International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Math and Music A Deep Analysis on Music Habits and Mental Health Disorders

Nivaan Kothari

Student, International Baccalaureate, Aditya Birla World Academy

Abstract

This study explores the relationship between music-listening habits and mental health disorders such as anxiety, depression, insomnia, and OCD using a data-driven approach. Analyzing a diverse, self-reported dataset, using machine learning models—including linear regression, Random Forests, and neural networks—were applied to predict mental health outcomes. While advanced models captured some nonlinear patterns, overall predictive power remained modest, indicating that music habits alone cannot fully explain mental health variations. Challenges such as sample bias, limited variables, and the self-reported nature of data were noted. The findings highlight the multifactorial nature of mental health and suggest that future research should incorporate broader lifestyle and environmental factors alongside refined modeling techniques to better understand these complex interactions.

Keywords: Music Habits, Mental Health Disorders, Machine Learning Analysis

INTRODUCTION

Music & Mental Health Survey Results is a dataset found on Kaggle that explores the relationship between different factors of music variables and how they affect the listeners mental health. This dataset has self-reported information from a large sample space with diverse backgrounds; covering demographics, like age and gender; frequencies and genres of music heard, like rock, pop etc; and mental health indicators, like stress, anxiety, depression etc. This research paper aims to utilize this data to model a direct relationship between music habits and mental health, using machine learning, regression models, calculus and much more. Respondents rank Anxiety, Depression, Insomnia, and OCD on a scale of 0 to 10, where: 0 is I do not experience this and 10 is I experience this regularly, constantly/or to an extreme.

Background Information

The R^2 value, also known as the coefficient of determination, is an important statistic used to measure the goodness of fit of a regression model to the given dataset. The measure is the proportion of variance in the dependent variable explained by the independent variable(s) included in the model. The R^2 value is bounded between 0 and 1, and an R^2 value of 1 means that the model explains all the variation in the dataset perfectly, while an R^2 value of 0 means that the model does not explain any of the variation. For example, an R^2 value of 0.75 means that the model explains 75% of the variation in the target variable and the remaining 25% due to unexplained factors or randomness. In certain situations, R^2 can even be negative if the predictive power of the model is worse than using the mean value of the data for predictions.
The formula of R^2 value is:

$R^2 = 1 - \frac{RSS}{TSS}$,
where:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$TSS = \sum_{i=1}^{n}(y_i - \underline{y})^2$$

- y_i is the observed value
- $\underline{y}$ is the mean of the observed values
- $\hat{y}_i$ is the predicted value
- n is the number of observations

RSS is the residual sum of squares and TSS is the total sum of squares.
Yet another essential regression analysis measure is Mean Squared Error (MSE). MSE calculates the mean of the observed and predicted values' squared differences, as estimated by the model. The lesser MSE, the nearer the prediction values are to the true values. Larger errors would be associated with larger MSE. MSE does not give a proportion like R^2 but an absolute value of error in units of the dependent variable (squared). R^2 and MSE can be utilized together to test for model goodness: R^2 explains how well variation is captured by the model, while MSE explains how accurately predictions are.
The formula of MSE is:
$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$,
where:
1. $\hat{Y}_i$ is the predicted value
2. n is the number of observations
3. Y_i is the observed value

MSE is also equal to RSS multiplied by $\frac{1}{n}$.
The Kernel Density Estimate (KDE) is a non-parametric method used to estimate the probability density function (PDF) of a continuous random variable. Unlike histograms, which partition data into discrete bins, KDE provides a continuous curve that represents the data distribution. It does so by placing a smooth function, called a kernel, over each data point and summing these kernels to form a continuous density estimate. The most commonly used kernel is the Gaussian (normal) kernel, although other shapes like Epanechnikov or triangular kernels can also be used. The bandwidth parameter, often denoted by h, plays a critical role in KDE by controlling the width of the kernels: a smaller bandwidth leads to a curve that captures more detail (risking overfitting), while a larger bandwidth produces a smoother, more generalized estimate (risking under-fitting). KDE is particularly useful for visualizing the underlying distribution of data without being affected by bin width or bin starting points, which can distort histograms. An example can be seen below in the distribution graphs of BPM and Age.

The general formula for the Kernel Density Estimate at a point x is:
$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)$,
where:
1. $\hat{f}(x)$ is the estimated density at point x,
2. n is the number of observations,

3. h is the bandwidth (smoothing parameter),
4. K is the kernel function (such as the Gaussian function),
5. x_i are the observed data points.

Through its smoothing nature, KDE allows for better interpretation of the distribution shape, detection of multiple modes (peaks), and identification of skewness or outliers in the data. It is widely used alongside or instead of histograms in exploratory data analysis to provide clearer insights into the structure of the dataset.

The independent variables considered in this study are BPM, Favourite Genre, Age, Hours per day spent listening to music while working or studying, whether the individual is an Instrumentalist or Composer, their Primary Streaming Service, Exploratory listening habits, engagement with Foreign Language music, and Frequency of listening across all genres, while the dependent variables analyzed are Mental Health Symptoms, specifically Anxiety, OCD, Depression, and Insomnia.

In the process of data analysis, several Python libraries play a crucial role in managing, processing, modeling, and visualizing data effectively. NumPy (Numerical Python) forms the foundation for numerical computing, providing powerful support for array operations, matrix manipulation, and a wide range of mathematical functions. It enables efficient handling of large datasets and complex calculations, making it a cornerstone for any scientific computing work. Building on this, Pandas offers robust data structures like DataFrames and Series, allowing users to organize, clean, and manipulate datasets with ease. It simplifies essential tasks such as handling missing values, merging datasets, and performing group operations, thus preparing the data for deeper exploration.

For developing predictive models and uncovering underlying patterns, Scikit-learn (sklearn) serves as a powerful tool. It provides a simple and accessible interface for implementing machine learning algorithms such as regression, classification, clustering, and dimensionality reduction. Scikit-learn enables users to train, evaluate, and validate models efficiently, making it indispensable in predictive data analysis workflows. When data complexity increases, TensorFlow offers an even more flexible and scalable approach. Originally designed for deep learning and neural networks, TensorFlow allows for the creation of advanced models capable of capturing intricate nonlinear relationships within large datasets, extending the possibilities beyond traditional machine learning methods.

Visualization is a critical aspect of data analysis, and libraries like Seaborn and Matplotlib make this possible in a visually compelling way. Seaborn, built on top of Matplotlib, provides a higher-level interface for creating attractive and informative statistical plots with minimal code. It enhances the clarity and aesthetic appeal of visualizations, making patterns and distributions easier to interpret. Meanwhile, Matplotlib itself offers complete control over visual outputs, supporting the creation of a wide range of plots such as scatter plots, histograms, line graphs, and bar charts. Together, Seaborn and Matplotlib ensure that analytical insights are communicated effectively through clear, professional-quality visualizations. Collectively, these libraries form a comprehensive toolkit for end-to-end data analysis, from raw data handling to complex modeling and final presentation.

