Identifying the relation of music listening habits to mental health

I) Introduction

Music is frequently used as a form of emotional regulation, with many individuals engaging in solitary listening to manage their emotions. Clinical music therapists have long recognized the influence of music on emotional states and have integrated it into therapeutic interventions for alleviating symptoms of psychiatric mood disorders, such as depression. A study conducted by researchers at the Centre for Interdisciplinary Music Research (University of Jyväskylä), Aalto University, and Aarhus University investigated the relationship between mental health, music listening habits, and neural responses to music-evoked emotions. Their findings indicated that individuals who predominantly listened to sad or aggressive music to express negative emotions exhibited higher levels of anxiety and neuroticism. In this project, we aim to further explore the association between music listening habits and self-reported mental health concerns, including anxiety and depression, to assess the potential of these habits as indicators for identifying mental health states.

The report is divided into 5 sections: The current one is section 1, which explains the background of our ML application and project structure. Section 2 discusses problem formulation and the dataset. Section 3 discusses data processing, feature selection, model selection, loss function and model validation. In section 4 we discuss and compare the results. Section 5 summaries the findings and discusses how the project can be improved and the limitations.

II) Problem formulation

In this project we attempt to predict a person's mental health states through their music listening habits and relating factors. We will be using an open source dataset from Kaggle. The data has been gathered via public surveys posted to different social media platforms such as Reddit and Discord as well as links to the survey in public locations. The Age, Hours per day, BPM, Anxiety, OCD, Depression, Insomnia columns have continuous variables while the rest are categorical. To be more specific on the categorical columns, While working, Exploratory, Composer, Foreign Language, Instrumentalist have ["Yes", "No"] values. The 16 columns for how frequently they listen to the 16 genres have ["Never", "Rarely", "Sometimes", "Very frequently"] values.

III) Methods

a. Preparing the data

The dataset columns can be divided into 4 types, each of them representing a different aspect of the respondent in regards to their relationship with music:

- Relating to music listening habits: Hours per day, While working, Fav genre, Exploratory, Foreign Languages, BPM and how frequently they listen to the 16 genres
- Relating to musical background: Whether they are an instrumentalist and/or composer.
- Self reported mental state: OCD, Depression, Anxiety and Insomnia (which is ranked from on a scale of 0 to 10 by the respondent)
- Others: Respondent's age and Timestamp (of when the respondent filled in the survey)

The whole dataset has 737 rows of data. Which means that 737 people filled in the survey. In total each datapoint has 33 columns of data. Some rows have missing values so we omitted them all using dropna(). This leaves us with 623 working data points. To make data processing easier, we

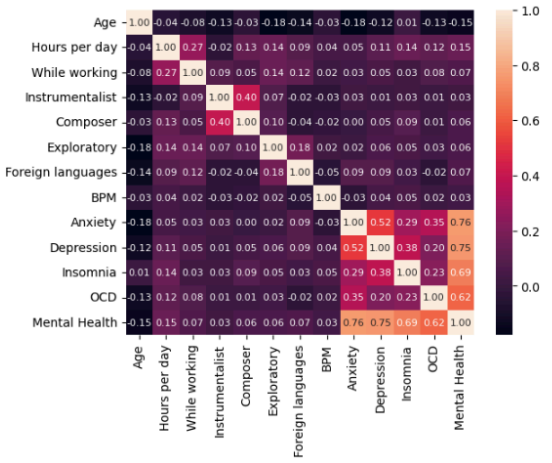
transform the columns that have “Yes” - “No” strings as their values into binary values. With “Yes” as 1 and “No” as zero. Now these columns can be taken into account when creating our ML models.

b. Labels

This is a supervised machine learning project. We decided to take the self reported mental state columns, which are Depression, Anxiety, Insomnia and OCD as our labels. This means that we will use each of them as a label for our model to see how accurately we can use music listening habits to predict the state of different mental health conditions. Different mental conditions may display different relationships with music listening habits. The respondents rate them from 0 to 10 in integer increments of 1, so strictly speaking, the label values are ordinal, but we will treat them as if they’re continuous.

c. Feature selection

We decide to omit the 16 columns that refer to how frequently the respondent listens to the 16 genres, Timestamp and Primary streaming service because of their lack of relevance and to prevent overfitting. That leaves us with 11 possible features to choose from. To reduce noise and select relevant features, we created a heatmap from possible features with numerical values to check for the correlation.



