Case Study Analysis on Music and Mental Health Survey

Siyad N,
MSc Computer Science (Data Analytics),
Rajagiri College of Social Sciences (Autonomous), Rajagiri P.O
Kalamassery, Kochi, Kerala,683104

Abstract:
The relationship between music and mental health has long been studied, and this survey aims to gather more information about the topic. The survey was conducted online, and a total of 736 participants completed the survey. This Study deals with how Music affects one's mental health. Music is effective to regulate our moods and emotions. Listening to music can uplift our mood, decrease feelings of anxiety, and promote relaxation. Listening to calming music has been shown to reduce stress levels, lower cortisol levels, and improve our overall sense of well-being. These are our basic thought about how music affects our mental health. In this study, we will be looking forward to analyzing how music actually affects our mental health. This dataset has the target variable "Music Effect" which will give information about our mental status by dividing the classes into three categories Improve, Not affect, and Worsen. In the Overall study of the data set, a statistical report says that 74.43% of people's mental health improves by hearing music and 23.44% of people's mental health has no effect on music and 2.33% of people's mental health was Worsen by music. The project aims to build multiple classification models using the Orange tool and select the model with the best accuracy.

Keyword:
[Objectives], [Data description], [Preprocessing], [Visualization],[Model Building],[Evaluating Results]

I. INTRODUCTION

Music and mental health are closely connected, and research has shown that listening to music can have a significant impact on our mental well-being. Music has been found to help reduce stress, anxiety, and depression, and it can also improve mood, increase motivation, and promote relaxation.

One reason why music is so effective at promoting mental health is that it has the ability to affect the brain and the body in a variety of ways. For example, certain types of music can activate the release of dopamine, a neurotransmitter associated with pleasure and reward, which can lead to feelings of happiness and satisfaction.

The Music and Mental Health Survey data set is a collection of responses from a survey conducted to investigate the relationship between music and mental health. The survey was distributed online and in person and collected responses from individuals of different age groups, genders, and cultural backgrounds. The survey consists of questions related to music listening habits, preferred genres of music, frequency of listening, and the impact of music on mental health. It also includes questions about the respondents' mental health, such as their Anxiety, depression, insomnia, and OCD.

The data set Includes a range of variables, including demographic information, music-related variables, and mental health-related variables. The data set can be used to explore the relationship between music and mental health, as well as other related insights from the dataset, and also to identify patterns and trends in music listening habits and mental health symptoms. Overall, this data set provides a valuable resource for researchers and mental health professionals interested in understanding the role of music in promoting mental well-being.

II. LITERATURE REVIEW

Music is known to improve mental wellness. Listening to and making music improves mood, emotions, and cognition, according to research. Music and mental health have become more popular in recent years, and various literature studies have summarized the present level of knowledge.

[1] Van der Wal et al. (2018) reviewed the evidence that music improves mental health. Music therapy helps alleviate sadness, anxiety, and PTSD, according to the authors (PTSD). Music therapies can improve

self-esteem, social functioning, and quality of life, according to the review.
[2] Fancourt and Finn (2019) examined how music prevents and treats youth mental health issues. Music-based therapies improved anxiety, depression, and self-esteem, the scientists found. Music therapy can also help trauma survivors.
[3] Särkämö et al. (2019) examined how music therapies affect cognitive functioning in neurological illnesses including dementia and stroke. Music-based therapies improve cognitive performance, memory, attention, mood, and quality of life.
[4] Moore et al. (2021) reviewed music and mental health during the COVID-19 epidemic. The review noted that music can bring comfort and connection during social isolation and promote mental health and well-being under stress and uncertainty.

III. OBJECTIVES

- To find which music genre is mostly affected by Mental health
- To find the comparison of Mental health status between people who were listening to music greater than 12 hours and less than 12 hours
- To find which age groups are mostly affected by Mental health-related problems.
- To find the impact of music on Music health.

IV. METHODOLOGY

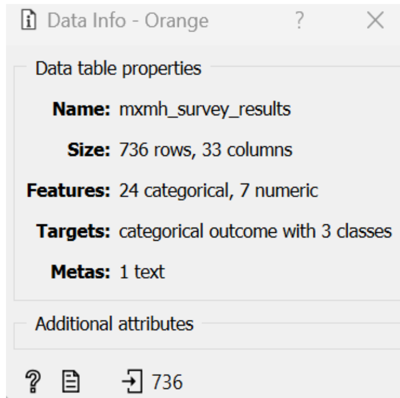
Tool Used- Orange.
Orange (3.310) is an open-source data mining and visualization toolkit. It is used for explorative rapid qualitative data analysis and interactive data visualization.
Before getting into the process of methodologies and machine learning model building, first of all, we have to understand the dataset and do many kinds of studies.

The methodology will explain the following steps,

- About the data
- Data Description
- Pre-processing
- Visualization
- Model Construction
- Evaluation of results

1. ABOUT THE DATA

Data collection was managed by @catherinerasgaitis via a Google Form. Respondents were not restricted by age or location.
The form was posted in various Reddit forums, Discord servers, and social media platforms. Posters and "business cards" were also used to advertise the form in libraries, parks, and other public locations.
The form was relatively brief so that respondents would be more likely to finish the survey. "Harder" questions (such as BPM) were left optional for the same reason.



As we can see that this data set consists of a total of 736 rows and 33 columns,24 categorical variables, and 7 numerical variables, and also target variable with three classes and one meta data.

