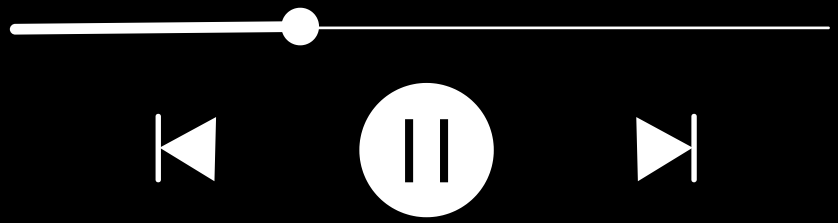
SC1015 Mini Project:

Predicting Song Popularity



CHUA JING JIE, JUSTIN
JESSIE ANG KAI PIN
LIM ZHE XUN
DSF 3, GROUP 10

SAMPLE COLLECTION

PRACTICAL MOTIVATION

EXISTING STATISTICS

- 60,000 songs are uploaded to Spotify each day
- > 365 million monthly users
- > 11 million artists on the platform
- Spotify pays $0.033 - $0.054 per stream

SAMPLE COLLECTION

PRACTICAL MOTIVATION

EXISTING STATISTICS

- 0.7% artists on Spotify earn 90% of the total revenue
- Spotify pays $0.033 - $0.054 per stream

SAMPLE COLLECTION

PRACTICAL MOTIVATION

DATASET:

Spotify's Million Playlist Dataset
Over 2 million unique songs, from 300,000 artists

VARIABLES FROM SPOTIFY API:

Danceability	Energy	Key	Loudness
Mode	Speechiness	Acousticness	Instrumentalness
Liveness	Valence	Tempo	Type
Duration_ms	Time_signature	Popularity	Artist genre

DEFINING POPULARITY (WITHIN DATASET)

- based on the total number of plays the track has had and how recent those plays are according to Spotify
- discrete numerical score from 0-100

SAMPLE COLLECTION

PRACTICAL MOTIVATION

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	type	duration_ms	time_signature	name	popularity	artist	genre
0	0.723	0.809	7	-3.081	0	0.0625	0.00346	0.00123	0.565	0.274	98.007	audio_features	176561	4	Lean On (feat. MØ & DJ Snake)	0	Major Lazer	dance pop/edm/elec
5	0.642	0.734	2	-3.924	1	0.03	0.0759	0.0	0.0937	0.903	155.031	audio_features	171707	4	Cut Me Some Slack	0	Chris Janson	contemporary count
0	0.419	0.793	2	-6.131	0	0.125	0.00857	0.0357	0.414	0.0712	127.944	audio_features	406907	4	Raining (feat. SunSun) - Dance Love Edit	0	Kaskade	edm/electro house/p
47	0.622	0.735	11	-7.538	0	0.0408	0.225	0.464	0.138	0.117	126.979	audio_features	219253	3	4 AM - Adam K & Soha Radio Edit	0	Kaskade	edm/electro house/p
8	0.615	0.758	6	-5.024	1	0.032	0.0667	0.0	0.199	0.541	106.904	audio_features	183760	4	Georgia Clay	0	Josh Kelley	acoustic pop/indiec
46	0.63	0.777	2	-5.787	0	0.0273	0.0142	8.71E-05	0.0827	0.52	124.985	audio_features	281053	4	Room for Happiness (feat. Skylar Grey) (feat. Skylar Grey)	0	Kaskade	edm/electro house/p
10	0.712	0.568	9	-8.295	1	0.031	0.136	0.000364	0.961	0.522	104.474	audio_features	285653	4	Oklahoma Breakdown	0	Stoney LaRue	heartland rock/oklah
45	0.814	0.604	2	-5.793	0	0.0653	0.0622	1.62E-05	0.0942	0.741	128.011	audio_features	159907	4	Fire in Your New Shoes	0	Kaskade	edm/electro house/p
44	0.709	0.677	6	-7.194	0	0.0383	0.0654	0.00318	0.239	0.359	128.013	audio_features	276333	4	Don't Stop Dancing	0	Kaskade	edm/electro house/p
43	0.573	0.631	9	-7.521	0	0.0414	0.0038	2.5E-06	0.102	0.157	125.972	audio_features	227453	4	Angel On My Shoulder	0	Kaskade	edm/electro house/p

