SL Project

This is a suggested approach that ChatGPT gave us (not bad at all)

1. Define the Research Questions

Clearly state the research questions you want to answer through your analysis. For example:

- How do the Spotify statistics (e.g., streams, audio features) of a song influence its probability of winning a Grammy?
- How do the YouTube statistics (e.g., views) of a song influence its probability of winning a Grammy?

2. Data Collection

3. Data Preparation

Once you have collected the data, you'll need to prepare it for analysis. This step may involve cleaning the data, handling missing values, and transforming the data into a suitable format for analysis. Ensure that the data from both Spotify and YouTube are aligned with each other and can be matched based on common identifiers (e.g., song title, artist, year).

We start of by reading the CSV files we have previously downloaded.

```
sp_y <- read.csv("Spotify_Youtube.csv", sep=',')
sp_y$unique <- paste(sp_y$Artist, "_", sp_y$Track)
grammy <- read.csv("grammySongs.csv", sep=',')
grammy$unique <- paste(grammy$Artist, "_", grammy$Name)
```

We only need to know if the song won an award or not, no matter which one it is. That is why we only keep the field 'unique' from the grammy dataframe and we remove all the duplicates.

```
grammy <- grammy[c('unique')]
grammy <- grammy[!duplicated(grammy), ]
```

We create a new column 'winner' in the sp__y dataframe. This is a boolean type column which value is true if the song has been awarded a grammy.

```
sp_y$winner <- ifelse(sp_y$unique %in% grammy, TRUE , FALSE)
```

```
colnames(sp_y)
```

```
##  [1] "X"                "Artist"           "Url_spotify"      "Track"
##  [5] "Album"            "Album_type"       "Uri"              "Danceability"
##  [9] "Energy"           "Key"              "Loudness"         "Speechiness"
## [13] "Acousticness"     "Instrumentalness" "Liveness"         "Valence"
## [17] "Tempo"            "Duration_ms"      "Url_youtube"      "Title"
## [21] "Channel"          "Views"            "Likes"            "Comments"
## [25] "Description"      "Licensed"         "official_video"   "Stream"
## [29] "unique"           "winner"
```

Then we get rid of all the columns that do not provide any information.

4. Exploratory Data Analysis

Perform exploratory data analysis to understand the characteristics of your data. This step will help you identify any patterns or insights that may guide your analysis. Visualize the data, calculate summary statistics, and look for correlations between different variables.

5. Feature Selection

Determine which variables from both the Spotify and YouTube datasets are relevant for your analysis. You can use techniques such as correlation analysis or feature importance methods to identify the most influential features.

6. Model Development

Build a predictive model to assess the influence of Spotify and YouTube statistics on the probability of winning a Grammy. You can use techniques such as logistic regression, random forests, or gradient boosting algorithms. The Grammy win (1 or 0) will be the target variable, and the relevant features from both datasets will be the input variables.

7. Model Evaluation

Evaluate the performance of your model using appropriate evaluation metrics such as accuracy, precision, recall, or area under the ROC curve. Cross-validation techniques can help ensure the robustness of your results.

8. Interpretation and Conclusion

Interpret the results of your analysis and draw conclusions based on your findings. Discuss the significance and influence of Spotify and YouTube statistics on the probability of winning a Grammy.

9. Documentation and Visualization

Document your project, including the steps you followed, the data sources, the analysis techniques used, and the results obtained. Use visualizations such as charts, graphs, or tables to present your findings effectively.