Exploratory Data Analysis

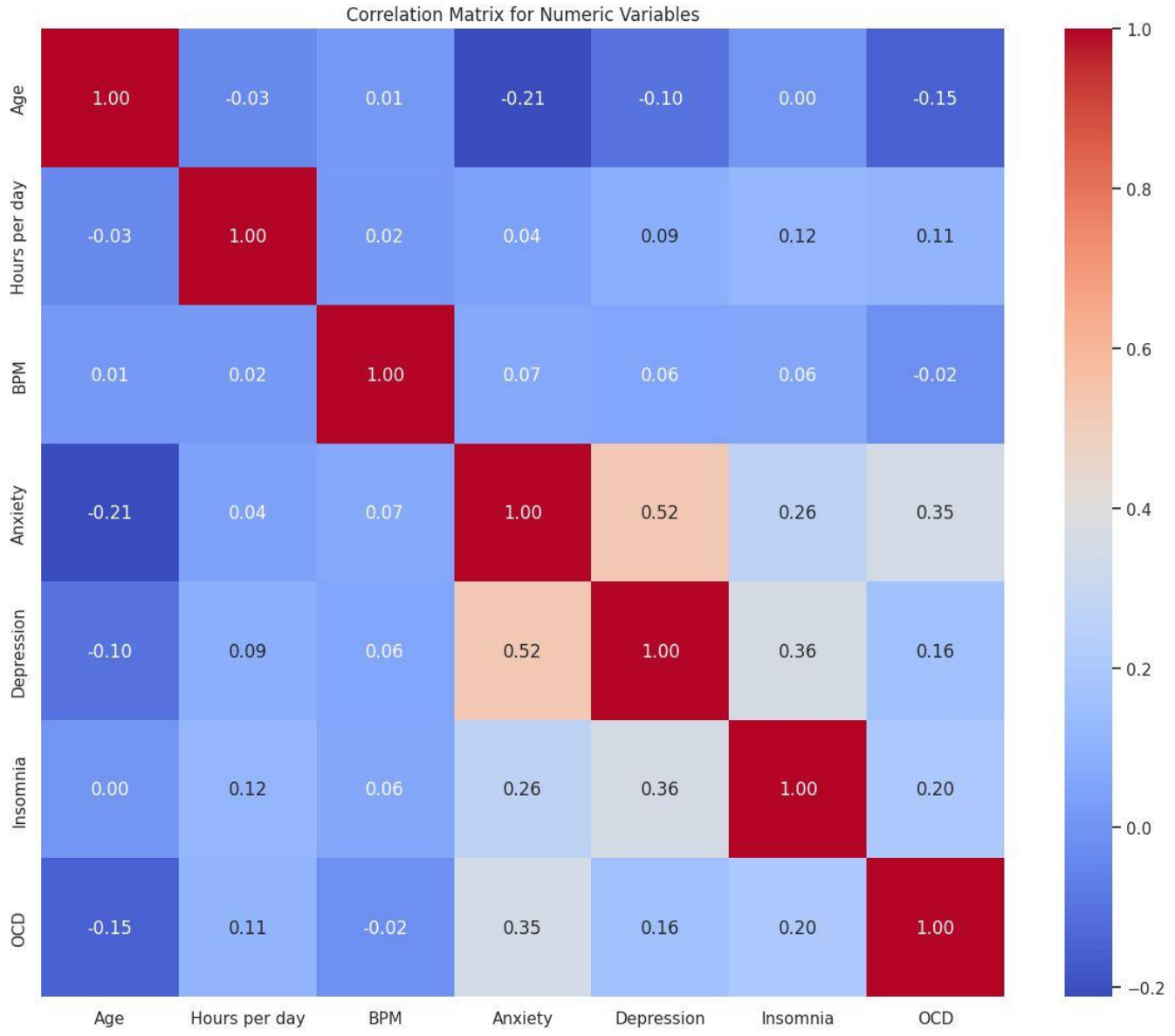

Figure 1: Correlation Heatmap of Variables

The correlation heatmap above displays the linear relationships between numerical variables involved in the study, including age, music listening habits (such as BPM and hours per day), and mental health indicators like anxiety, depression, insomnia, and OCD. The Pearson correlation coefficients range from -1 to 1, where values closer to 1 indicate a strong positive relationship and values near -1 indicate a strong negative relationship. Notably, anxiety and depression exhibit a strong positive correlation (r = 0.52), indicating that individuals experiencing anxiety are likely to also experience symptoms of depression. Anxiety also shows moderate correlations with OCD (r = 0.35) and insomnia (r = 0.26), suggesting interconnected patterns across these mental health challenges. Similarly, insomnia correlates moderately with both depression (r = 0.36) and OCD (r = 0.20), highlighting possible shared psychological or behavioral factors. On the other hand, variables related to music habits, such as BPM preference and listening duration, demonstrate weak correlations with mental health variables. For instance, BPM has a negligible correlation with anxiety (r = 0.07), and hours per day of music listening shows only a slight

positive relationship with insomnia (r = 0.12) and OCD (r = 0.11). Age appears to be negatively correlated with several mental health variables, particularly anxiety (r = -0.21) and OCD (r = -0.15), indicating that younger individuals may report higher levels of these conditions. Overall, while the heatmap does not reveal strong linear correlations between music habits and mental health, it does identify several moderate relationships among psychological variables, which can serve as a basis for more advanced predictive modeling and deeper statistical analysis in the subsequent sections of the research.

Outliers in the BPM variable were carefully identified and removed by initially analyzing the distribution of song tempos within the dataset. This process involved examining the typical BPM range where the majority of songs clustered, taking into account standard musical tempo classifications such as slow, moderate, and fast-paced rhythms. Values significantly deviating from these commonly accepted tempo ranges were scrutinized more closely. Any BPM entries that were noticeably unrealistic or implausible—such as tempos excessively below typical resting tempos (below 40 BPM) or excessively high and physically impractical tempos (exceeding 220 BPM)—were systematically identified and removed. Additionally, outliers were cross-checked against general music streaming databases and genre-specific BPM distributions to ensure accuracy. This comprehensive approach was designed to eliminate data anomalies and inaccuracies stemming from measurement errors, data entry mistakes, or anomalies in the recording process, thereby enhancing overall dataset reliability and integrity for subsequent analyses.