VARIABLES FROM DATASET:

Danceability, Energy, Key, Loudness, Mode, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo, Type, Duration_ms, Time_signature, Name, Popularity, Artist genre

SAMPLE COLLECTION

PRACTICAL MOTIVATION

POPULARITY == REVENUE

WILL A SONG BE WELL RECIEVED?

WHAT SONGS TO PUT IN AN ALBUM?

DATA PREPARATION

PROBLEM FORMULATION

IS IT POSSIBLE TO PREDICT IF A SONG IS POPULAR?

EXPLORATORY DATA ANALYSIS

EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION

BASIC DATA CLEANING:

```
RangeIndex: 63032 entries, 0 to 63031
Data columns (total 16 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   danceability      63032 non-null  float64
 1   energy            63032 non-null  float64
 2   key               63032 non-null  int64
 3   loudness          63032 non-null  float64
 4   mode              63032 non-null  int64
 5   speechiness       63032 non-null  float64
 6   acousticness      63032 non-null  float64
 7   instrumentalness  63032 non-null  float64
 8   liveness          63032 non-null  float64
 9   valence           63032 non-null  float64
 10  tempo             63032 non-null  float64
 11  duration_ms       63032 non-null  int64
 12  name              63031 non-null  object
 13  popularity        63032 non-null  int64
 14  artist            63032 non-null  object
 15  genre             58497 non-null  object
dtypes: float64(9), int64(4), object(3)
memory usage: 7.7+ MB
There are 63032 unique titles, 16964 unique artists, and 9339 unique genres
```