2. DATA DESCRIPTION

NO	ATTRIBUTES	DESCRIPTION
1	Timestamp	Date and time when form was submitted.
2	Age	Respondent's age.
3	Primary Streaming Service	Respondent's primary streaming Service.
4	Hours per day	Number of hours the Respondent listen to Music per day.
5	While working	Does the Respondent listen to the music while working/studying.
6	Instrumentalist	Does the Respondent play an instrument regularly?
7	Composer	Does the Respondent Compose music?
8	Fav genre	Respondent's favorite or top genre.

9	Exploratory	Does the Respondent actively explore new artists/genres?
10	Foreign language	Does the Respondent regularly listen to music with lyrics in a language they are not fluent in?
11	BPM	Beat per minute of the favorite genre.
12	Frequency[Classical]	How frequently the respondent listens to classical music
13	Frequency [Country]	How frequently the respondent listens to country music
14	Frequency [EDM]	How frequently the respondent listens to EDM music
15	Frequency [Folk]	How frequently the respondent listens to Folk music
16	Frequency [Gospel]	How frequently the respondent listens to Gospel music
17	Frequency [Hip hop]	How frequently the respondent listens to HIP hop music
18	Frequency [jazz]	How frequently the respondent listens to jazz music
19	Frequency [K pop]	How frequently the respondent listens to K pop music
20	Frequency [Latin]	How frequently the respondent listens to Latin music
21	Frequency [Lofi]	How frequently the respondent listens to Lofi music
22	Frequency [Metal]	How frequently the respondent listens to Metal music
23	Frequency [Pop]	How frequently the respondent listens to Pop music
24	Frequency [R&B]	How frequently the respondent listens to R&B music
25	Frequency [Rap]	How frequently the respondent listens to Rap music
26	Frequency [Rock]	How frequently the respondent listens to Rock music
27	Frequency [Video game]	How frequently the respondent listens to Video game music
28	Anxiety	Self-reported anxiety, on a scale 0-10
29	Depression	Self-reported Depression, on a scale 0-10
30	Insomnia	Self-reported Insomnia, on a scale 0-10
31	OCD	Self-reported OCD, on a scale 0-10
32	Music effects	Does music improve/No effect/worsen respondent's mental health condition?
33	Permissions	Permissions to publicize data

3. PRE-PROCESSING

An essential phase in the data mining process is data preparation. It describes the processes of preparing data for analysis by cleansing, converting, and integrating it. The purpose of data preprocessing is to enhance the data's quality and suitability for the particular data mining operation. The main objective of data pre-processing is that it helps to build an efficient model.
Information about the data is given below.

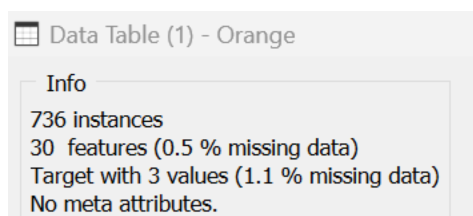


Figure 3.1

From figure 3.1 we can see that this dataset has missing values.
There are several steps involved in data pre-processing they are:
STEP1: As the first step of pre-processing we have to first remove 2 columns that are Timestamp(metadata) and Permissions (As this column does not play an important role in the entire dataset).

STEP2: Then we have to step forward with the imputation of missing values. As we can see that this dataset has a certain amount of missing values that have to be resolved with the help of Imputer.

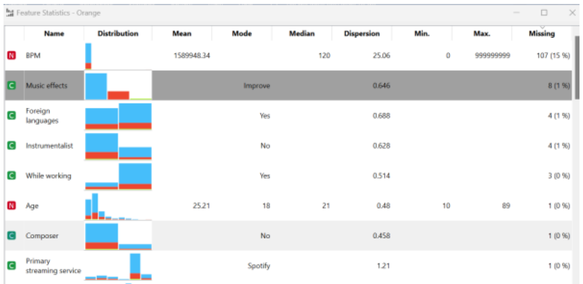


Figure 3.2

From figure 3.2 we can evident that 8 columns have missing values out of which BPM has the highest missing values about 107. So in order to impute the missing values columns other than BPM have very few missing values so that we can drop those values. But in the BPM column, we will impute the missing values with a model-based imputer(simple tree). This model is fitted to the observed data before being used to produce imputations for the values that are missing.

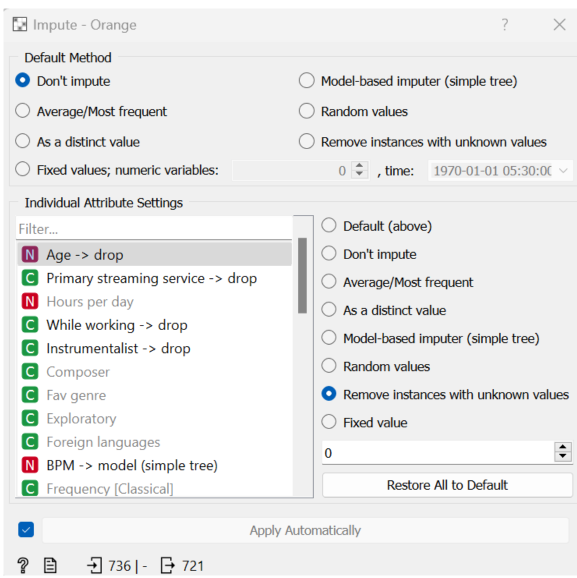


Figure 3.3

Figure 3.3 shows the imputation of missing values.

