, tokenize the text input and prepare dataloaders. After training we use the model to predict the primary emotion conveyed by the lyrics, aiming to add it as an additional feature to our dataset.

In our first approach we make use of t5-small, optimizing for loss and logging the exact match between expected and predicted output (as it is a generative rather than a classification task theoretically) during 10 epochs, utilizing Adam and a learning rate of 5e-5.

While the model performed well during train and test (test $_loss = 7.68e-07$), the results after prediction are rather unexpected and not usable, because instead of an emotion they are repeating parts of the respective lyrics. Potential causes for this problem could lie in the dataset size used for training comparing to the amount of classes present, dataset differences (tweets vs. lyrics) or in too little training time (therefore computational resource limits).

To overcome this issue, we facilitate the class division to three classes: positive/neutral/negative. Furthermore, we upgrade the model to T5-base to boost transfer learning capabilities of the model itself. Nevertheless, this comes with the drawback of the need to reduce the number of epochs to three due to computational resource limits. The setup of the rest of the model and the task identifier stays the same.

While the model again performed well during train and test (test $_{loss} = 8.3e-05$), the results after prediction are still not usable, repeating parts of the respective lyrics again. We presume that the Data difference between the Tweets and Lyrics could be too big, and as we are not able to increase our computational resources we need to change the approach and model.

Song lyrics sentiment analysis Approach #2: LLM

Given the unsatisfactory results of the T-5 predictions above, we decide to switch towards an LLM model. This has the beauty of enabling us to ask multiple purposeful questions. We extract the top 3 emotions of the lyrics which we now only use to selectively check whether the sentiment of the song was correctly grasped. More importantly as OpenAI has a deep contextual understanding and interpretation ability, we directly check if there is any indication of each of the four mental health concerns in the lyrics, because we assume that listeners could identify with the story told in it.

For this we use langchain to pass the lyrics alongside our prompt to the model and expect an output in binary format (textual for the top three emotions). The model we use is gpt-3.5-turbo, where we set the temperature and top_p equal to zero to ensure prediction accuracy and a higher level of replicability.

Our prompt is targeted at priming the model for our task to heighten the accuracy and at ensuring the reliability of the output format of the answers to facilitate further processing (example excerpt – for more details pls refer to code):

'Please set yourself in the position of a language expert, musician, songwriter and psychologist. Given the following songlyrics {topic} please answer my following questions. [...] Take a deep breath before you start'

'Please ensure that the answers are in the format of Q1: your_answer /n Q2: your_answer /n Q3: your_answer /n Q4: your_answer /n Q5: your_answer.'

'Q1 What is the main emotion transferred through this lyric? Please display only the top 3 emotion names (as adjectives) without wrapping it in a sentence (i.e. happy, energized, sad).'

'Q2 Is there any indication of depression in these lyrics? Please answer numerically with 1 if its true and 0 if its false.'

We pass all lyrics collected across the different genres to the model. We used Q2-4 to create 4 additional binary columns in our dataset for each mental issues reported by the users, based on the output of the prompt. The column indicated connotations associated with each of the disorders as a "1", e.g. a song whose lyrics is associated with anxiety would have a 1 label in the anxiety column and 0 otherwise. (Q1 was used to sense-check the results to ensure consistency and accuracy of the findings.) This approach yielded much better results than T-5 overall.

2.3. Merging process

The last step in the data collection is to merge the songlevel information (Spotify song features and the binary lyrics sentiment collected via LLM) with the survey data to obtain a user-level dataset with combined information.

First we prepare the survey data by converting the categorical columns of genre frequency listening behaviour into ordinal numerical values (where "Never" equals 0 and "Very frequently" equals 3). As this metric is still measuring in absolute values, we create a second set of normalised listening frequency to reflect the proportion of each genre in user's listening profile. This prevents skewed results based on higher listening volumes of some users

overall and provides us with relative genre preference profiles per user (where all genres listened to add up to one).

Second, we prepare the song metrics and mental health indicators from the lyrics data for merging by aggregating the metrics based on genre to calculate the mean per metric per genre.