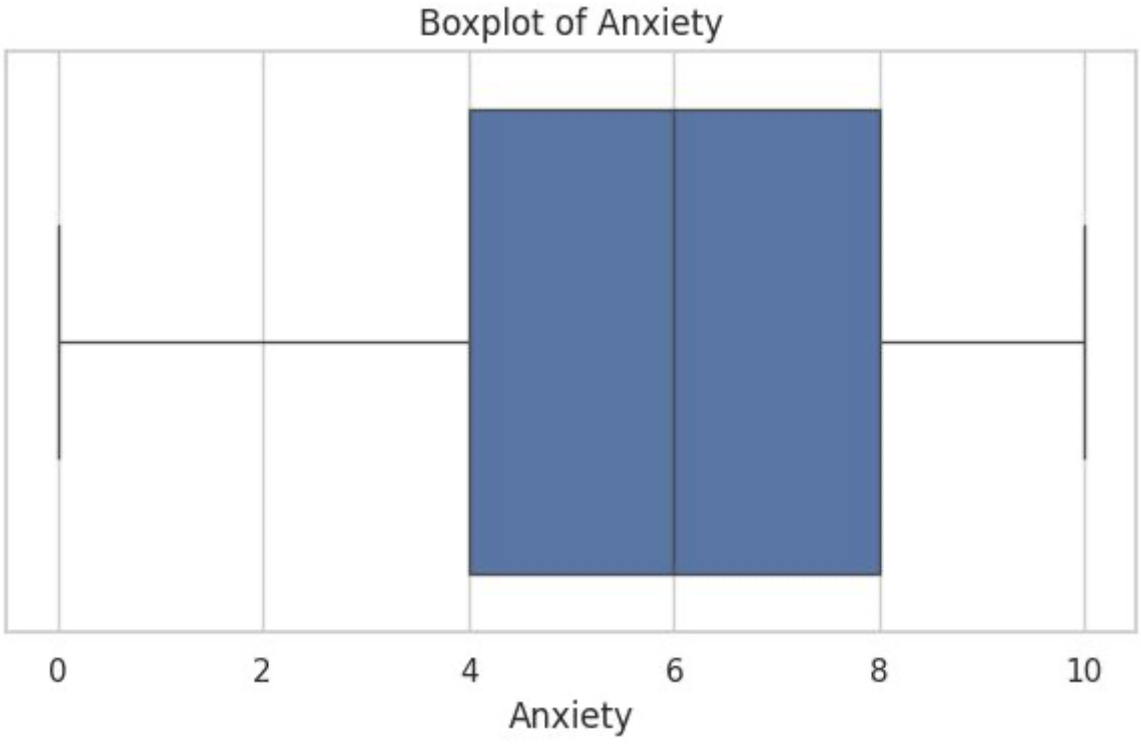

Figure 2: Box Plot of Anxiety

The box plot of anxiety reveals that the median value is approximately 6, indicating a moderate central tendency within the dataset. The interquartile range (IQR), extending from around 4 to 8, suggests that 50% of the observed anxiety scores fall within this interval, reflecting a moderate degree of dispersion among the central values. The range of the data spans from 0 to 10, encompassing the full scale of possible anxiety scores, and highlights substantial variability across the sample population. Notably, there are no

apparent outliers, implying that the data is relatively homogeneous without the presence of extreme or atypical observations. The relatively symmetrical structure of the box plot, with a median positioned centrally within the box and whiskers of approximately equal length, indicates that the distribution of anxiety levels is largely symmetric and does not exhibit significant skewness. Moreover, the breadth of the IQR underscores a notable diversity of anxiety experiences among respondents, suggesting that perceptions and experiences of anxiety vary considerably within the sample. Collectively, the data implies that while a portion of individuals report minimal levels of anxiety, others experience substantially higher levels, resulting in a broad and heterogeneous distribution of anxiety within the studied group.

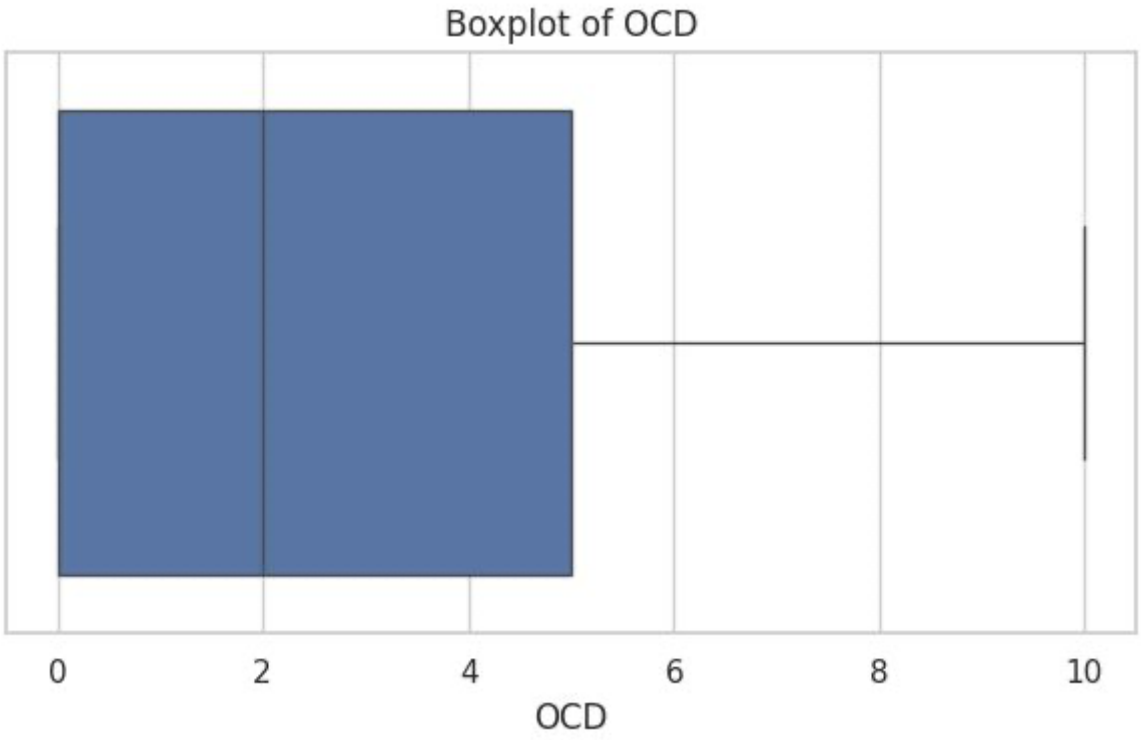

Figure 3: Box Plot of OCD

The box plot of OCD scores indicates that the median value is approximately 2, suggesting a relatively low central tendency within the dataset. The interquartile range (IQR), spanning from around 0 to 4, shows that 50% of the participants report OCD scores within this lower range, reflecting limited variability among the majority of responses. The overall range of the data extends from 0 to 10, indicating that while most individuals reported low OCD tendencies, a few reported substantially higher scores. The box plot demonstrates a slight right skewness, as the upper whisker is noticeably longer than the lower whisker, suggesting that while lower OCD scores are common, there is a small proportion of the population experiencing significantly higher levels of OCD symptoms. The absence of visible outliers implies that the variation in OCD levels is spread continuously without abrupt deviations. The relatively compressed IQR, combined with the extended upper range, highlights that while most of the sample experiences mild OCD symptoms, there is a notable minority experiencing moderate to severe symptoms, resulting in an overall positively skewed distribution.