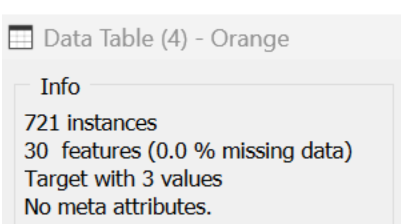


Figure 3.4

Figure 3.4 shows that the dataset has no missing values.

STEP3: This step deals with the discretization of the Column Age. So that we have to first convert the datatype of the age from Numerical to Categorical. By doing so at the stage of visualization we can attain very good inferences. Age is classified into three categories.

- Teenagers - up to 19 years old
- Adults – between 20 to 50 years old
- Seniors – above 51 years old

Age	Age groups
18	Teenegers
61	Seniors
18	Teenegers
18	Teenegers
18	Teenegers
21	Adults

Before *After*

STEP4: In this Step Removal of Outliers take place. An observation that differs abnormally from other values in a population-based random sample is referred to as an outlier. In the Orange tool with the help of the Outlier widget, we can remove the outliers. Here removal of outliers is based on Euclidean distance.

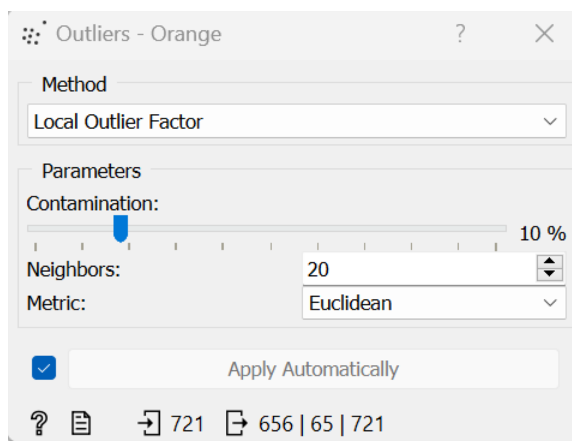


Figure 3.4

In figure 3.4 we can clear that out of 721 data about 65 data are considered as outliers based on Euclidean distance.

STEP5: In this step replace the Boolean values ie, Yes/No values with 1 and 0 respectively for the columns(While working, Instrumentalist, Composer, Exploratory, and Foreign Languages) in order to make the data in a more standard format.

- Yes – 1
- No – 0

While working	Instrumentalist	Composer	Exploratory
No	No	No	No
Yes	No	Yes	Yes
Yes	No	No	Yes
Yes	Yes	Yes	Yes
Yes	Yes	No	Yes
Yes	No	No	Yes

Figure 3.5.a

While working	Instrumentalist	Composer	Exploratory
0	0	0	0
1	0	1	1
1	0	0	1
1	1	1	1
1	1	0	1
1	0	0	1

Figure 3.5.b

Figure 3.5.a will give the values with Yes/no before conversion and Figure 3.5.b gives the values with 1 and 0 after conversion.

STEP6: In this step mainly normalization of certain columns takes place. In the dataset, we can have the columns(Anxiety, Depression, insomnia, OCD) which are some of the mental health problems, values ranging from 0-10. The least value is 0 which means that mental health problem is unaffected by the person and the highest value is 10 which affect quite largely. Normalization makes your database more navigable and enhances the accuracy and integrity of your data. The values of the columns are Normalized to (0 to 1).

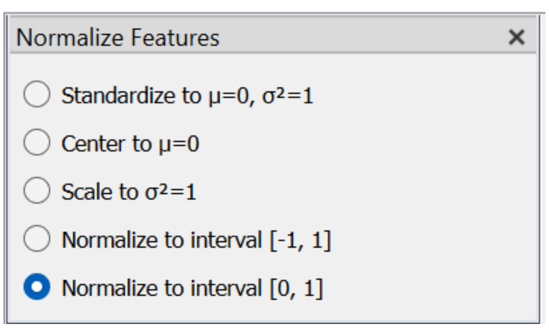


Figure 3.6

Anxiety	Depression	Insomnia	OCD
7.0	7.0	10.0	2.0
9.0	7.0	3.0	3.0
7.0	2.0	5.0	9.0
8.0	8.0	7.0	7.0
4.0	8.0	6.0	0.0
5.0	3.0	5.0	3.0

Anxiety	Depression	Insomnia	OCD
0.70	0.70	1.00	0.20
0.90	0.70	0.30	0.30
0.70	0.20	0.50	0.90
0.80	0.80	0.70	0.70
0.40	0.80	0.60	0.00
0.50	0.30	0.50	0.30

Anxiety	Depression	Insomnia	OCD
0.70	0.70	1.00	0.20
0.90	0.70	0.30	0.30
0.70	0.20	0.50	0.90
0.80	0.80	0.70	0.70
0.40	0.80	0.60	0.00
0.50	0.30	0.50	0.30

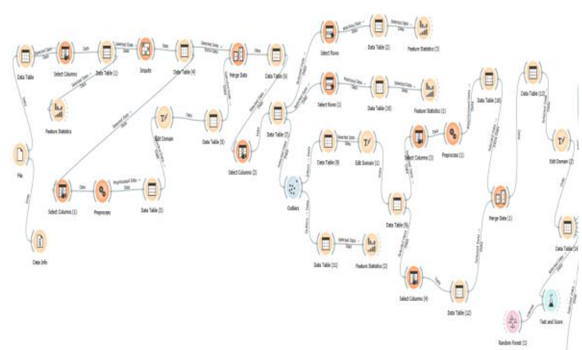
Figure 3.7.b

Figure 3.7.a gives the information before Normalization whereas figure 3.7.b gives the information after normalization.

