**Data: Explain how you collected the data. Explain any data cleaning or data transformations you had to do. Explain how you dealt with missing data, if applicable. Present some statistics about the data using tables and/or graphs.**

The goal of this project is to predict people’s mental health based on their song preference. To conduct this study, a google form survey was sent out asking people to list three songs that they listen to currently. From these 3 songs, song attributes such as energy factor, danceability factor, valence, tempo, popularity, instrumentalness, acousticness, and liveness of the song. The definition of the song attributes are shown in Table … In addition, the survey asked people five questions in relation to their mental health. Users rated their ability to enjoy life, resilience, balance in their lifestyle, emotional flexibility and self-actualization using a Likert scale. The summation of their ratings was used to determine the person’s mental health. To control the mental health score, the subjects were asked if they have recently experienced any out of the ordinary events. All of the questions required an answer in order to submit the survey.

|  |  |  |  |
| --- | --- | --- | --- |
| **Mental Health Score** | **Number** **of people** | **Categorical Mental Health** | **Percentage of Dataset (%)** |
| 0-5 | 1 | Low-Health | 0.4% |
| 5-10 | 21 | Low-Health | 8.2% |
| 10-15 | 46 | Low-Health | 18.0% |
| 15-20 | 92 | Medium-Health | 36.1% |
| 20-25 | 80 | High-Health | 31.4% |
| 25-30 | 15 | High-Health | 5.9% |

The survey was given out twice as not enough data was collected the first round of survey distribution. The survey was advertised by group members to several universities, families, and reddit. A total of 343 people filled out the survey. Once all the data was collected, an “Identification” column was made to keep track of each record in the dataset. All of the song names had to be formatted by placing the song name followed by the artist name. For example, “Take Care Drake”. There were 1012 songs that were parsed using Microsoft Excel functions or by manually formatting the song. Songs that could not be formatted due to invalid or inappropriate inputs were left blank.

Song attributes mentioned above were then collected for each song using Spotify’s API. Out of 1012 songs, Spotify was unable to find song attributes for 160 songs. This was due misspelled artist name or song name, song did not exist in Spotify’s database, or the lyrics of the songs were inputted as the song name by the person. To further clean the 160 songs, each song was searched on Google. Google would then output the correct song name or artist name which would be copied and pasted in a Microsoft Excel document with their respective identification number. For example, an incorrect input such as “Work Rhiinaa” was searched on Google and Google outputted “Work Rihanna” which is the correct spelling. Spotify’s API was used again to attempt to get the song attributes for the remaining 160 songs. However, song attributes for 53 songs were not collected due to data entry errors or existents of the song in Spotify’s database.

Out of 343 people, 88 people were missing at least 1 song. These 83 records were scrapped from the dataset. Therefore, there were a total of 255 clean records, each with song attributes for 3 individual songs. The average of each of the song attribute acquired using Spotify’s API for each person was taken. For example, record “1” tempo song attribute has values of 100, 120 and 130. The tempo value is changed to 116.67 which is the average of the three song tempos.

|  |  |  |
| --- | --- | --- |
| **Song Attribute** | **Definition** | **Average (based of 255 records)** |
| Energy | Factor between 0-1 that measures the intensity and activity in the track [2]. High energy tracks have high intensity [2]. | 0.656 |
| Danceability | Measures the ability to dance based on various track attributes [2]. Factor between 0-1, where 1 means the track has a high danceability [2]. | 0.597 |
| Valence | Factor between 0-1 that measures the musical positiveness of the track [2]. | 0.456 |
| Tempo | Beats per minute of the track [2]. | 123.5 |
| Popularity | Measure between 0-100 of the tracks popularity based on number of plays and year [2]. | 61.6 |
| Instrumentalness | Factor from 0-1 which measures the vocals present in a track [2]. No vocals has an instrumentalness factor of 0. | 0.061 |
| Acousticness | Confidence measure from 0-1 if the track is acoustic [2]. The more acoustic the track, the higher the measure. | 0.222 |
| Liveness | Confidence measure from 0-1 if the track is played live [2]. Detects the presence of audience in the track [2]. | 0.184 |