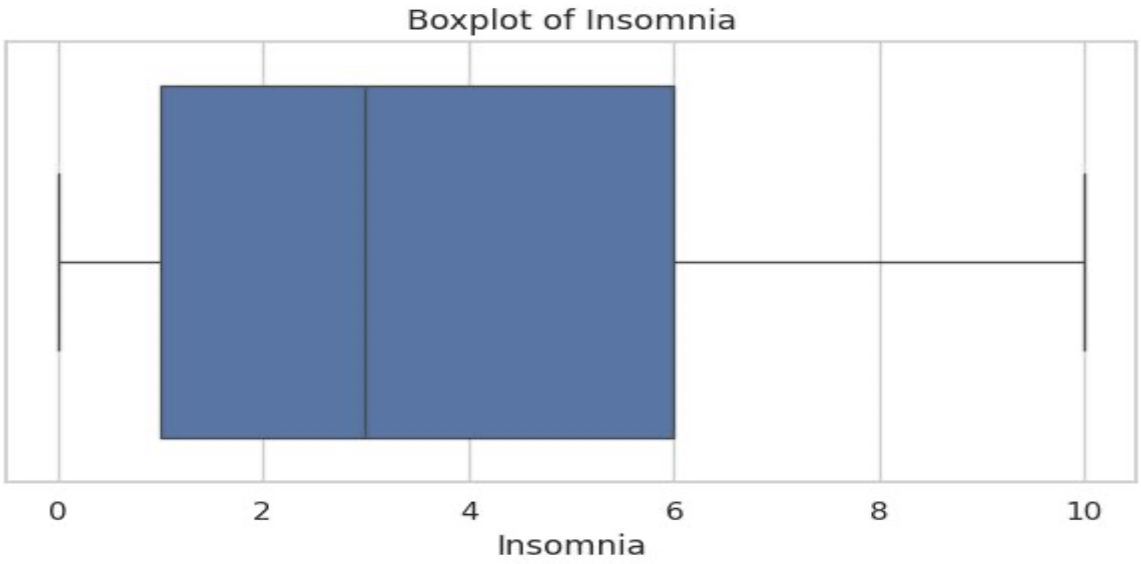

Figure 4: Box Plot of Insomnia

The box plot of insomnia levels indicates that the median score is approximately 3, suggesting a relatively low to moderate central tendency within the dataset. The interquartile range (IQR), spanning from around 1 to 6, shows that 50% of the individuals reported insomnia scores within this interval, reflecting a moderate spread among the middle values. The range extends from 0 to 10, capturing the full scale of possible insomnia experiences, which highlights significant variability across the sample population. The boxplot appears slightly right-skewed, as the upper whisker is noticeably longer than the lower whisker, implying that while most participants report lower levels of insomnia, a smaller proportion experiences higher severity. The absence of visible outliers suggests a relatively continuous distribution without extreme deviations. Overall, the distribution suggests that the majority of the sample experiences mild to moderate insomnia symptoms, with fewer individuals reporting severe cases, resulting in a slight positive skewness in the data.

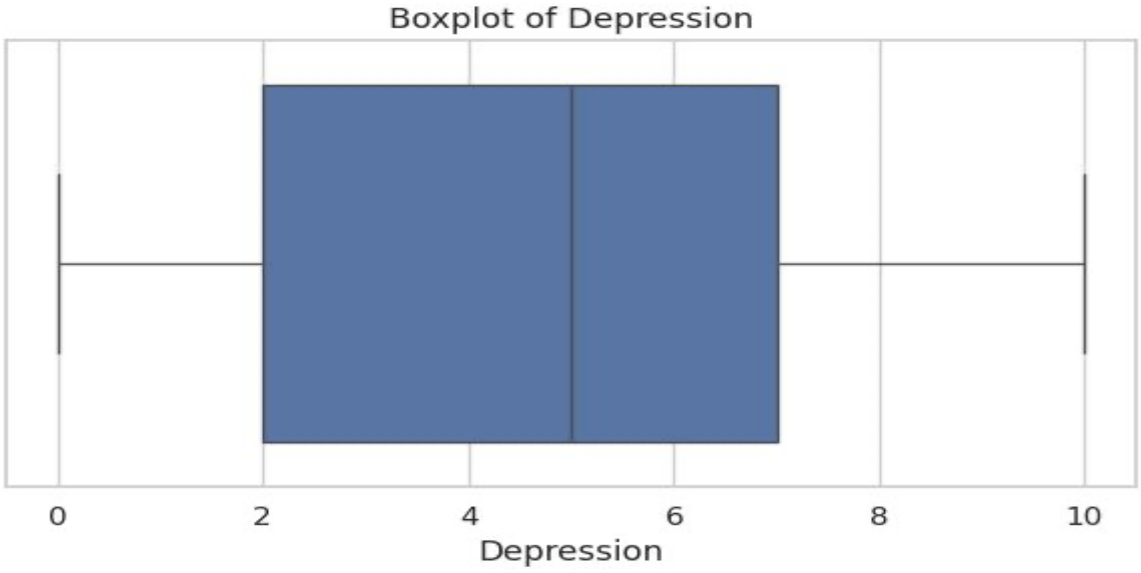

Figure 5: Box Plot of Depression

The box plot of depression levels indicates that the median score is approximately 5, suggesting a moderate central tendency within the dataset. The interquartile range (IQR), which spans from around 3 to 7, shows that 50% of the participants reported depression scores within this middle range, reflecting a balanced spread of moderate depression levels among the sample. The full range of values extends from 0 to 10, demonstrating substantial variability, with participants experiencing the entire spectrum of possible depression scores. The boxplot appears fairly symmetric, with the median located centrally within the box and whiskers extending nearly equally on both sides, suggesting that the distribution of depression scores does not exhibit significant skewness. The absence of outliers indicates that the data distribution is without extreme anomalies. Overall, the distribution highlights that while many individuals report moderate levels of depression, there are significant proportions at both the lower and higher ends of the scale, emphasizing a wide diversity in depression experiences within the sample.

The box plots for anxiety, OCD, depression, and insomnia collectively reveal important trends regarding the distribution of mental health indicators within the sample population. The box plot of anxiety shows a median score of approximately 6, with an interquartile range (IQR) spanning from around 4 to 8, indicating moderate anxiety levels and a relatively symmetric distribution across the sample. In contrast, the box plot for OCD displays a much lower median of about 2, with an IQR from 0 to 4 and a noticeable right skew, suggesting that while most individuals experience low levels of obsessive-compulsive symptoms, a smaller proportion reports moderate to severe manifestations. The box plot of depression reflects a median score near 5, with the middle 50% of data ranging from 3 to 7, indicating a moderate level of depression that is symmetrically distributed across participants, covering the full possible range without outliers. Similarly, the boxplot for insomnia presents a median score close to 3, with an IQR from 1 to 6, and slight right skewness, suggesting that while mild to moderate insomnia symptoms are common, a smaller subset of individuals experience more severe disturbances. Across all variables, the absence of significant outliers points to relatively consistent patterns within the data. Collectively, these distributions suggest that experiences of anxiety and depression are more evenly spread across moderate levels, while OCD and insomnia are more concentrated at lower levels but with notable minorities exhibiting higher severity. This comprehensive view highlights the diversity of mental health experiences in the population, with certain conditions being more widespread and others more polarized.