STEP7: The frequency of each genre column consists of 4 values that are (Never, Rarely, Sometimes, and Very frequently). These values were replaced with certain numeric values they are represented below,

- Never – 0
- Rarely – 1
- Sometimes – 2
- Very frequently – 3

These type of pre-processing is done because of a better understanding of data.



Overall view of Pre-processing Steps

4. DATA VISUALIZATION

Data visualization is the representation of data in graphical or pictorial form, using charts, graphs, and other visual elements. The purpose of data visualization is to make complex data more accessible, understandable, and useful to people who need to analyze and interpret it. By visually presenting data, it becomes easier to identify patterns, trends, and relationships between different data points.

Age group:

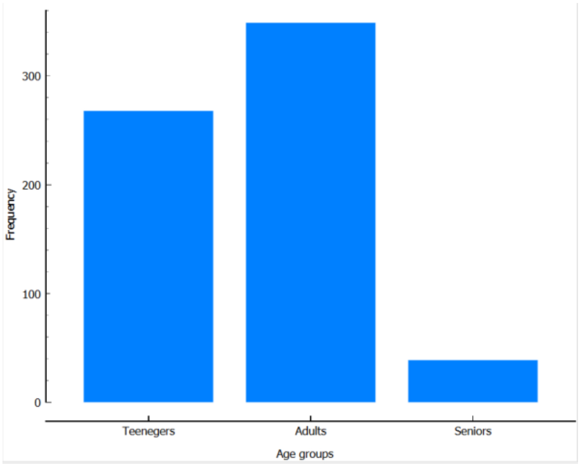


Figure 4.1

From this distribution, we can clearly identify that Adults are the most ones who have responded more than Teenagers and Seniors

- Teenagers – 40.85%
- Adults – 53.20%
- Seniors – 5.95%

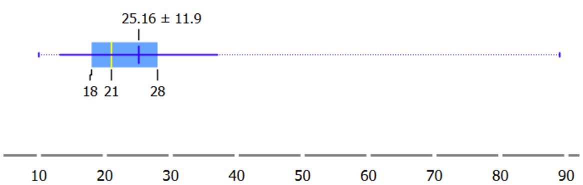


Figure 4.2

From this Box plot, we can clearly identify that most people having aged between 18 and 28. Also, we can see that there are certain outliers after 40 years of age that are removed in the pre-processing step.

Primary streaming service:

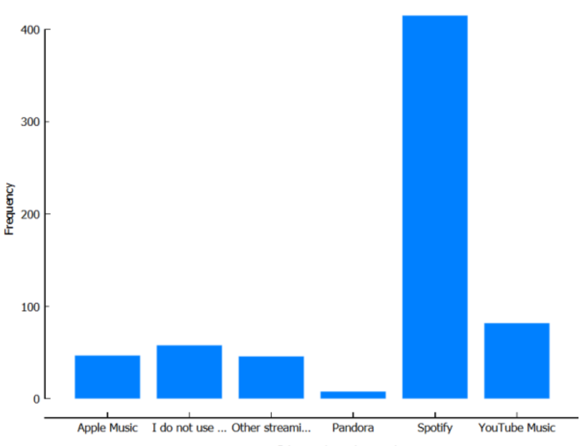


Figure 4.3

From this visualization, we can clearly identify that most users prefer Spotify as their primary streaming service as compared with other streaming platforms.

Hours per day:

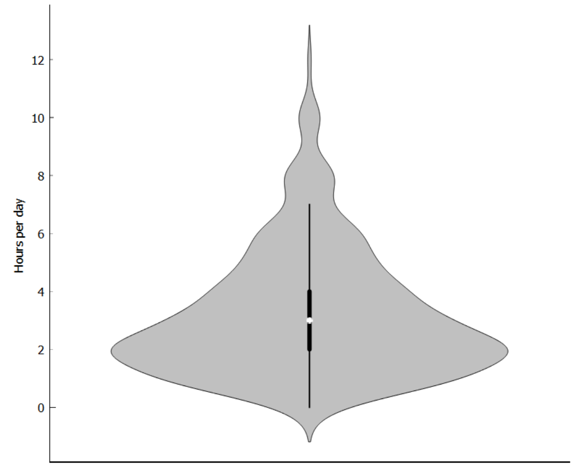


Figure 4.4

From this Violin plot, we can evident that most people spend 2 to 4 hours a day, listening to music. From that people mostly prefer 3 hours of listening to music.

Favorite genre:

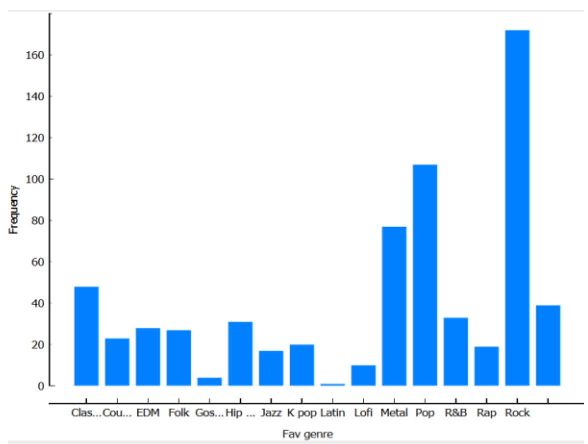


Figure 4.5

From this visualization, we can see that most people choose Rock, Pop, and metal as their favorite genre out of which Rock is the most popular one.

