HITPREDICT: PREDICTING BILLBOARD HITS USING SPOTIFY

01

INTRODUCTION:

The Billboard Hot 100 is a chart that ranks the best-performing singles of the United States. Its data, published by Billboard magazine and compiled by Nielsen SoundScan, is based collectively on each single's weekly physical and digital sales,as well as airplay and streaming. At the end of a year, Billboard will publish an annual list of the 100 most successful songs throughout that year on the Hot 100 chart based on the information.
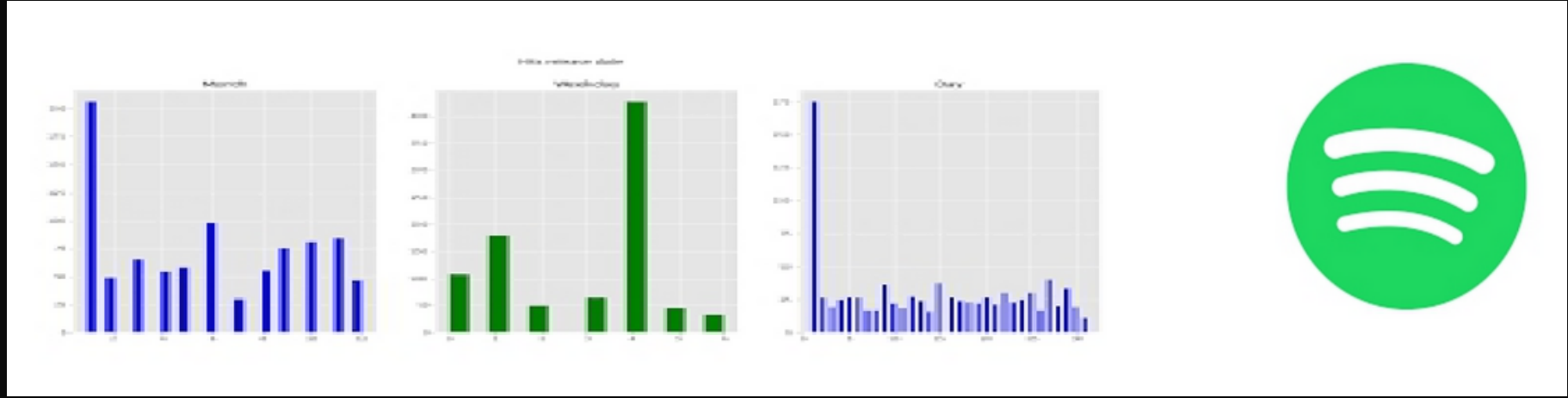
INTRODUCTION:

```
billboard = pd.read_csv('billboardtop.csv', parse_dates=["date"])
billboard.head()
```

]:

	id	date	artist	song	weeks	link
0	1041	2015-01-17	Mark Ronson featuring Bruno Mars	"Uptown Funk"♪ [70]	14	[71]
1	1042	2015-04-25	Wiz Khalifa featuring Charlie Puth	"See You Again"	12	[72]
2	1043	2015-06-06	Taylor Swift featuring Kendrick Lamar	"Bad Blood"	1	[73]
3	1044	2015-07-25	Omi	"Cheerleader"	6	[74]
4	1045	2015-08-22	The Weeknd	"Can't Feel My Face"	3	[75]