Dropping 'type' & 'time_signature' columns and N/A & duplicate values

EXPLORATORY ANALYSIS

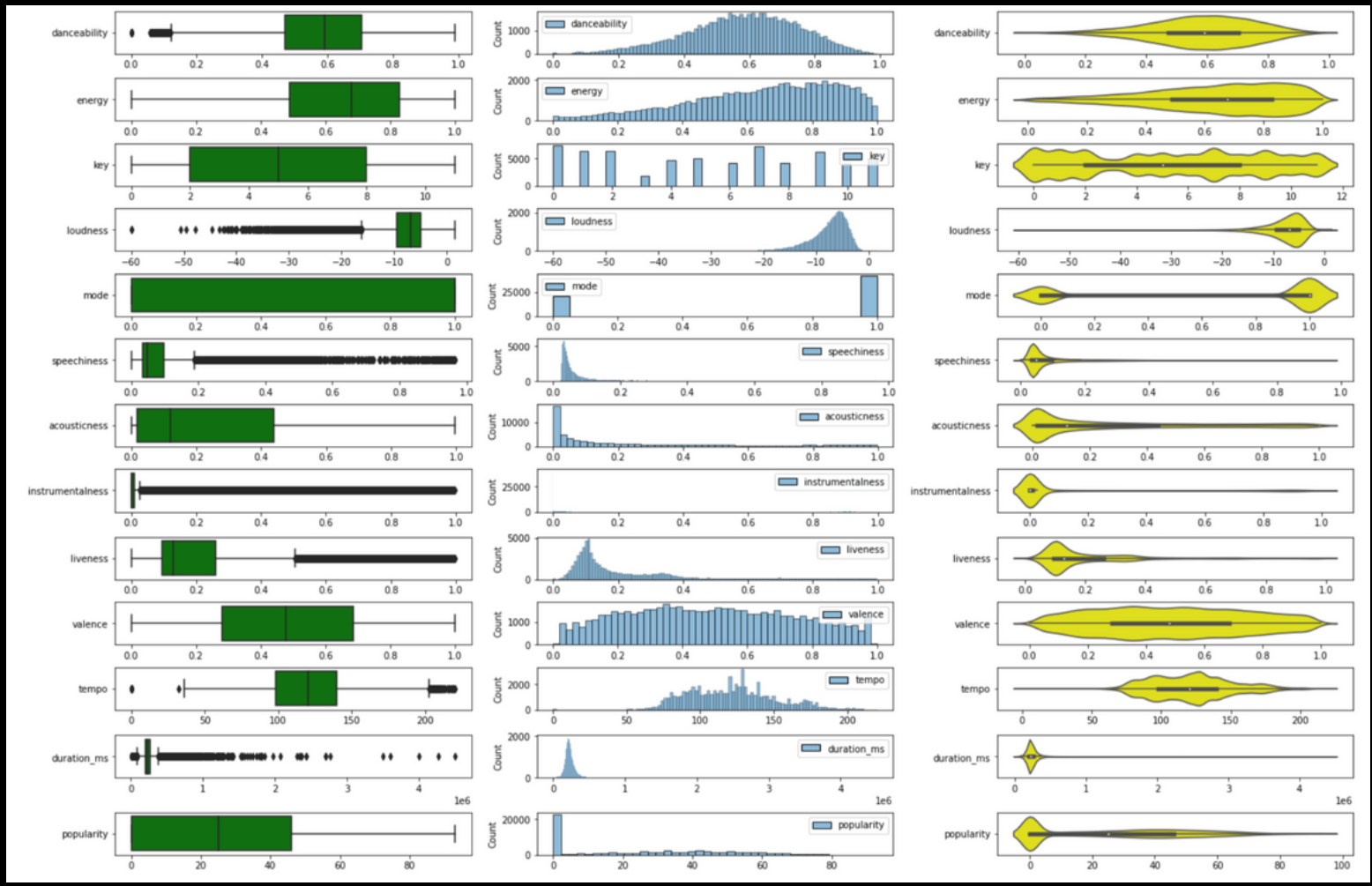
STATISTICAL DESCRIPTION

Numerical	Categorical
Danceability Energy Key Loudness Mode Speechiness Acousticness Instrumentalness Liveness Valence Tempo Duration_ms Popularity	Artist Genre

EXPLORATORY ANALYSIS

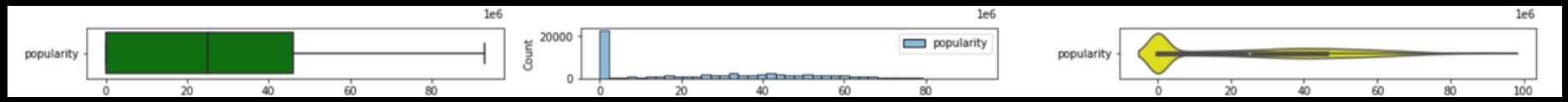
STATISTICAL DESCRIPTION

BASIC ANALYSIS (NUMERICAL)

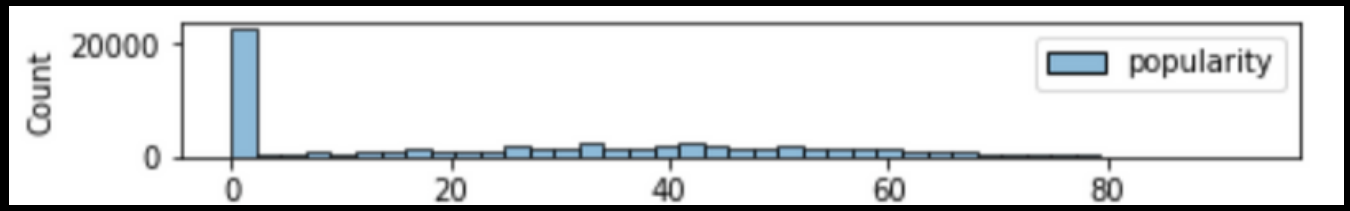


EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION



Disproportionate number of tracks with 0 popularity



0.0712	127.944	audio_features	406907	4	Raining (feat. SunSun) - Dance Love Edit	0	Kaskade
0.1170	126.979	audio_features	219253	3	4 AM - Adam K & Soha Radio Edit	0	Kaskade