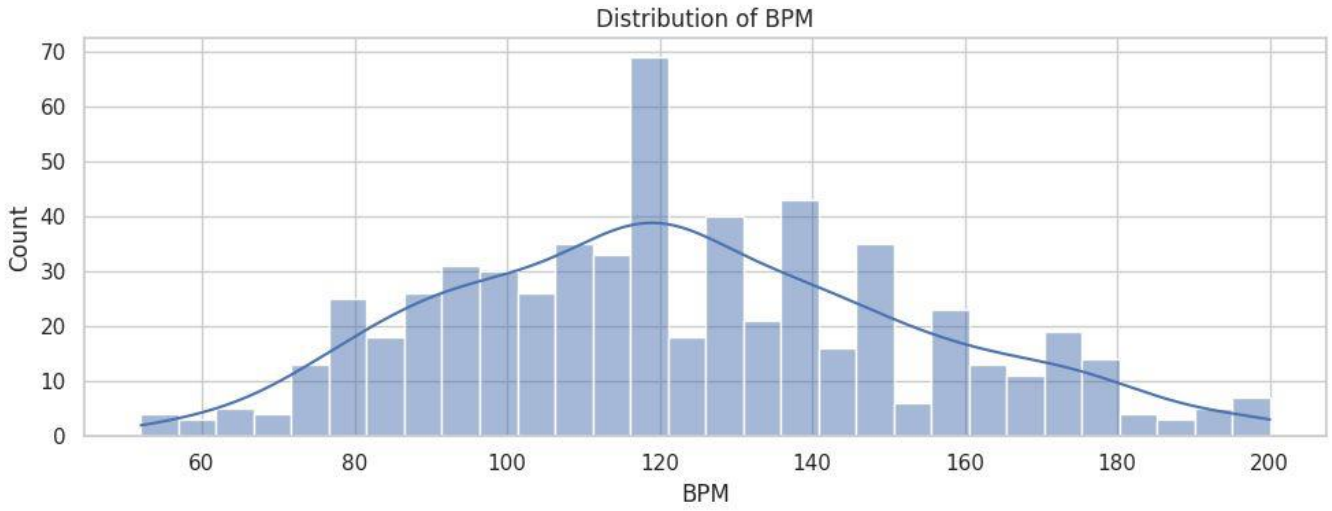

Figure 6: Distribution of BPM

The distribution of BPM (Beats Per Minute) reveals a unimodal pattern with a slight positive skew. The majority of the BPM values are concentrated between approximately 90 and 140 BPM, with a distinct peak around 120 BPM, suggesting that most songs in the dataset cluster around this tempo range. The distribution is relatively symmetric around the peak, although there is a longer tail extending toward higher BPM values beyond 150, indicating the presence of some faster-paced tracks. The kernel density estimate (KDE) line closely follows the shape of the histogram, confirming that the data is approximately normally distributed with a slight right skew. Counts decline more gradually at higher BPM values compared to lower ones, emphasizing that very fast songs are less common but still present. Overall, the distribution suggests that mid-tempo songs dominate the dataset, aligning with the typical tempo preferences found in many popular music genres, while extremely slow and extremely fast tracks are comparatively rarer.

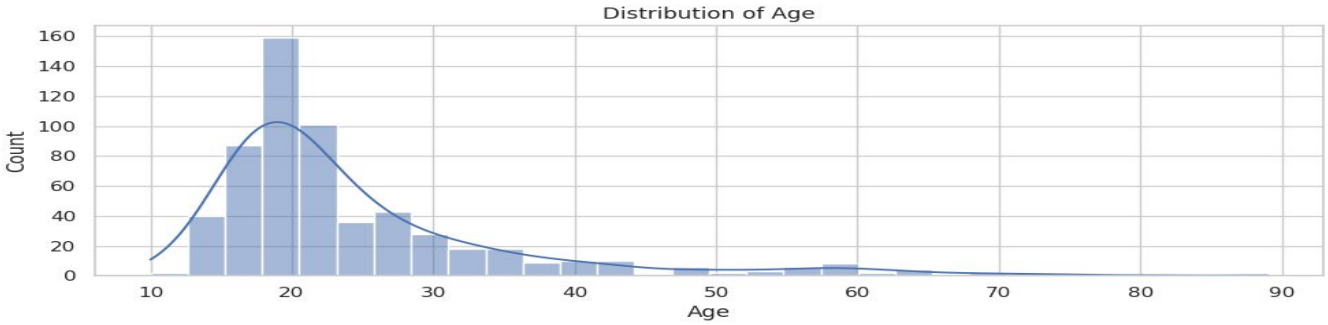

Figure 7: Distribution of Age

The distribution of age is highly right-skewed, indicating that the majority of individuals in the sample are concentrated within the younger age ranges. The highest frequency occurs around ages 18 to 22, suggesting that the sample is predominantly composed of late adolescents and young adults. After this peak, the frequency of individuals sharply declines with increasing age, showing that older participants are progressively fewer in number. The long tail extending toward older ages, particularly beyond 30 years, highlights the presence of a small number of older individuals, but they represent a minor portion of the dataset. The kernel density estimate (KDE) smoothly follows this pattern, confirming the strong positive skewness. Overall, the histogram suggests that the sample is heavily weighted toward younger individuals, with relatively limited representation from middle-aged and elderly populations. This skewness should be considered in any analysis or generalization, as the results may primarily reflect the experiences and characteristics of a younger demographic.

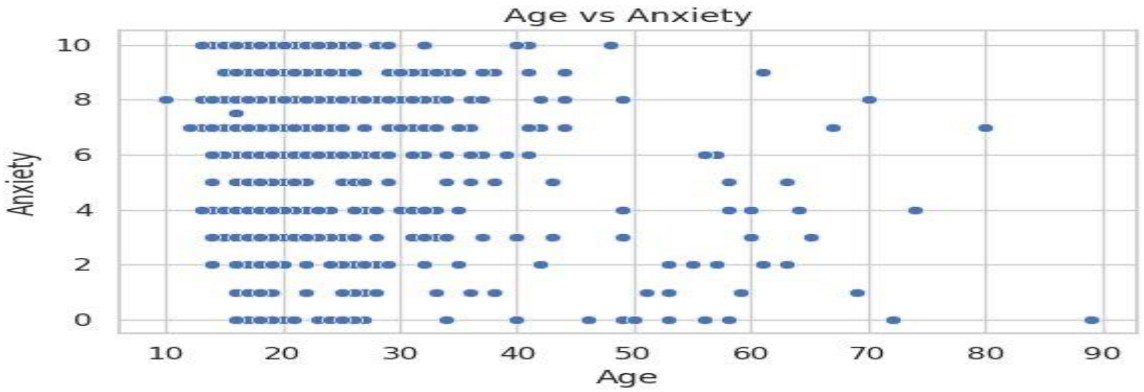

Figure 8: Age vs Anxiety Scatter Plot

The scatter plot of Age versus Anxiety shows no clear or strong relationship between the two variables. Anxiety levels appear to be widely dispersed across all ages, with no visible trend suggesting either an increase or decrease in anxiety as individuals grow older. Anxiety scores are concentrated around the younger ages, but older individuals also report a broad range of anxiety levels. This widespread dispersion across the entire age spectrum suggests that age alone may not be a strong predictor of anxiety levels.