Finally, we calculate the weighted metrics per genre for each user. The normalised listening frequencies are used for weighting the genre metrics, resulting in a unique listening pattern per user appended to the original survey data – our final dataset.

(Please note that especially this step is loaded with generalisations and assumptions, ideally in later steps it would be very beneficial to directly collect the individual listening behaviour.)

3. Analysis

3.1. EDA and pre-processing

3.1.1. Handling Outliers

We looked at the distribution of each of the variable and features and checked for the presence of potential outliers or potentially erroneous data. The combined dataset has 736 rows, each corresponding to a specific user. We first check for outliers among the numerical columns using the z-score method, i.e. check the distance for the mean, setting a threshold of 3.

Figure 1: outliers for Age, Hours per day and BPM



Based on the plots in figure 1 above, we notice that that there are several outliers in the data, chiefly "age", where individuals over the age of 70 can be considered outliers, listening time where some individuals listened to music more than 12 hours a day (This could be either a data error, or potentially someone having multiple sessions of streaming platforms open simultaneously, i.e. listening to Spotify on both phone and computer). Lastly, the BPM (Beats Per Minute for each song) variables contains some peculiar outliers as well, with three songs that have BPM of 999999999.0, which is unrealistic. To clean the dataset, we drop such BPM values that are over 300 BPM (Based on general knowledge of music, this is unlikely to be representative / possible). Similarly, we drop the users which are over 70 (given domain knowledge about Spotify / streaming platform adoption rates) and listened to more than 12 hours.

3.1.2. NaN values and imputations

Next-step is to identify missing values in the data and: We notice several NaN values among multiple categorical and numerical columns ("Age", "Primary streaming service", "Instrumentalist". "While working", "Composer". "Foreign languages", "BPM" and "Music effects"). We look at the distributions of such variables (Figure 2 below) to decide on the best course of action: As a general rule, we input the mean for normally-distributed numerical variables, the median for skewed numerical distributions and mode for categorical variables. The "Age" distribution is significantly skewed towards younger individuals, which is expected considering music streaming platforms are relatively recent and tend to be more widespread among younger generations. We therefore replace missing values with the median. Most of the categorical variables ("Primary streaming service", "While working", "Instrumentalist", "Composer"), have skewed distributions (e.g. the frequency of Spotify is much higher than any other streaming services among participants) thus we input their respective modes for NaN values.

Figure 2: Distribution of variables



For the remaining NaN in "Foreign languages", "BPM" and "Music effects", the imputation process is more complex as we expect these values to be related to one another (i.e. BPM should have an impact on the mental health of the individual) and inputting mode, media or

mean arbitrarily there could potentially lead to errors and misinterpretation. As result we chose to input missing values using a random forest iterative imputer, to avoid multicollinearity among those three variables.

3.1.3. Feature encoding

Multiple feature encoding techniques were employed to handle both the categorical as well as the ordinal data. The columns that were encoded are as follows:

Feature	Values	Encoding
Music effects	No effect, Improve, Worsen	Ordinal
Foreign languages	Yes, No	One-hot
Frequency (for each genre)	Never, Rarely, Sometimes, Very frequently	Ordinal
While working	Yes, No	One-hot
Instrumentalist	Yes, No	One-hot
Composer	Yes, No	One-hot
Exploratory	Yes, No	One-hot
Fav Genre	Country, EDM, Folk, Gospel, Latin, Lofi, Hip hop, Jazz, K pop, Metal, R&B, Rap, Rock, Video game music, Pop	One-hot

Figure 3: Encoding logic by feature

For each feature with "Yes / No" values, we used one-hot encoding to translate values to 1 and 0, respectively. We used ordinal encoding for listening frequency ('Never': 0, 'Rarely': 1, 'Sometimes': 2, 'Very frequently': 3) (already done earlier as part of the merging process in order to calculate the weighted average values per user). We also used ordinal encoding for Music effects as it has a hierarchy of importance. The Fav genre was one hot encoded to represent all the genres accurately.