Music effect (target variable):

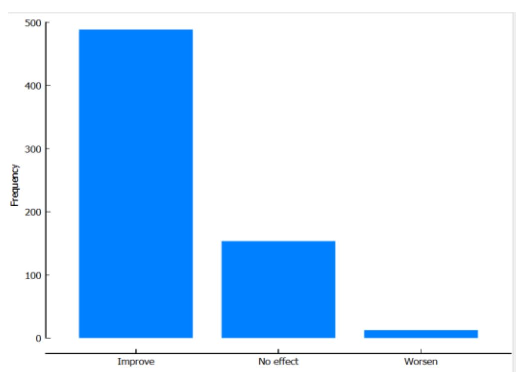


Figure 4.6

From this visualization, we can see that by listening to music most people's mental health improves. We can clearly see that the distribution among the classes of the target variable is quietly unbalanced, which means that Improve dominated the other two values which are, No effect and Worsen.

- Improve – 74.54%
- No effect – 23.48%
- Worsen – 1.98%

Other Inference gain through Visualization:

1. Age group vs Primary streaming service:

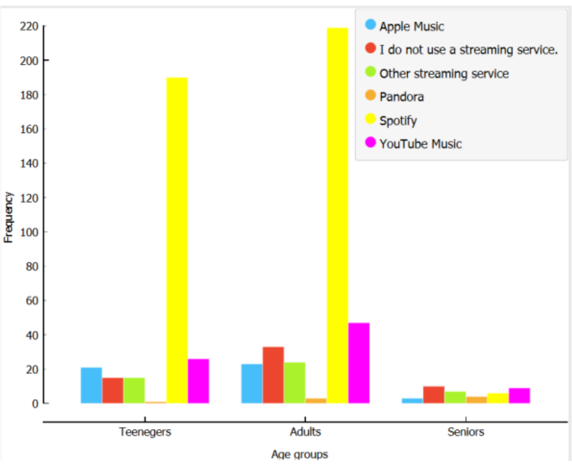


Figure 4.7

We can clearly see from figure 4.7 that Teenagers and Adults mostly prefer Spotify as their primary Streaming service whereas Seniors mainly prefer Youtube music or they do not mainly stick to a particular platform.

2. Mental health vs Age groups:

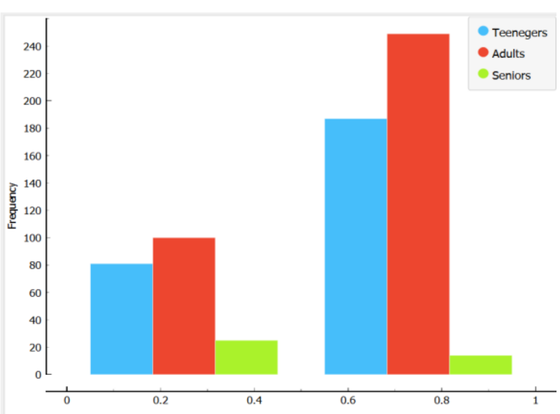


Figure 4.8

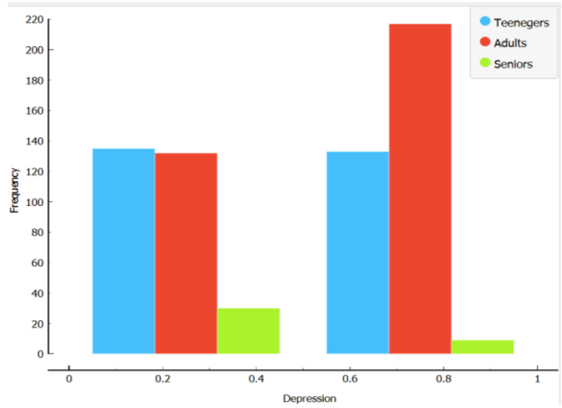


Figure 4.9

Figure 4.8 gives the distribution of Anxiety with respect to the age group whereas Figure 4.9 gives the distribution of Depression with respect to the age group. From these two graph we can get the clear idea that Most of the adults are Suffered from Anxiety and depression as compared to Teenagers. Also, People who have an age greater than 50 years(Seniors) suffer less from mental health problems.

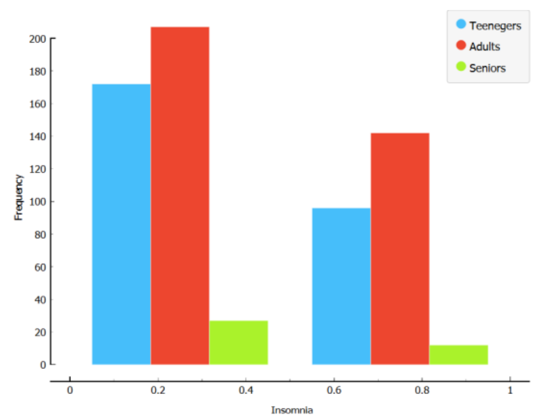


Figure 4.10

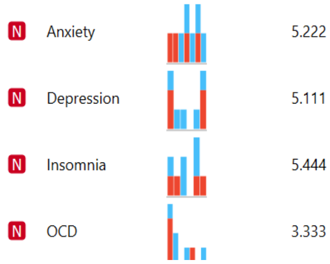
Figure 4.10 gives information about Insomnia with respect to age group. From the graph, it is clear that majority of the adults, teenagers, and seniors are less affected by insomnia as compared with anxiety and depression.