We decide to choose the features with a correlation above 0.03 for at least 2 mental health conditions. According to the study mentioned in the introduction paragraph, individuals who predominantly listened to sad or aggressive music to express negative emotions exhibited higher levels of anxiety and neuroticism, which suggest that genre is also an important factor. Fav genre is a categorical feature, so we used get_dummies() to convert it into indicator variables.
We did a lot of transformation and processing with the data so here’s a comprehensive table:

Features	Description	Data type
Hours per day	How many hours per day do they listen to music?	Continuous
While working	Do they listen to music while working or not?	Binary
Composer	Are they composers?	Binary
BPM	Beats per minute of favourite genre	Continuous
Foreign languages	Do they often listen to music in a language they’re not good at?	Binary
Age	How old are they?	Continuous
Exploratory	Does the respondent actively explore new artists/genres?	Binary
Fav genre_[genre]	Is [genre] their favourite?	Binary

d. Model validation

The standard split of training data and testing data in sklearn’s train_test_split function is 0.75/0.25. We felt that this was a suitable split for our data as well, since various sources on the internet suggest splits from 70/30 to 80/20. Using a single split can maximise the limited amount of data available for training. Larger datasets could opt with even tighter splits, but since ours isn’t that comprehensive, we chose the 0.75/0.25. We decided not to use k-fold cross-validation because it may lead to high variance in performance estimates due to the reduced number of training samples in each fold.

e. Models

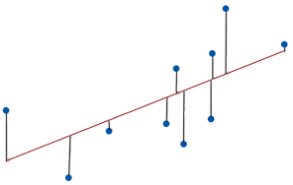
Linear Regression: This is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features) to find the best-fitting linear equation that predicts the target variable as a function of the independent variables. In the case of this project with multiple features, the model attempts to find the best-fitting hyperplane that minimises the difference between the actual values of the target variable and the predicted values. Therefore the model should work well when finding a relationship between various factors of a person's music listening habits, as the features (which has been elaborated above) and their mental health state. It is a straightforward and interpretable model, making it easy to understand the relationship between the features and the label. Moreover, since our label values are continuous, linear regression is well-suited to predict real-valued outputs.

Lasso Regression: Lasso regression (least absolute shrinkage and selection operator) is similar to linear regression, but differs from it by implementing variable selection and regularisation. We felt that lasso regression would be another good machine learning model to use for our task, since the regularisation would help reduce overfitting in addition to our dataset being quite high dimensioned, which can cause issues in some other models.

Lasso regression is a good fit for our task since by selecting a good alpha value, we can control how smaller, possibly less important values affect our result. An additional factor that pushed us towards lasso regression was the premade methods for it in the sklearn library.

f. Loss Function

Mean squared error: Linear Regression typically uses MSE as the loss function during training. With the formula $\frac{1}{n}\Sigma(y_i - \widehat{y_i})^2$, the function assesses the average squared difference between the observed and predicted values. As the data fall closer to the regression line the lower the error is, which means MSE decreases. MSE penalises large errors between predicted data and actual more heavily, emphasising the minimization of these significant errors. The goal of Linear Regression during training is to find the values of the parameters that minimise the MSE.



L1 norm: Lasso regression uses the L1 norm as its loss function in addition to a regularisation factor. It seeks to optimise the function:

$\frac{1}{2\,(2 * n\_samples))} * ||y - Xw_2||^2 + \alpha * ||w_1||$, where α is variable in [0, ∞ that controls the strength of the regularisation applied. The higher the value, the more the parameters are smoothed, meaning that the only parameters with more important weights are considered. This helps avoid the use of unwanted or unrelated variables within the model. The coefficient can also reduce values to zero unlike in some other models.

IV) Results

For linear regression, it yields the following error for each label in each method in the following tables:

Training errors for linear regression:

	MSE	R^2
Depression	6.907122	0.02968048702725723
Anxiety	7.540244	-0.01821227684605642
OCD	8.192183	-0.01812589730696712
Insomnia	10.33683	-0.0679682132633399

Training errors for lasso regression:

	MSE	R^2
Depression	6.965335	0.021502
Anxiety	7.156269	0.033638
OCD	7.824458	0.027574
Insomnia	9.657900	0.002177

It seems that for Depression and Anxiety, the model yields lower error than those of OCD and Insomnia, suggesting a higher correlation between music listening habits and Depression and Anxiety than OCD and Insomnia. This falls in line with what the research mentioned in the introduction paragraph suggested.

We used mean squared error and the R^2 score to compare our results from each model. Mean squared error is a common way of measuring the accuracy of a regression model, which is why we chose it as one of the quality measures.. The R^2 score provides another similar score to compare, though it doesn't provide much new information besides being able to compare the fit between unrelated models as well. After trial and error we settled for an alpha value of 0.09 for the lasso regression method. When using values less than 0.8 or so, the mean squared error would rise to the thousands.

Based on these values, despite our estimates lasso regression provided a better fit for our dataset. The scores for mse are comparatively high when compared to similar methods used on other datasets. This indicates that either our method is unsuitable for our dataset or there is little or no real correlation between the variables in our data. This will be discussed further in the results section.