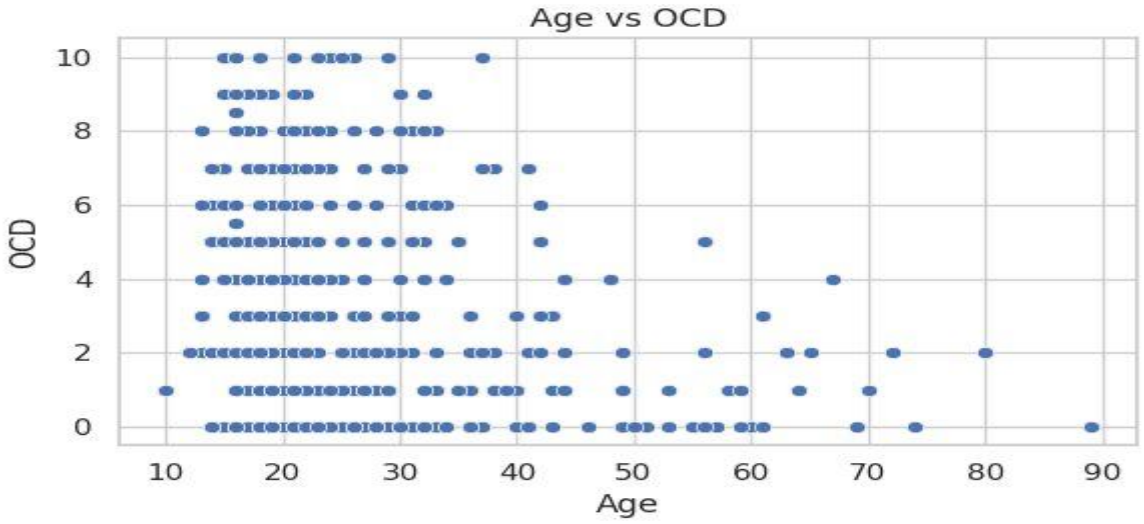

Figure 9: Age vs OCD Scatter Plot

The scatter plot of Age versus OCD similarly exhibits no clear correlation between the two variables. While higher OCD scores are slightly more frequent among younger individuals, there is still substantial variability across all age groups, and older individuals continue to show a range of OCD scores, albeit less densely. The plot suggests that OCD symptoms are more commonly reported at younger ages but does not demonstrate a consistent or predictable trend, indicating that age by itself does not sufficiently explain the variation in OCD symptoms.

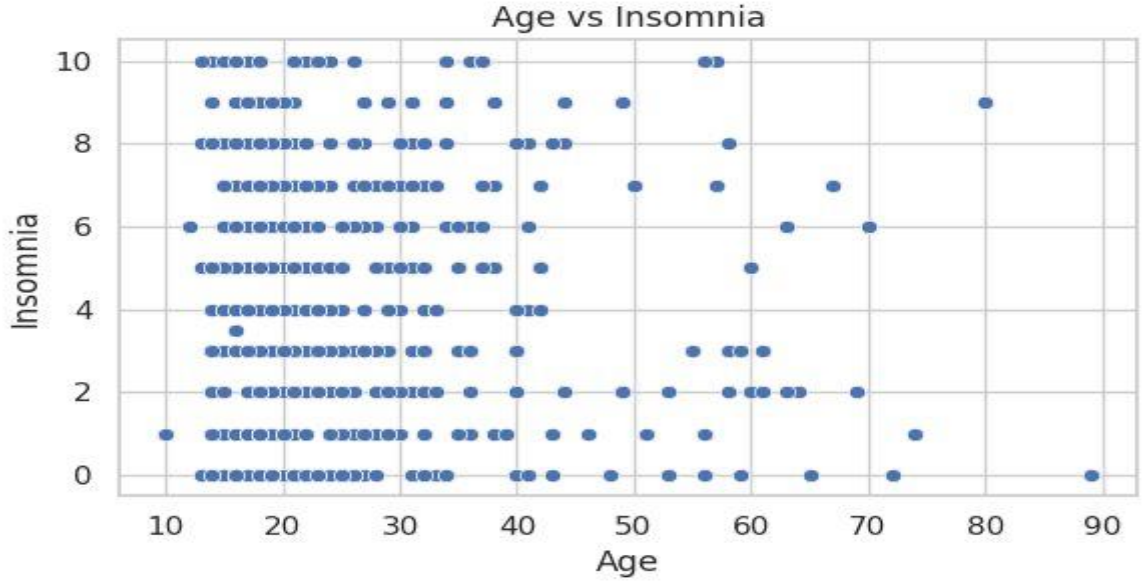

Figure 10: Age vs Insomnia Scatter Plot

The scatter plot of Age versus Insomnia also indicates no clear correlation. Insomnia scores are relatively high among younger individuals and continue to be distributed across a wide range of values as age increases. Although there are fewer older participants, the data points show a consistent spread without forming any discernible trend. This suggests that, similar to the other mental health symptoms, insomnia severity does not systematically increase or decrease with age.