3. Mental health vs Hours per day

The average values of the user's mental health listening to music for less than 12 hours are given below

N	Anxiety	5.839
N	Depression	4.771
N	Insomnia	3.681
N	OCD	2.608

The average values of the user's mental health listening to music greater than 12 hours are given below.



From the figure, we can evident that there is a change in the average value of insomnia and OCD among the people who have listened to music above and below 12 hours. Insomnia and OCD are much greater in people who have listening to music greater than 12 hours per day. So we can say that people who have insomnia stay awake for much long time and listen to music.

4. <u>Mental health vs Favorite genre:</u>

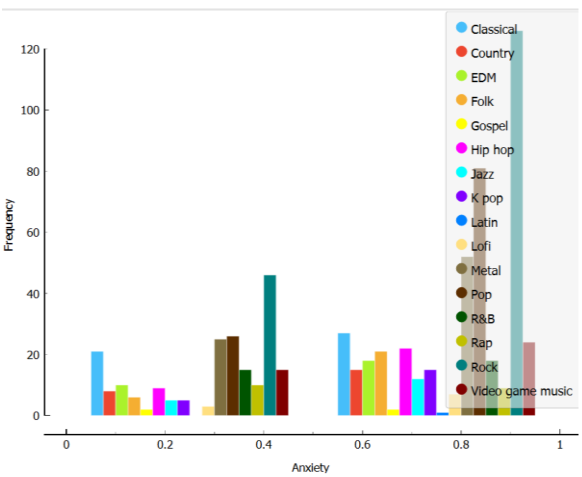


Figure 4.11

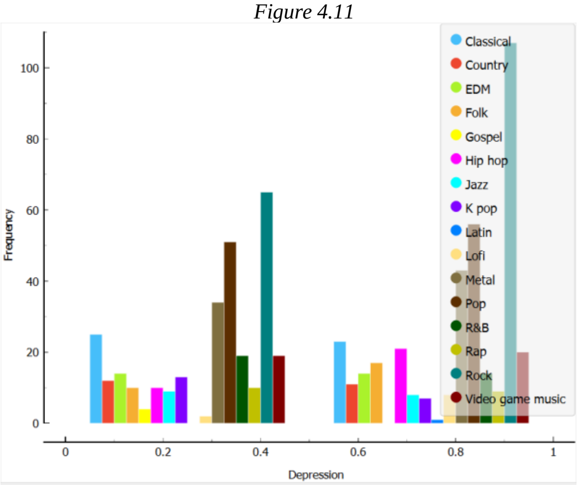


Figure 4.12

From Figures 4.11 and 4.12, we can say that people who have high anxiety and depression usually prefer Rock, Pop, and metal as their favorite genres as compared with all other genres.

5. <u>Mental health vs Music effect:</u>

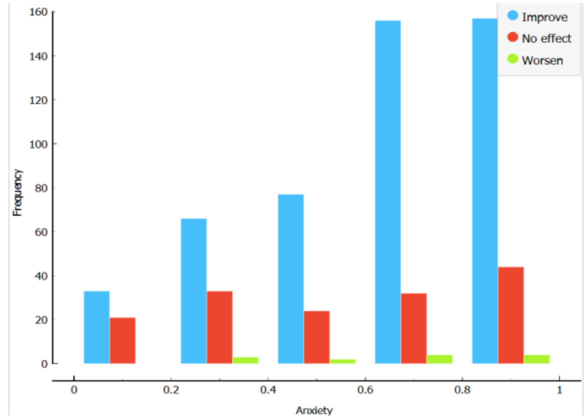


Figure 4.13

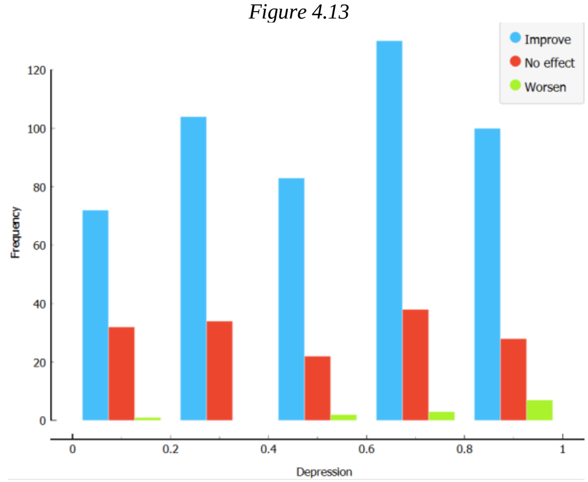


Figure 4.14

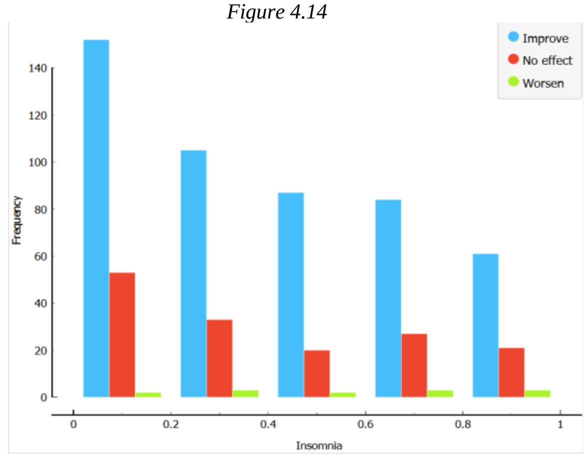


Figure 4.15

From the above analysis of visualization, it is evident that most people who have high anxiety, depression, and insomnia improve their mental health condition by listening to music. So Listening to music has great impact on people who were mentally down.