3.2. Machine learning models

Using the combined dataset above, we will now attempt to train and test multiple models with **1. Mental health state as outcome:** Use our combined set of features to help predict each of the four potential mental health issues (depression, anxiety insomnia and OCD) as well as their degree of severity and **2. Music effect as outcome:** Implement a model that can effectively predict whether specific songs, genres and listening patterns can lead improve or worsen user's mental state.

3.2.1. Multiple Linear Regression (Base)

Given that our outcome is of multi-class nature (i.e. 4 mental health indicators), we start our analysis using a Multiple Linear regression model, which has the advantage of providing a visual aspect for interpretation as well as enabling us to identify the most important features along

with the added benefit of seeing if the relationship between the predictors and the target are not complex.

The model performance was extremely poor overall with all features included, resulting in an MSE score of 9.9 within the output range of the target in the range of 1-10, showing the model's inability to properly predict the target variables. The results are also shown in the graph below.

Figure 4: Multiple linear regression output



The graph has no meaningful patterns to decipher even with sorting. There could be several root causes for the poor performance: Given the large number of features included in the mode, some of them could be insignificant. Additionally, there could be multicollinearity especially among the many song metrics. This is also a strong indication that complex relationships exist between the predictors and the targets which make it difficult for simple models to interpret.

This leaves us with two improvement options: Improving the Regression model itself or changing towards a different model.

3.2.2. Regression – Feature Selection

Our first step is attempting to improve the regression model by first limiting the number of features to statistically significant ones only and then perform backward selection on these. To extract the significant features, we choose different approaches depending on the nature of feature and target.

Figure 5: Identification of significant features

	Numerical Target (Mental health states)	Categorical Target (Music Effects)
Numerical column	Pearson correlation matrix	ANOVA-Test
Categorical column	ANOVA - Test	Chi-Square-Test

Given the high degree of correlation between each of the target variables, we ensure to drop the remaining other targets from the set and conduct the further selection process for each target variable individually. To obtain the final feature sets we perform backward selection. As for the target "Depression" we encounter high multicollinearity, we calculate the Variance Inflation Factor to identify and remove the features that suffer most from multicollinearity. We therefore arrive at the following model inputs and run separate linear regressions for each mental health indicator and a logistic regression for music effects as target variable:

Tgt. variable	Features Selected	R-squared
Anxiety	Frequency [Folk], Age	0.044
Depression	Frequency [Metal], Frequency [K pop]_nor, Frequency [Country]_nor, weighted_valence, Age	0.063
OCD	Hours per day, Frequency [EDM], Frequency [Country], Age	0.04
Insomnia	Hours per day, Frequency [Metal], Frequency [Metal]_nor, Frequency [Country]	0.057
Music Effects	while working [T.Yes], Exploratory [T.Yes], Instrumentalist [T.Yes], Frequency [R&B]	(pseudo) 0.0533

Figure 6: Backward selection output

Overall performance across all metrics remains subpar, with a very low R-squared ranging from 4-6%, showing little explanatory power from the regression model.

This underlines the challenge that simple regression models are not able to correctly predict the targets, potentially stemming from not well-defined decision boundaries.

3.2.3. Support Vector machines

Given the results from the regression above, we continue our analysis by implementing several different models in order to identify a potential top performer: We first try the SVM model to gauge the decision boundaries and help us decide which ensembles to use. This model also performs poorly at 8.981 although slightly better than MLR.

Figure 7: SVM Output



The boundaries are not distinct, and we can see severe overlap in the target columns. This situation is best for treebased ensembles that can learn these complex relationships and perform better on smaller datasets than Deep Neural Networks.

3.2.4. Ensembles

Given the constraints and poor performance associated with the simple models above, we move to tree ensemble models. To start, we use AdaBoost, XGBoost and Random Forest classifiers with the default parameters on the original dataset.