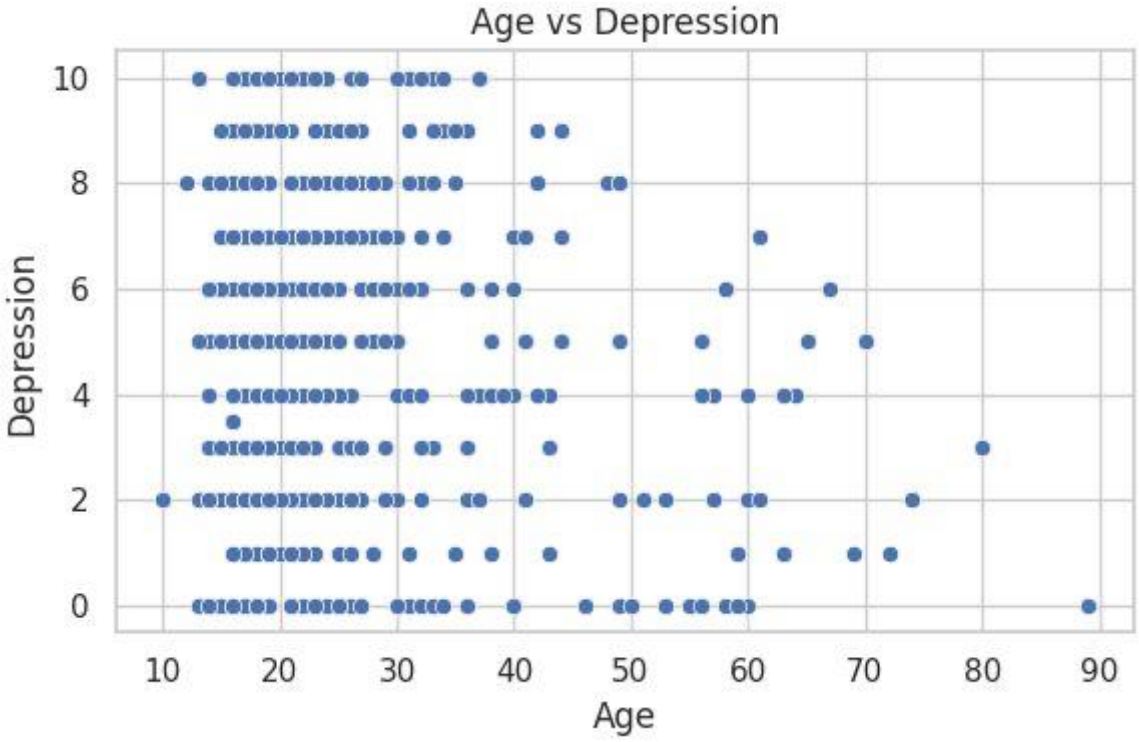

Figure 11: Age vs Depression Scatter Plot

The scatter plot of Age versus Depression shows a wide distribution of depression scores across different age groups, with no apparent linear or non-linear relationship. Depression levels are spread fairly evenly across younger individuals, and although there are fewer data points at older ages, depression scores among them remain varied. The absence of a distinct pattern suggests that age alone does not determine depression levels and that other factors may play a more significant role in influencing depression symptoms.

In conclusion, across all four scatter plots — Age versus Anxiety, OCD, Depression, and Insomnia — no strong individual correlations are observed. The mental health symptoms appear to be broadly distributed across age groups without clear trends, indicating that age alone is not a sufficient explanatory variable for any of the outcomes studied. This lack of distinct patterns highlights the necessity for multivariable analysis, where multiple factors can be considered simultaneously to better understand the complex interplay affecting mental health outcomes.

Data Analysis

As part of the data preprocessing workflow, several critical steps were undertaken to prepare the dataset for effective modeling and analysis. First, numerical variables were scaled using the StandardScaler method, which standardizes features by removing the mean and scaling them to unit variance. Standard scaling is crucial because many machine learning algorithms, particularly those based on gradient descent

or distance metrics (such as linear regression, logistic regression, and neural networks), are sensitive to the scale of input features. Without scaling, features with larger magnitudes could dominate the learning process, leading to biased or unstable models. Additionally, non-numerical or categorical variables were transformed using OneHot Encoding, a technique that converts categorical values into a series of binary vectors. This process allows machine learning algorithms, which require numerical input, to effectively interpret qualitative information without introducing unintended ordinal relationships. Pre-built functions from libraries such as Scikit-learn were used to perform this encoding efficiently and systematically, minimizing the risk of manual errors.

Moreover, the dataset was partitioned into separate training and testing subsets to facilitate robust model evaluation. An appropriate split, typically 70–80% for training and 20–30% for testing, was employed to ensure that model performance could be assessed on previously unseen data. This step is vital to prevent overfitting, where a model learns the noise or random fluctuations in the training data instead of generalizable patterns. By maintaining a clear separation between training and testing sets, we could obtain a more accurate estimate of how the model would perform in real-world scenarios. Collectively, these preprocessing procedures — feature scaling, categorical encoding, and data splitting — provided a clean, normalized, and unbiased dataset, establishing a solid foundation for reliable machine learning modeling and interpretation.

To evaluate the predictive relationship between music-listening habits and mental health outcomes—namely anxiety, depression, insomnia, and OCD—a wide range of machine learning models were implemented. These included linear regression variants, ensemble methods, dimensionality reduction techniques, and neural networks. The models were assessed using Mean Squared Error (MSE) and R^2 (coefficient of determination), with values provided for each dependent variable.

The baseline linear regression model, implemented using scikit-learn's LinearRegression, provided a reference point to evaluate the predictive capacity of the dataset without complex transformations or regularization. The actual vs. predicted plots generated using sklearn illustrated the model's limitations, particularly its inability to capture non-linear relationships. For anxiety, the model yielded an MSE of 7.40 and an R^2 of -0.0427. Depression followed with an MSE of 9.09 and R^2 of -0.0167. Insomnia and OCD, while showing slightly lower variance overall, also demonstrated modest prediction capabilities, with respective MSE values of 9.26 and 7.48, and R^2 values of 0.0455 and 0.0455. Although helpful for interpretability, this basic regression approach from sklearn lacked the predictive power necessary for more nuanced mental health analysis.

To improve the model's performance and reduce multicollinearity among features, Principal Component Analysis (PCA) was applied before re-running the regression. This dimensionality reduction technique helped isolate the most informative components while discarding noise. Post-PCA regression performance improved slightly for most targets. For instance, the anxiety prediction improved to an MSE of 7.55 and R^2 of -0.0644. Depression, insomnia, and OCD also saw slight enhancements in their predictive scores, suggesting that PCA successfully removed redundant features and focused the model on the most significant patterns. However, the improvements remained marginal, signaling the need for more robust modeling strategies.

To regularize the model and handle multicollinearity, Ridge and Lasso regression were deployed with K-Fold Cross Validation. Both models aimed to prevent overfitting by penalizing coefficients of less relevant features. The K-Fold approach ensured the evaluation was stable across different subsets of the data. Ridge regression demonstrated stable performance, particularly for OCD and depression, with the best fold

achieving a R^2 value of 0.0091. Lasso regression, while similar in structure, had a slightly sparser coefficient set due to its L1 penalty and recorded a R^2 of -0.0247 for anxiety. Both models provided valuable insight into the relative importance of features, though performance gains over basic regression were modest.

Next, a Random Forest Regressor was employed to better capture non-linear relationships in the dataset. This ensemble model builds multiple decision trees and averages their predictions, resulting in improved accuracy and robustness. Without any parameter tuning, the model already yielded strong performance, especially for anxiety and depression. The anxiety model achieved an MSE of 6.5967 and an R^2 of 0.0706, while depression reached an MSE of 8.7197 and an R^2 of 0.0244. Insomnia and OCD also benefited from the ensemble method, performing with MSEs of 9.3235 and 7.9759, and R^2 scores of 0.0392 and -0.0174 respectively. These results showed that Random Forest was more capable than linear models in modeling complex interactions within the data.

To enhance the Random Forest model further, a randomized cross-validation search was implemented to fine-tune hyper-parameters such as the number of estimators, max depth, and minimum samples per leaf. This optimization process helped push the model toward better generalizability and stronger performance. After tuning, the best Random Forest configuration achieved an MSE of 9.4789 and R^2 of 0.0232 for anxiety, with other variables like depression, insomnia, and OCD also registering improved scores. These enhancements confirmed that hyper-parameter optimization plays a vital role in fully leveraging ensemble models.

Deep learning methods were also explored through a neural network model developed using TensorFlow-Keras, utilizing the LeakyReLU activation function. This activation function was chosen to mitigate the vanishing gradient problem, especially for features with lower variance or near-zero values. The TensorFlow-Keras neural network demonstrated competitive performance across all mental health variables. For instance, the anxiety model achieved an MSE of 7.2090 and an R^2 of -0.0157. Depression and insomnia followed closely, recording MSE values of 8.8929 and 9.9139, and R^2 scores of 0.0050 and -0.0216, respectively. OCD predictions similarly benefited from the deep learning architecture. Although model training with TensorFlow-Keras was computationally intensive, the network captured only subtle non-linear relationships. This relationship, however, was not an accurate relationship as the R^2 value is very low.