V) Conclusion

According to the result section we see that lasso regression is the slightly better performing model. Although looking at the high error value, it's safe to say that it doesn't really matter which model we choose because at this point we can conclude that just music listening habit alone isn't sufficient enough to evaluate a person's state of mental health. Thinking intuitively we can easily understand why this is the case. Mental health evaluation is known to be a multi faceted issue, affected by a multitude of physical, social, cultural and environmental issues over an extended period of time. Music just plays a minor role in revealing a person's state of mind, this is also indicated in the correlation matrix above, with the highest correlation being just 0,15. Moreover, we are training both of our models on a fairly small data set, with just 623 valid data points, so it is expected that the models won't yield high accuracy predictions. Mental health professionals should not rely on music listening habits as a trustworthy metre for patients, but it is a possibility that they can use patients music listening habits to further probe their patient's mental state after they have done extensive analysis of other more significant aspects such as the patient's lifestyle and relationships.

In the future, both of the models can definitely benefit from collecting more training data. One other possible ML model we can use when there's a bigger dataset is Random Forest Regression because it combines multiple decision trees, which improves accuracy and reduces overfitting, especially when there's a lot of data to learn from diverse patterns. More research should be done on effects on music on the brain to enable higher rate of relevance in information of publicly collected data, which would help produce more appropriate features for ML models in other similar topics to this.

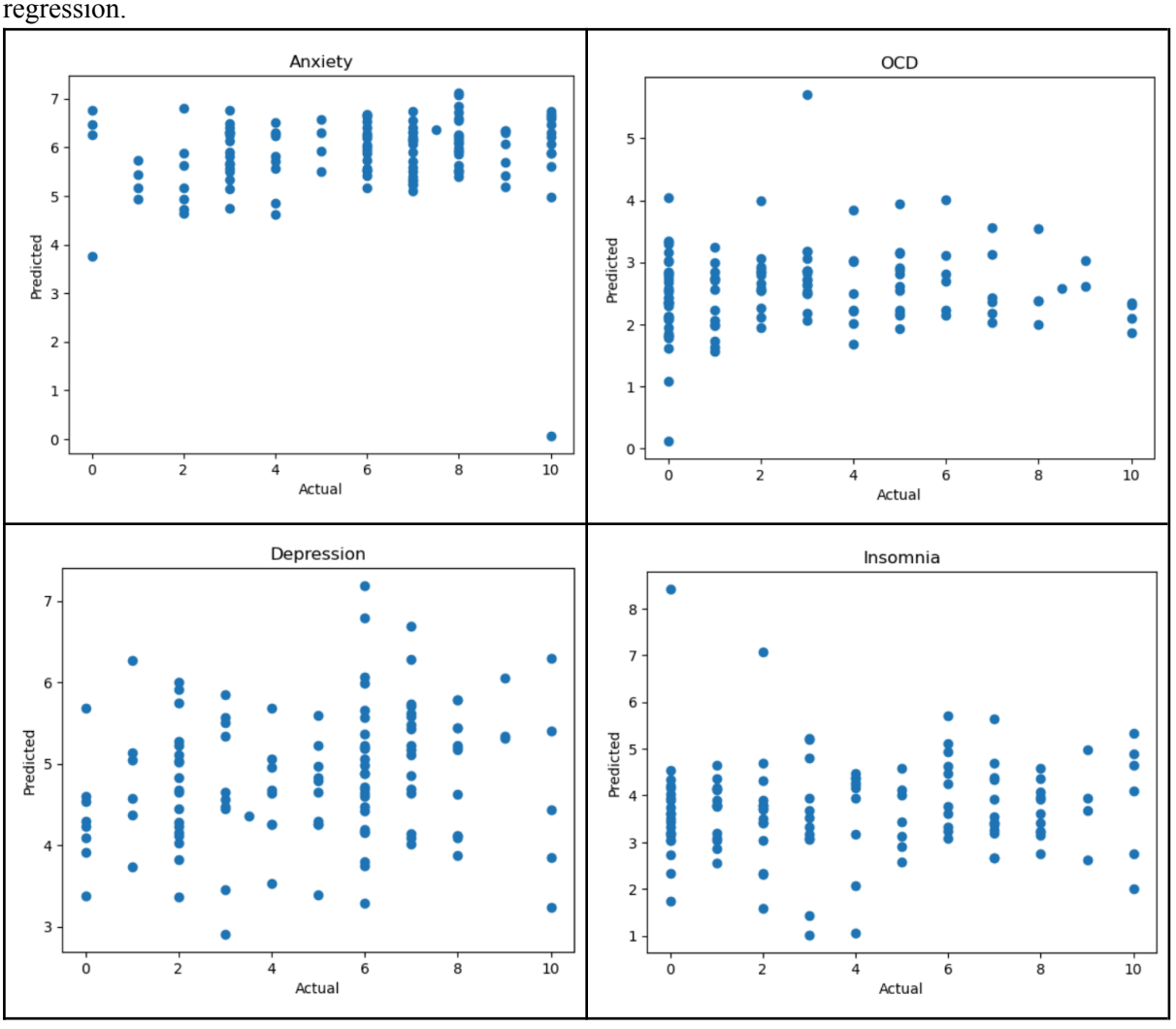
VI) Bibliography

Sources:

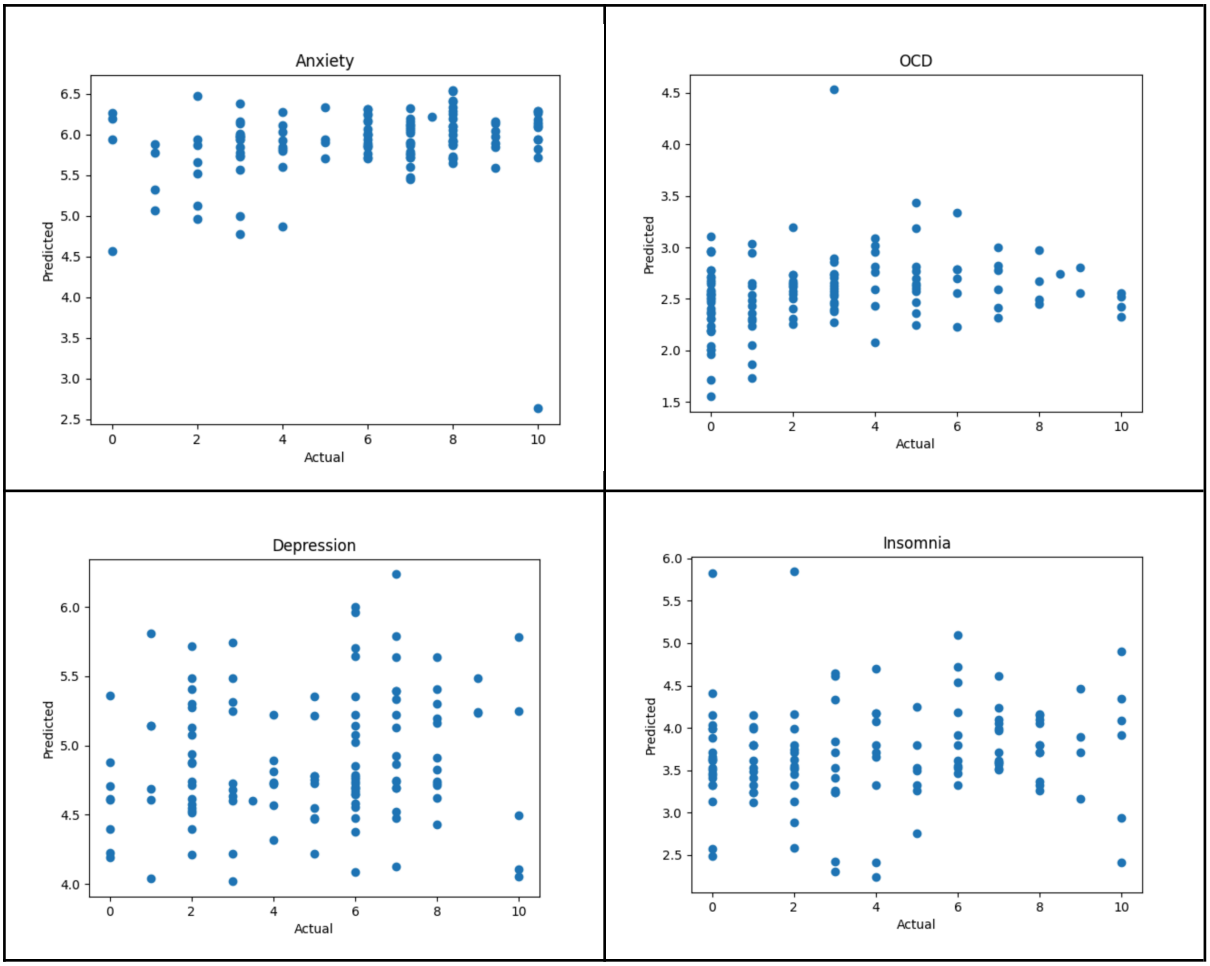
Dataset: https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results/data
Code: https://github.com/RedknanRonin/ml-24
Lasso regression: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.Lasso.html
https://www.ibm.com/topics/lasso-regression
Linear regression: https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.LinearRegression.html

VII) Appendix

Charts depicting the relation of predicted mental health scores and the actual scores from linear regression.



Charts depicting the relation of predicted mental health scores and the actual scores from lasso regression.


Anxiety
Predicted
Actual
OCD
Predicted
Actual
Depression
Predicted
Actual
Insomnia
Predicted
Actual