ADVANCED DATA CLEANING (NUMERICAL)

- Remove data with 0 popularity
- Normalise the data using min-max so each numerical value is between 0-1

EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION

BASIC ANALYSIS (CATEGORICAL)

ARTIST

```
Streets                    157
Eminem                     136
Pearl Jam                  128
Dave Matthews Band         103
Rascal Flatts              102
                           ...
Los Leones Del Norte         1
Alagoas                      1
The Low Life                 1
Boston Pops Orchestra        1
Offset                       1
Name: artist, Length: 16964, dtype: int64
```

~17000 unique 'Artist' values

EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION

BASIC ANALYSIS (CATEGORICAL)

GENRE

```
contemporary country/country/country road                                    541
contemporary country/country/country road/modern country rock                406
dance pop/pop/post-teen pop                                                  376
dance pop/pop                                                                186
beats                                                                        172
grunge/permanent wave/rock                                                   159
adult standards/easy listening/lounge                                        155
modern rock/rock                                                             152
ccm/christian alternative rock/christian music/world worship/worship         149
k-pop/k-pop boy group                                                        145
reggae fusion                                                                143
latin/latin hip hop/reggaeton/trap latino                                    137
detroit hip hop/hip hop/rap                                                  136
ccm/christian alternative rock/christian music/worship                       135
christian hip hop/christian trap                                             125
gospel                                                                       116
atl hip hop/dirty south rap/gangster rap/hip hop/pop rap/rap/southern hip hop/trap  114
latin/latin hip hop/reggaeton/reggaeton flow/trap latino                     113
indie pop rap/pop rap                                                        111
jam band/neo mellow/pop rock                                                 103
Name: genre, dtype: int64
9338
```

before

```
['classic rock',
 'post-teen pop',
 'electro house',
 'electropop',
 'stomp and holler',
 'pop dance',
 'urban contemporary',
 'alternative rock',
 'indie rock',
 'gangster rap',
 'country road',
 'r&b',
 'contemporary country',
 'southern hip hop',
 'country',
 'trap',
 'edm',
 'modern rock',
 'pop rock',
 'hip hop',
 'pop rap',
 'rap',
 'rock',
 'dance pop',
 'pop']
```

after

EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION

ADVANCED DATA CLEANING (CATEGORY)

- Replace 'na' with 'no genre'
- Took top 25 genres as they represent 50% of the dataset

```
POP                      3528
ROCK                     3325
ALTERNATIVE ROCK         2331
HIP HOP                  1918
COUNTRY ROAD             1727
GANGSTER RAP             1521
URBAN CONTEMPORARY       1412
CLASSIC ROCK             1409
POST-TEEN POP            1350
TRAP                     1287
STOMP AND HOLLER         1137
ELECTRO HOUSE            1077
ELECTROPOP               1073
POP ROCK                 1009
R&B                       883
SOUTHERN HIP HOP          782
RAP                       771
COUNTRY                   717
MODERN ROCK               583
INDIE ROCK                384
DANCE POP                 346
POP RAP                   339
POP DANCE                 334
EDM                       291
CONTEMPORARY COUNTRY      238
Name: genre, dtype: int64
```