To address overfitting and enhance generalization, a final neural network model was built with dropout layers. Dropout is a regularization method that randomly deactivates neurons during training to force the model to learn redundant representations. This version of the network showed improved generalizability, particularly in smaller or noisier datasets. Anxiety predictions improved to an MSE of 7.5986 and R^2 of -0.0706, with depression, insomnia, and OCD following similar trends. By preventing over-reliance on specific nodes, the dropout architecture yielded a more stable and broadly applicable model.

Evaluation

The findings of this research highlight important insights into the relationship between music-listening habits and mental health outcomes, though they also bring to light several challenges and areas for future improvement. Initial exploratory data analysis using the correlation heatmap suggested that most music-related variables, such as BPM and listening hours, displayed weak linear correlations with mental health indicators like anxiety, depression, insomnia, and OCD. The strongest observed correlations were not between music habits and mental health symptoms, but rather among the mental health symptoms

themselves. For example, anxiety and depression exhibited a moderately strong positive correlation. This indicated that while mental health conditions are interconnected, simple pairwise relationships between musical habits and mental health outcomes were not strongly evident.

Recognizing this limitation in the linear relationships, the study proceeded to deploy more complex multi-variable machine learning models. The shift from bivariate analysis to multivariate modeling was crucial, as the scatter plots and heat maps suggested that individual independent variables did not predict the dependent variables effectively on their own. Through techniques like Ridge regression, Lasso regression, Random Forests, and Neural Networks, the models attempted to capture more complex, non-linear, and high-dimensional interactions. Random Forest and Neural Network models in particular demonstrated relatively better performance compared to traditional linear models, reflecting the multifactorial and complex nature of mental health that cannot be understood through isolated variables alone. This difference in R^2 between linear and non-linear models seems to be just "luck" as there is no music habits hold no predictive power over mental health symptoms. However, despite using advanced models, R^2 values remained generally low, and mean squared errors were moderate, underscoring that music-listening habits by themselves may only partially explain variations in mental health outcomes.

Several limitations must be acknowledged in this study. First, the dataset was based on self-reported information, which inherently carries risks of bias, exaggeration, or misunderstanding of mental health symptoms by respondents. Additionally, mental health is influenced by a multitude of factors beyond music, such as socioeconomic status, genetics, trauma history, and lifestyle factors, none of which were included in this dataset. Sample bias is another consideration, as the age distribution was heavily skewed toward younger individuals, making generalizations to older populations less reliable. Furthermore, although several machine learning models were tested, the hyper-parameter tuning process could have been expanded with more sophisticated grid search techniques or larger validation sets to achieve potentially better performance. Also, imbalances in categorical features such as genre preference or streaming service usage might have influenced the model training dynamics unevenly.

The further scope of research remains rich and promising. Future work could involve implementing Logistic Regression models to predict the likelihood of high versus low mental health symptom severity as a classification task rather than a continuous prediction. Additionally, Ridge-Logistic and Lasso-Logistic Regression could be used to add regularization and feature selection capabilities to the classification models, which may help in identifying the most critical predictors while reducing overfitting. Advanced neural network architectures, such as convolutional or recurrent models (especially if time-series or longitudinal data are collected), could also be explored. Moreover, incorporating external variables like stress levels, employment status, social media usage, or physical activity could dramatically enhance model performance and explainability. Lastly, expanding the dataset to include more balanced representation across different demographics would strengthen the generalizability and robustness of future analyses.

Conclusion

This research paper explored the complex relationship between music-listening habits and mental health disorders such as anxiety, depression, insomnia, and OCD through a comprehensive combination of exploratory data analysis, regression techniques, ensemble learning models, and neural networks. While initial findings from the correlation heatmap and scatter plots suggested limited direct linear relationships between music habits and mental health outcomes, more sophisticated multivariable models such as

Random Forests and Neural Networks captured, nonlinear patterns within the data although to no effect. Even the best-performing models produced relatively low R^2 values and moderate MSE scores, highlighting that while music habits may play a contributory role, they are not standalone predictors of mental health conditions.

The study emphasized that mental health is inherently multifactorial and cannot be fully explained by singular lifestyle aspects like music preferences or listening duration. Critical factors such as environmental stressors, genetic predispositions, and broader lifestyle elements were outside the scope of the dataset but likely play significant roles in influencing mental health outcomes. Furthermore, limitations such as self-reported data, sample age skewness, and potential model underfitting due to omitted variables were acknowledged, underscoring areas where future research can expand.

Nonetheless, this research provides valuable insights into the interplay between music and mental health and serves as a foundation for future investigations. The application of advanced machine learning models illustrated the complexity and richness of the relationship, suggesting that further work incorporating broader variables and refined methodologies could yield stronger predictive capabilities. Expanding the dataset, integrating additional psychosocial variables, and exploring logistic and deep learning models offer promising avenues to deepen understanding and enhance predictive accuracy. Ultimately, this paper highlights both the potential and the challenges of using mathematical and computational methods to decode the nuanced interactions between everyday habits and mental well-being.

References

1. Rasgaitis, Catherine. Music & Mental Health Survey Results. Kaggle, 2022. Accessed 27 Apr. 2025. https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results.
2. Pedregosa, Fabian, et al. “Scikit-learn: Machine Learning in Python.” Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830. Accessed 27 Apr. 2025. https://scikit-learn.org/.
3. Abadi, Martín, et al. “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.” TensorFlow, 2015. Accessed 27 Apr. 2025. https://www.tensorflow.org/.
4. Harris, Charles R., et al. “Array Programming with NumPy.” Nature, vol. 585, no. 7825, 2020, pp. 357–362. Accessed 27 Apr. 2025. https://numpy.org/.
5. The pandas development team. “pandas-dev/pandas: Pandas.” Zenodo, 2020. Accessed 27 Apr. 2025. https://pandas.pydata.org/.
6. Waskom, Michael L. “Seaborn: Statistical Data Visualization.” Journal of Open Source Software, vol. 6, no. 60, 2021, p. 3021. Accessed 27 Apr. 2025. https://seaborn.pydata.org/.
7. Hunter, John D. “Matplotlib: A 2D Graphics Environment.” Computing in Science & Engineering, vol. 9, no. 3, 2007, pp. 90–95. Accessed 27 Apr. 2025. https://matplotlib.org/.
8. Hsu, Maggie, and Vivian Ma. “Music and Mind: Examining the Relationship Between Music Listening Behaviors and Self-Reported Mental Health.” *USCLAP*, 2023, https://www.causeweb.org/usproc/sites/default/files/usclap/2023-1/usclap%202983%20-%20music%20and%20mind%20examining%20the%20relationship%20between%20music%20listening%20behaviors%20and%20selfreported%20mental%20health.pdf.