6. MODEL CONSTRUCTION

A classification model is used for Model construction as the dataset has the categorical target variable having three classes. Improve, Not effect, and Worsen. So this project consists of a few classification techniques. Classification models are a type of machine learning algorithm that is used to predict a categorical (discrete) target variable based on a set of input features. The goal of a classification model is to find a function that maps input variables to discrete output values, which are typically labels or classes.

Model	AUC	CA	F1	Precision	Recall
Tree	0.522	0.624	0.623	0.621	0.624
Random Forest (1)	0.539	0.719	0.650	0.623	0.719
Naive Bayes	0.531	0.165	0.246	0.673	0.165
Logistic Regression	0.586	0.682	0.660	0.644	0.682
Gradient Boosting	0.603	0.711	0.671	0.653	0.711

Figure 5.1

Figure 5.1 shows the accuracy of different models. As we can see from the figure the accuracy of the model was very low and it will result in poor performance. This is due to the dataset having unbalanced data so in order to build a better model we have to balance the data. Balancing of data is done either by up-sampling or down-sampling.

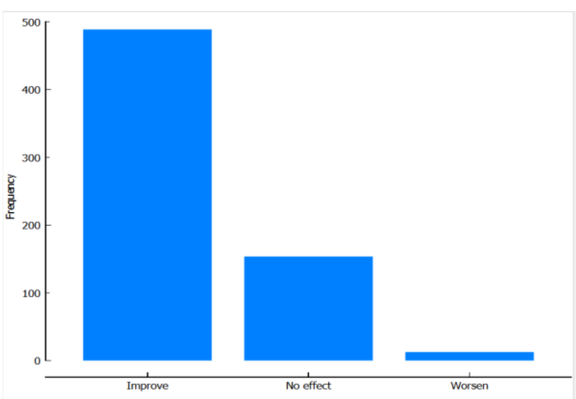


Figure 5.2

Figure 5.2 gives the distribution of the target variable before up-sampling the data.

- Improve -74.54%
- No effect - 23.48%
- Worsen – 1.98%

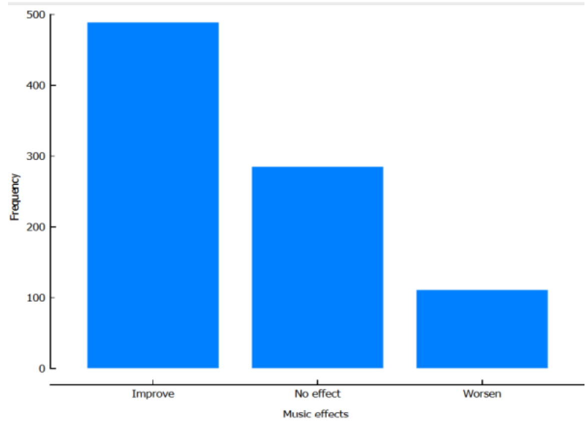


Figure 5.3

Figure 5.3 gives the distribution of the target variable after up-sampling the data

- Improve – 55.25%
- No effect – 32.30%
- Worsen – 12.54%

Now the data are more balanced than before and now we will construct a model for this balanced data and check whether the model can give better performance. There are various classification algorithms are there, such as Decision Tree, Random forest, SVM, Naïve Bayes, Gradient Boosting, and Neural networks. The data set is split into a 75:25 ratio for train and test.

i) <u>Decision tree Algorithm:</u>

The decision tree algorithm is a popular classification algorithm that works by recursively splitting the data into subsets based on the features of the data until the subsets are as pure as possible in terms of classification. The decision tree algorithm does not have a single equation or formula. Instead, it uses a set of rules that are learned from the data to make predictions or classify new instances. These rules are represented as a tree-like structure, where each node represents a decision based on a feature, and each branch represents the outcome of that decision. The leaves of the tree represent the final prediction or class label.

Model	AUC	CA	F1	Precision	Recall
Tree	0.809	0.734	0.733	0.733	0.734

Decision tree Evaluation parameters

ii) <u>Random Forest:</u>

Random forest is a machine learning algorithm that is used for classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. The algorithm works by creating a large number of decision trees and then aggregating their predictions to make a final prediction.

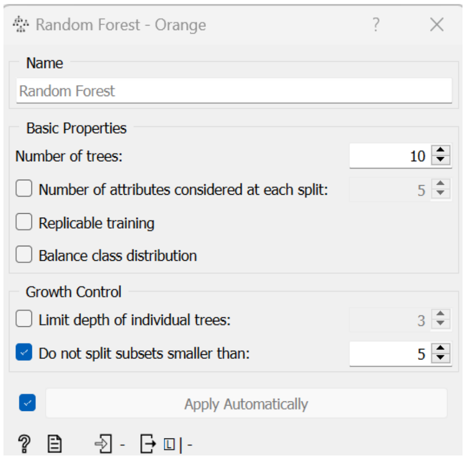


Figure 5.4

Model	AUC	CA	F1	Precision	Recall
Random Forest	0.884	0.806	0.802	0.802	0.806

Random forest Evaluation parameters

iii) <u>Support Vector Machine (SVM):</u>

Support vector machine, or SVM for short, is a well-liked machine learning technique used for both regression and classification applications. It operates by identifying the hyperplane that divides the data into distinct classes the most effectively.