Figure 8: Ensemble model results (examplatory for Depression)

	Accu	racy	F1-Score		
	Train	Test	Train	Test	
AdaBoost	0.191033	0.122172	0.193504	0.1188016	
XGBoost	1.000000	0.117647	1.000000	0.1011043	
RandomForest	1.000000	0.171946	1.000000	0.1598497	

We can see severe overfitting on XGBoost and RandomForest and underfitting on AdaBoost with very poor performance on the test dataset across.

The main reason for this could be the scale of the target columns (0 to 10) which makes it difficult for the model to predict as there are potentially too many categories for too little data. Thus, our next course of action is to scale the target columns down.

3.2.5. Ensembles on the scaled down target columns

For better model performance (and readability) we scale down the target columns for mental health states to a scale of 0 to 4, signifying fine / low presence/ medium / high.

3.2.6. Fine tuning with Optuna

To further optimize to reduce overfitting, we use optima and then define the objective function as the crossvalidation score and aim to maximise this value over 100 studies.

The increase in AUROC gives us a new value of 0.55, which is better but still unsatisfactory. The best hyperparameters we found are as follows:

- n_estimators = 100
- $max_depth = 5$
- learning_rate = 0.01406235974348568
- subsample = 0.6041048627531311
- $max_{features} = 0.7582181029284326$
- min_samples_split = 9
- $min_samples_leaf = 4$

3.2.7. Predicting Mental health

Figure 9: Model Performance on Mental Health indicators (*Anxiety, Depression, Insomnia, OCD*)

	Accuracy		F1-Score		AUROC	
	Train	Test	Train	Test	Train	Test
AdaBoost	0.4327	0.3303	0.3696	0.2516	0.7434	0.4675
XGBoost	1.0	0.3303	1.0	0.2508	1.0	0.5119
RandomForest	1.0	0.3891	1.0	0.2696	1.0	0.5393
LGBM	1.0	0.3394	1.0	0.2547	1.0	0.5405
CatBoost	1.0	0.3891	1.0	0.2687	1.0	0.5334
Extra Trees	1.0	0.3756	1.0	0.2662	1.0	0.5470
GBM	0.9669	0.3484	0.9683	0.2508	0.9979	0.5229

	Accuracy		F1-Score		AUROC	
	Train	Test	Train	Test	Train	Test
AdaBoost	0.4639	0.3167	0.4250	0.2539	0.7462	0.5691
XGBoost	1.0	0.3575	1.0	0.3096	1.0	0.6066
RandomForest	1.0	0.3891	1.0	0.2837	1.0	0.5680
LGBM	1.0	0.3937	1.0	0.36	1.0	0.6034
CatBoost	1.0	0.3891	1.0	0.3189	1.0	0.5761
Extra Trees	1.0	0.3665	1.0	0.2697	1.0	0.6072
GBM	0.9454	0.3982	0.9526	0.3408	0.9966	0.6158

	Accuracy		F1-S	Score	AUROC	
	Train	Test	Train	Test	Train	Test
AdaBoost	0.5087	0.4706	0.3888	0.2932	0.7597	0.5446
XGBoost	1.0	0.4615	1.0	0.2412	1.0	0.5165
RandomForest	1.0	0.5656	1.0	0.2281	1.0	0.5324
LGBM	1.0	0.4706	1.0	0.2381	1.0	0.5486
CatBoost	1.0	0.5113	1.0	0.2351	1.0	0.5079

Extra Trees	1.0	0.5294	1.0	0.2289	1.0	0.4942
GBM	0.9513	0.4706	0.9591	0.242	0.9974	0.5546

			F1- Score AUROC			
Train						T e ^{Fra} Tes Frai Fes s t n t t
AdaBoost	0.64 72	0.59 28	0.31 73	0.24 14	0.76 9	0.5378
XGBoost	1.0	0.64 71	1.0	0.25 45	1.0	0.5108
RandomFo rest	1.0	0.66 06	1.0	0.20 97	1.0	0.4993
LGBM	1.0	0.63 35	1.0	0.24 77	1.0	0.5307
CatBoost	1.0	0.64 71	1.0	0.22 44	1.0	0.5314
Extra Trees	1.0	0.63 35	1.0	0.20 47	1.0	0.5024
GBM	0.96 1	0.60 63	0.95 12	0.22 87	0.99 99	0.5232

XGB, RF, LGBM, CatBoost, ExtraTrees and GBM suffer from overfitting and Adaboost is underfit. However, there is an increase in the model's performance. The best one – GBM - has an average of 0.55 test AUROC across all the target columns with the highest for depression at 0.62.