All of the song attributes and mental health ratings are numeric variables. In order to run association algorithms, the data had to be changed to categorical. Factors such as energy, danceability, valence, instrumentalness, acousticness, and liveness were categorised based on their averages. If a record had an energy factor less than or equal to the average, the numeric number was changed to “Low-Energy”, else it would be “High-Energy”. For tempo, the attribute was divided into four categories: “Fast-Tempo”, “Medium-Fast-Tempo”, “Medium-Slow-Tempo” and “Slow-Tempo”. This was completed by viewing the distribution and average of the tempos attribute for all 255 records. Similarly, popularity was divided into three categories by evening out the distribution of the popularity attribute. The three categories were “High-Popularity”, “Medium-Popularity” and “Low-Popularity”. The distribution of tyempo and popularity categorical variables are shown in Table…

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tempo** | | | | **Popularity** | | |
| **Fast-Tempo** | **Medium-Fast-Tempo** | **Medium-Slow-Tempo** | **Slow-Tempo** | **High-Popularity** | **Medium-Popularity** | **Low-Popularity** |
| 86 | 100 | 64 | 5 | 54 | 136 | 65 |
| 33.7% | 39.2% | 25.1% | 2.0% | 21.2% | 53.3% | 25.5% |

General information about the people who filled out the survey is shown in Table…

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Survey Questions** | **Gender** | | | **Age Range** | | | **Out of the Ordinary** | | **Hours Listening to Music (per day)** | | |
| **Categories** | **Male** | **Female** | **Not Specified** | **17 & Under** | **18-30** | **30+** | **Yes** | **No** | **2+ hours** | **1-2 hours** | **0 - 1 hours** |
| **Percentage** | 41.6% | 57.6% | 0.8% | 8.6% | 87.5% | 3.9% | 35.7% | 64.3% | 41.6% | 34.1% | 24.3% |
| **Number of Records** | 106 | 147 | 2 | 22 | 223 | 10 | 91 | 164 | 106 | 87 | 62 |

**Association**

The Apriori algorithm was used for association rule mining. This algorithm was chosen as it is useful for finding patterns in the data along with analyzing the relationships between various attributes of the data. Since there were only 255 unique records, a minimum support of ten percent was chosen so that each rule will cover at least 25 records. If there was a much larger dataset, a lower support would have been chosen. For example, for a million records, a support of one percent would be chosen as it covers 10,000 records. A support of 10-20 percent is a reasonable assumption for the current dataset.

Furthermore, a high confidence level of 90% was chosen initially and then decreased to see the additional association rules with the categorical variable mental health as a consequent. An association rule is similar to an if statement. An antecedent (if) is the item(s) found in the data and a consequent (then) is an item found in combination with the antecedent [1]. The effectiveness of the Apriori algorithm on the dataset was determined based on the confidence and support values. If the dataset has rules that have a high support and high confidence, it means that there are pattern in the data associating song attributes with a person mental health.

Various combinations of support and confidence level were tested in order to check for the validity of the Apriori algorithm on the dataset. This is seen in the table below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Support** | **%** | **10%** | | | **15%** | | | **20%** | | |
| **Confidence** | **%** | **80%** | **70%** | **40%** | **80%** | **70%** | **40%** | **80%** | **70%** | **50%** |
| **# of Rules Associated with Mental Health Categorical Variable as a Consequent** | | 0 | 0 | 65 | 0 | 0 | 11 | 0 | 0 | 0 |

First, a support of 20% and a confidence of 80%, 70% and 50% was used for the Apriori algorithm. However, there were zero rules associated with mental health as a consequent in the rule. This the support and confidence was lowered accordingly to analyze the relationship between mental health and various song attributes. The result shows that patterns in the dataset involving mental health were only derived by the algorithm when the confidence and support percentage was dropped to 40% and 15% respectively. This illustrates that the generated association rules are not good for predicting mental health of a person given song attributes, age, and gender of the person.

However, the algorithm did provide rules that adequately associated song attributes with each other. For example, if a song has a high popularity, then it has a low instrumentalness factor. This rule is consistent with the real world as the billboard top 100 chart almost never have any instrumental tracks.

There are several reasons as to why the Apriori algorithm did not generate good association rules based on the dataset. One factor could be that music may not be a dominant influence on a person’s mental health leading to poor association rules. Another factor could be the lack of song data. A total of 3 songs were collected per person and the song attributes were averaged for each person. If a single record consisted of a playlist of songs instead of 3 songs, the averaged song attributes may have given better associations between the song attributes and the person’s mental health. Finally, lack of data may also be a result of poor association rules.

[1] <http://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining>

[2] https://developer.spotify.com/web-api/get-audio-features/