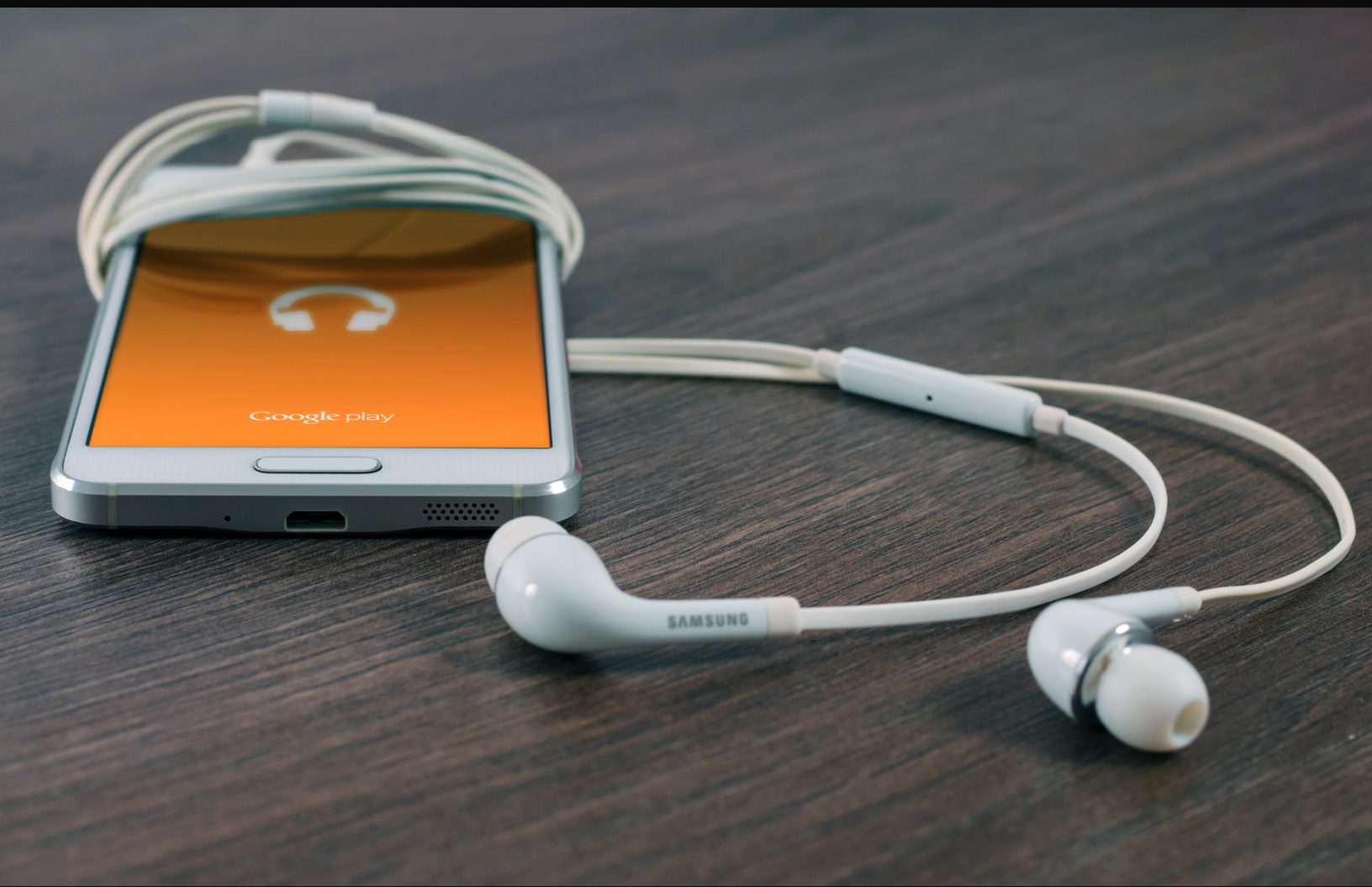# FEATURES AND DATA METHODS, ALGORITHMS RESULTS:

IMPORT LIBRARIES:

```python
# Load necessary libraries

import os
import numpy as np # numerical calc package
import pandas as pd # holds data
import matplotlib.pyplot as plt # plotting library
import seaborn as sns # pretty plotting
import warnings
warnings.filterwarnings('ignore')

# plotting config
sns.set(style='white', rc={'figure.figsize':(20,10)})
```

DATA PREPARATION AND ANALYSIS:

DATA PREPATION IS THE PROCESS PROCESS OF CLEANING AND TRANSFORMING RAW DATA PRIOR TO PROCESSING AND ANALYSIS.

Listen on all devices.

DATA CLEANING:

DATA CLEANING IS THE PROCEES OF FIXING OR REMOVING INCORRECT, CORRUPTED, INCORRECTLY FORMATTED, DUPLICATE, OR INCOMPLETE DATA WITHIN A DATASET.

Put your favorite beats on loop

Create and share as many playlists as you like

Listen to playlists made for you based on your streaming history

FEATURE ENGINEERING:

FEATURE ENGINEERING IS THE PROCESS OF USING DOMAIN KNOWLEDGE TO EXTRACT FEATURES FROM RAW DATA.

Download our mobile app.

Stream music via our website.

Plug in Audios Co. Streaming App on your smart TV

SELECTING THRESHOLD FOR CLASSIFICATION:

THE CLASSIFICATION THRESHOLD IS SIMPLY SET TO 0.5, WHICH IS USUALLY UNSUITABLE FOR AN IMBALANCED CLASSIFICATION. IN THIS STUDY WE ANALYSE THE DRAWBACKS OF USING ROC AS THE SOLE MEASURE OF IMBALANCE IN DATA CLASSIFICATION PROBLEMS.

BEST MUSIC EXPERIENCE.

FINAL MODEL AND RESULTS:

```
>>> print(classification_report(y_true, y_pred, target_names=target_names))
                precision    recall  f1-score   support
<BLANKLINE>
    class 0          0.50      1.00      0.67         1
    class 1          0.00      0.00      0.00         1
    class 2          1.00      0.67      0.80         3
<BLANKLINE>
    accuracy                             0.60         5
   macro avg         0.50      0.56      0.49         5
weighted avg         0.70      0.60      0.61         5
```