3.2.8. Predicting Mood improvements

We also try to predict the mood improvements in the individuals by now using the "music effects" column as the target. The AdaBoost, XGBoost, RandomForest, LGBM, CatBoost, ExtraTrees and GBM classifiers were used. This too doesn't give us promising results.

Figure 10: Model Performance on Music Effect

	Accu	iracy	F1-8	Score	AUI	ROC
	Train	Test	Train	Test	Train	Test
AdaBoost	0.7466	0.7149	0.2138	0.2084	0.7146	0.474
XGBoost	1.0	0.7014	1.0	0.266	1.0	0.5853
RandomForest	1.0	0.7059	1.0	0.2146	1.0	0.5566
LGBM	1.0	0.7059	1.0	0.2713	1.0	0.637
CatBoost	1.0	0.724	1.0	0.2583	1.0	0.5472
Extra Trees	1.0	0.7285	1.0	0.2646	1.0	0.5493
GBM	0.9844	0.7149	0.9879	0.2692	0.99	0.6549

3.3. Findings and insights

Overall, we are unable to identify a model than can generalise the relationship between listening behaviours and mental health effectively: Simple regression models registered very high MSE and showed little explanatory power across all mental health indicators. Tree ensembles models yielded better results, but the performance remained poor overall, as all models suffered from overfitting to a varying extent, with low generalization ability on the test set across the board. The resulting AUROC scores revolve around the 50% range, showing that the model performance is not better than a random choice.

Despite the results above, we were able to identify the Gradient Boosting Classifier as the best performer for our dataset. The is likely due its inherent ability to capture complex patterns through sequential learning involving multiple weak learners, enabling the model to build on the errors made by previous learners and generating step by step improvements. Gradient Boosting can also handle feature importance effectively, which is advantageous in our case given the large number of features in the dataset. Should such an analysis be done again on a larger sample, the Gradient Boosting Classifier would likely perform better again given its inherent advantages that are very relevant in such a case of multiple interconnected features.

There are several potential root causes for the poor performance: First, we could highlight the representativeness issue as our survey is both relatively small (734 rows) and volunteer-based - This can reduce the ability of the model to capture interconnectivity and detail. Furthermore, mental health issues are notoriously difficult to diagnose and detect - The mental health data in the survey is self-reported and could thus differ from an official diagnosis provided by a certified psychiatrist. The level of information on each individual user is also limited and further data such as gender, salary, status, etc. could potentially help the model associate specific profile to listening patterns and draw better conclusions. In addition, we had to make some assumptions regarding the song selection for each user and assume their listening history would closely resemble the top playlist of each genre, which is not necessarily true as individual users could have different preferences and playlists within a specific genre. With all these combined issues in mind, we can argue that our sample survey is not fully representative of the complex relationship between mental health status and the listening habits predictors, leading to the observed poor performance.

Beyond the performance on the test subset, we also wanted to further test our model on real world data, to assess its ability to predict mental health in a different setting. Given the lack of similar datasets, with audio features and mental health information, we decided to create country-wide user summary based on top 50 country playlists readily available through Spotify as well as publicly available mental health information from the Word Happiness Index and official depression rates. We chose Singapore and Indonesia for this analysis, given their close geographic proximity but widely different top 50 playlist and mental health statistics: Singapore is #30 on the World Happiness Index, with 49% of Singaporeans reporting to have suffered from depression at least once. Indonesia on the other hand is #80 in Happiness Index but its depression rate is very low at a mere 6.1%.

