Final Report

Introduction

Music plays a crucial role in emotional expression and stress management. When learning about mental health, it's valuable to explore how music habits relate to well-being. This project analyzes survey data on music preferences and mental health indicators to identify patterns between listening behaviors and emotional states. Our motivation is to explore these relationships and determine how they may offer insight into the impact of music on mood and mental well-being, or vice versa. In this research project, we aim to answer the following question: *Is it possible to successfully predict a self-reported anxiety score based on an individual's age and music preferences?*

Methods

Our dataset is sourced from Kaggle, and data collection on music taste and self-reported mental health was conducted through public surveys using a Google Form distributed online across various social media platforms. The form was also advertised using posters and business cards in public locations such as libraries and parks. It contained questions about the primary music streaming service, hours listened per day, age, BPM (of favorite genre), favorite genre, listening frequency of certain genres, and self-reported mental health scores for various mental health disorders. To ensure it was ready for use in our models, we had to implement preliminary cleaning procedures. First, we dropped irrelevant columns: "timestamp," "permissions," "depression," "insomnia," and "OCD." Because our research question focuses only on anxiety, we did not need the other self-reported mental health columns, and the timestamp and permissions columns were not relevant to our research question. Next, we remapped responses for the music listening frequency questions as follows: "Never" = 1, "Rarely" = 1, "Sometimes" = 2, and "Very frequently" = 3, and we encoded the dataset using these values. We also replaced the boolean columns ("True" and "False") with 1 and 0, respectively. Finally, we manually cleaned the BPM column because the maximum reported value was 999999999.0, which is unrealistic. After conducting some research, we capped the BPM value at 250 so that unreasonable outliers would not affect model training or performance.

First, we trained the support vector machine and penalized linear models. To find the best penalized linear model, we tested regression with the following penalties: Ridge, Lasso, and ElasticNet. For each, we tried 100 parameters (alpha for Ridge and Lasso, and alpha and L1 ratio for ElasticNet) spaced logarithmically from 10^-4 to 10^4. We then used grid search and cross-validation to evaluate each model's performance on the test set. For SVM, we created a hyperparameter grid for tuning, testing different kernels, gamma types (for RBF), and values of C and epsilon. After running a grid search with cross-validation, we evaluated the best parameters on the training data. We also trained a Random Forest regressor to represent the ensemble portion of our approach. The model was implemented within a pipeline to ensure that preprocessing was applied identically across all folds of cross-validation. We used grid search and cross-validation to tune the hyperparameters, relying on mean squared error as the selection measure. This process provided an unbiased estimate of model performance and allowed us to identify the hyperparameter combination that generalized best. For the neural network model, we implemented a regression model using PyTorch. To preprocess the data, we used a scikit-learn pipeline that included standardization of features before applying the neural network. To evaluate and tune the model, we used 5-fold cross-validation with grid search over hyperparameters such as the number of hidden dimensions, learning rate, batch size, and number of epochs. The performance of all models was assessed using Mean Squared Error and Root Mean Squared Error from cross-validation to measure prediction error, and R-squared to see how much variance is explained by the model.

Results

	MSE	RMSE	R-Squared
Ridge Regression	7.64	2.71	0.05
Support Vector Machine	7.81	2.79	0.03
Gradient Boosting	7.51	2.74	0.02
Random Forest	7.50	2.73	0.03
Neural Network	8.81	2.97	-0.14

After evaluating the four models using 5-fold cross-validation and hyperparameter tuning, we determined that each model performed somewhat similarly based on MSE, RMSE, and R-squared. However, we observed that the Random Forest model outperformed the rest slightly. On the cross validation set, we observed that it had an R-Squared value of 0.03, and lower MSE of 7.50. Ultimately, this indicates that the other linear models were not able to capture many patterns in the data, but the Random Forest model was, likely due to its ability to capture non-linear relationships. After selecting this as the best model, we implemented the optimized Random Forest model on the test set. The performance showed that although the Random Forest model fit the training data decently, (Training R-Squared=0.666), very little of that learned structure transferred to unseen data - as the R-Squared on the test set was 0.017, indicating that the model is only slightly better than just predicting the mean. The increase in MAE and RMSE from training to testing, indicated that the model captured patterns specific to the training set. This suggests that the available features contain limited predictive information about anxiety levels and that much of the variance in anxiety is likely explained by factors not represented in the dataset. Despite its limited generalization, the Random Forest model displayed the highest performance across the models tested, providing an improvement compared to the linear and neural approaches.

Discussion

The results suggest that a mix of music preferences, listening habits, and brief demographic information contains some predictive signal for self-reported anxiety, but the strength of that signal is modest. Most of our models scored in a similar range for RMSE, having values around 2-3. This indicates that for self-reported anxiety levels on a scale of 1-10, our models are typically off by about 2-3 points. But, some models like the ensemble model performed slightly better than others, likely due to its ability to capture nonlinear relationships and complexity that linear models can't capture, and is more robust than neural networks on the medium-sized tabular survey data.

Some limitations of our analysis include the self-reported nature of the data, which may introduce biases, as anxiety levels are a subjective measure and can vary depending on mood, interpretation, and willingness to disclose personal information. Respondents might not accurately recall their music listening habits, leading to measurement error. Additionally, the dataset is not representative of the general population as it is skewed towards younger ages and heavier music listeners, limiting the generalizability of our results. Scaling numeric features reduces the distortion from different scales, but future work could explore larger datasets, additional features, or consider information from professionally administered tools to have a more objective measure of anxiety.