Model	AUC	CA	F1	Precision	Recall
SVM	0.876	0.768	0.770	0.773	0.768

SVM Evaluation parameters

iv) <u>Gradient Boosting:</u>

A potent ensemble method that belongs to the family of machine learning techniques is gradient boosting. To produce a stronger model, ensemble approaches combine several weak models. In gradient boosting, a series of weak models that each attempt to fix the flaws of the preceding model are trained using the residuals of the prior model. A weighted average of the weak models makes up the final model.

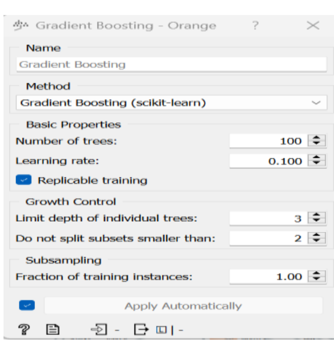


Figure 5.5

Model	AUC	CA	F1	Precision	Recall
Gradient Boosting	0.889	0.791	0.788	0.787	0.791

Gradient Boosting Evaluation parameters

v) <u>Neural Network:</u>

Neural networks are versatile classification techniques. Neural network categorization steps are summarized here. Stochastic gradient descent trains the neural network on the training set. Training reduces the loss function by adjusting neuron weights and biases. The validation set evaluates neural network performance. This lets us assess if the neural network overfits or under fits training data. To maximize validation set performance, neural network hyperparameters like learning rate, number of layers, and number of neurons can be modified.

After training and tuning the hyperparameters, the neural network can be evaluated on new data.

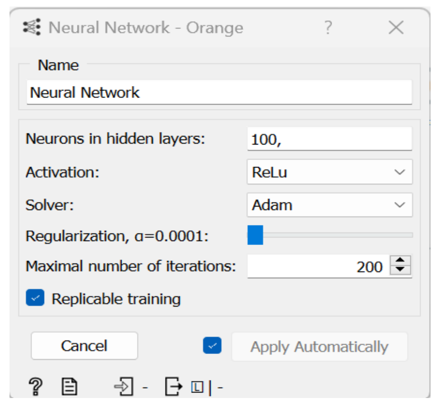
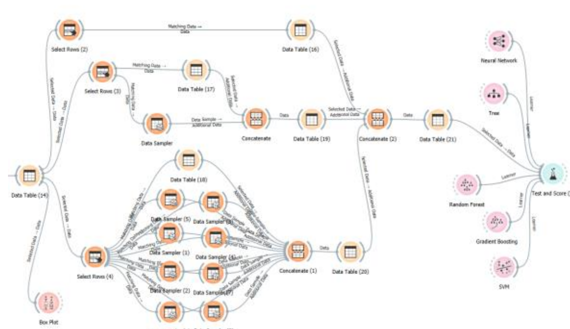


Figure 5.6

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.899	0.810	0.811	0.814	0.810

Neural network Evaluation parameters

Following is the Orange connection diagram:



7. EVALUATION OF RESULTS

The dataset "Music and mental health survey" consists of 885 rows and 30 columns (After up-sampling). The Selected target variable is the "Music effect". The data set is split into 75:25 ratios for train and test.
The different classification models applied in this dataset are Decision Tree, Random Forest, SVM, Gradient Boosting, and Neural Networks. The performance value obtained for each model is given below

Model	AUC	CA	F1	Precision	Recall
Tree	0.809	0.734	0.733	0.733	0.734
SVM	0.877	0.768	0.770	0.773	0.768
Random Forest	0.888	0.805	0.801	0.801	0.805
Neural Network	0.899	0.810	0.811	0.814	0.810
Gradient Boosting	0.889	0.791	0.788	0.787	0.791

Overall Evaluation result (After up-sampling)

We can evident from the above figure that after balancing the data the performance model becomes better as compared with unbalanced data which is why data balancing plays an important role in the performance of the model.
Among the applied classification models, the Neural network has the highest accuracy at about 81%. The Confusion matrix of the neural network is given below

Actual \ Predicted	Improve	No effect	Worsen	Σ
Improve	1178	287	5	1470
No effect	205	646	9	860
Worsen	0	0	330	330
Σ	1383	933	344	2660

Confusion matrix of neural network

V. CONCLUSION

According to studies, music can help people feel less stressed, anxious, and depressed. Moreover, it can raise motivation, elevate mood, and enhance cognitive performance. Many mental health issues, such as depression, anxiety, and post-traumatic stress disorder, have been demonstrated to be successfully treated by music therapy, which uses music to meet emotional, cognitive, and social requirements. From the analysis of this data, there were a lot of insights that are discovered which helps to identify the certain relationship among certain attributes such as Mental health with respect to age, Mental health with respect to the number of hours the music listen, per day, How mental health is affected with a favorite genre. From the overall analysis of data, we can get clear information about music improves people's mental health.

REFERENCES

- https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results
- https://www.nature.com/articles/s41598-019-49099-9
- https://www.kaggle.com/code/hazigin/music-mental-health-survey-results-eda
- https://www.researchgate.net/publication/358399496_Impact_of_Music_on_Mental_Health
- https://www.researchgate.net/publication/347131046_Music_and_Mental_Health
- https://www.researchgate.net/publication/318702182_EFFECT_OF_MUSIC_ON_HEALTH_MIND_MOOD_AND_EMOTION_A_SURVEY