We create aggregate song metrics using the top 50 playlists (extracted via Spotify API), using the same pre-processing as the survey and create average user information using Spotify and Statista to input average age, streaming hours, favourite genres, etc. as follows:

Country	Singapore	Indonesia
Age	27.0585	24.2145
Hours per day	2.533333	2.183333
While working	No	No
Instrumentalist	Yes	Yes
Composer	No	No
Fav genre	EDM	Рор
Exploratory	No	Yes
Foreign languages	Yes	Yes

Figure 11: User profile for Singapore and Indonesia

We then applied the fine-tuned Gradient Boosting Classifier to both countries yielding the following results:

Figure 12: Predictions for Singapore and Indonesia

Results	Singapore	Indonesia
Depression	2	1
Anxiety	1	0
OCD	0	0
Insomnia	0	0
feel better?	0	0

Interestingly, the model predicts a relatively higher depression score for Singapore despite the limitations discussed above, in line with the aforementioned official depression rates. However, it fails to identify a potential higher "feel better?" score in Singapore linked to the higher happiness index, possibly due to significant level of aggregation.

3.4. Conclusion and expansion areas

Conclusively, while the existing research does highlight a potentially strong link between mental well-being and music, translating such a link into a model has proven very difficult in the context of our survey, with both regression and ensembles models being unable to yield concrete results. Using such a model in a sensitive medical context would likely prove very dangerous for Spotify and its users, given the high risk behind false positives/negatives. As a result, subsequent research should focus on strongly improving model performance - This can be done through Representativeness improvements, with further surveys and studies, to create a larger overall data sample. Collecting data over time in a time series format could also prove very useful, as it would help to refine the relationship and variations over time between mental health and listening behaviours, highlighting causal patterns instead of pure correlations. Such larger data samples would also enable us to expand the model selection, leveraging Deep Neural Networks for additional perspective and insights. Once a high performing model is identified, we could further refine the analysis with explainable AI, to provide additional transparency on the decision process in a highly sensitive medical context.

5. References

GitHub: <u>https://github.com/papafffranku/5153-group-project</u>

5.1. Literature

- Burns JL, Labbé E, Arke B, Capeless K, Cooksey B, Steadman A, Gonzales C. The effects of different types of music on perceived and physiological measures of stress. J Music Ther. 2002 Summer;39(2):101-16. doi: 10.1093/jmt/39.2.101. PMID: 12213081.
- Dattani, S., Rodés-Guirao, L., Ritchie, H., & Roser, M. (2023). Mental Health. Our World in Data. Retrieved from <u>https://ourworldindata.org/mental-health</u>
- Farmer, C., Farrand, P. & O'Mahen, H. 'I am not a depressed person': How identity conflict affects help-seeking rates for major depressive disorder. BMC Psychiatry 12, 164 (2012). <u>https://doi.org/10.1186/1471-244X-12-164</u>
- Greenberg PE, Kessler RC, Birnbaum HG, Leong SA, Lowe SW, Berglund PA, Corey-Lisle PK. The economic burden of depression in the United States: how did it change between 1990 and 2000? J Clin Psychiatry. 2003 Dec;64(12):1465-75. doi: 10.4088/jcp.v64n1211. PMID: 14728109.
- Gustavson, D.E., Coleman, P.L., Iversen, J.R. et al. Mental health and music engagement: review, framework, and guidelines for future studies. Transl Psychiatry 11, 370 (2021). <u>https://doi.org/10.1038/s41398-021-01483-8</u>
- Labbé E, Schmidt N, Babin J, Pharr M. Coping with stress: the effectiveness of different types of music. Appl Psychophysiol Biofeedback. 2007 Dec;32(3-4):163-8. doi: 10.1007/s10484-007-9043-9. Epub 2007 Oct 27. PMID: 17965934.
- McCraty R, Barrios-Choplin B, Atkinson M, Tomasino D. The effects of different types of music on mood, tension, and mental clarity. Altern Ther Health Med. 1998 Jan;4(1):75-84. PMID: 9439023.
- NHS. (2021, November 4). Antidepressants Overview. Retrieved from <u>https://www.nhs.uk/mental-health/talking-therapies-medicine-treatments/medicines-and-psychiatry/antidepressants/overview/</u>
- Nochaiwong, S., Ruengorn, C., Thavorn, K. et al. Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: a systematic review and meta-analysis. Sci Rep 11, 10173 (2021). <u>https://doi.org/10.1038/s41598-021-89700-8</u>
- Schaeffer, K. (2022, April 25). In CDC survey, 37% of U.S. high school students report regular mental health struggles during COVID-19. Pew Research Center. Retrieved from https://www.pewresearch.org/short-

reads/2022/04/25/in-cdc-survey-37-of-u-s-high-schoolstudents-report-regular-mental-health-struggles-duringcovid-19/

5.2. Datasets

1. Kaggle Mental Health Survey:

https://www.kaggle.com/datasets/catherinerasgaitis/mxmh -survey-results

- 2. Spotify API: https://spotipy.readthedocs.io/en/2.22.1/
- 3. Azapi GitHub: https://github.com/elmoiv/azapi
- 4. Kaggle NLP training dataset:

https://www.kaggle.com/datasets/pashupatigupta/emotion -detection-from-text

5. Kaggle list of commonly found contractions: https://www.kaggle.com/datasets/ishivinal/contractions

6. Spotify playlists:

Classical:

https://open.spotify.com/playlist/37i9dQZF1DWWEJIAG A9gs0?si=e7d3dac97915451c

Country:

https://open.spotify.com/playlist/37i9dQZF1DX7aUUBC Kwo4Y?si=5635901a93084f14

EDM:

https://open.spotify.com/playlist/37i9dQZF1DX3Kdv0IC hEm9?si=8e964c6ffd4f42f4

Folk:

https://open.spotify.com/playlist/37i9dQZF1DWVmps5U 8gHNv?si=18bce18f2da44226

Gospel:

'https://open.spotify.com/playlist/37i9dQZF1DX7OIddoQ VdRt?si=2bed0b648ba44069

Hip hop:

https://open.spotify.com/playlist/37i9dQZF1DXbkfWVL d8wE3?si=afcdd0d007fa46c3

Jazz:

https://open.spotify.com/playlist/37i9dQZF1DXbITWG1 ZJKYt?si=guFyActQQw2vTGMobg6YSw&pi=aaPfVZWhHRmKW

K pop:

https://open.spotify.com/playlist/37i9dQZF1DWTTXfIhc SkXG?si=1d8413b68211490f Latin:

https://open.spotify.com/playlist/37i9dQZF1DX6ThddIj WuGT?si=ewnDCCmwQCmUuuW-OtcFQQ&pi=a-X8qfK7viSlOo

Lofi:

https://open.spotify.com/playlist/37i9dQZF1DX4t95PAs1 EpY?si=kDgr4_IoRBeOioE6RM4arA&pi=a-0nKUKT8bSHmJ

Metal:

'https://open.spotify.com/playlist/37i9dQZF1DWWOaP4 H0w5b0?si=Izrft9RVRT2t8X595Ud8OA&pi=a-6cDsTZS3SJm8

Pop:

https://open.spotify.com/playlist/37i9dQZF1DX2vTOtsQ 5Isl?si=a77c2bcb0ef24418

R&B:

https://open.spotify.com/playlist/37i9dQZF1DX4SBhb3fq CJd?si=eB_quwSsSbWxD4bUbE0PKw&pi=aweeDuYyNSOyM

Rap:

https://open.spotify.com/playlist/2shZ4OmoeOmtAabLw HUiy7?si=W6gE09OZR7eYxPgbAW7cnw&pi=a-6DtXvfgxQzmY

Rock: https://open.spotify.com/playlist/37i9dQZF1DWXRqgorJ j26U?si=w-IR2ajvRzmbmbMZmOAswA&pi=a-FoFG5A14Q6Og

Video game music: https://open.spotify.com/playlist/4Pf4e2VIW38Lg5jNUnp Wvg?si=0d93a6af0f954bc5

5.3. Models

T5: https://huggingface.co/docs/transformers/model_doc/t5

Github link