EXPLORATORY DATA ANALYSIS

```
Int64Index: 29772 entries, 22122 to 63031
Data columns (total 17 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   danceability      29772 non-null  float64
 1   energy            29772 non-null  float64
 2   key               29772 non-null  float64
 3   loudness          29772 non-null  float64
 4   mode              29772 non-null  float64
 5   speechiness       29772 non-null  float64
 6   acousticness      29772 non-null  float64
 7   instrumentalness  29772 non-null  float64
 8   liveness          29772 non-null  float64
 9   valence           29772 non-null  float64
 10  tempo             29772 non-null  float64
 11  duration_ms       29772 non-null  float64
 12  name              29772 non-null  object
 13  popularity        29772 non-null  float64
 14  artist            29772 non-null  object
 15  genre             29772 non-null  object
 16  pop               29772 non-null  category
dtypes: category(1), float64(13), object(3)
memory usage: 3.9+ MB
```

- Post cleaning
  - 29772 songs
  - 17 columns

- Data was split into 2 classes: "Pop" (>0.9) and "Not Pop" (<0.9)

EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION

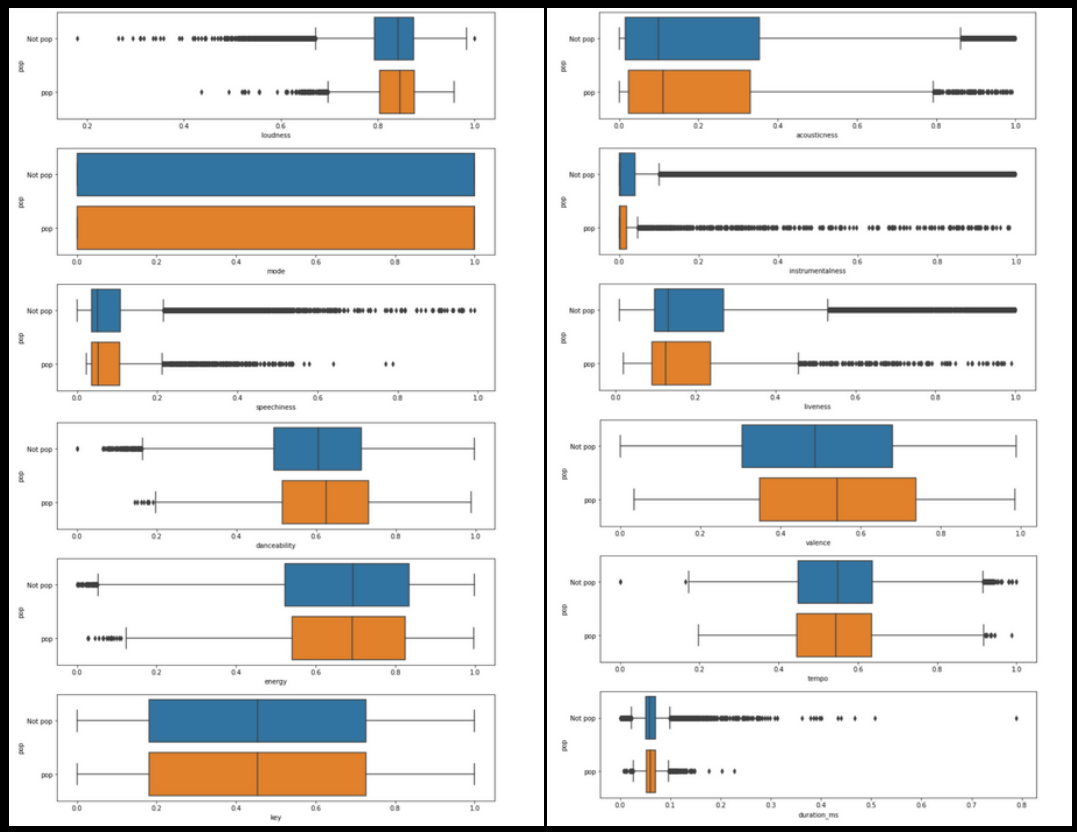
Relationship Visualisation between variables and popularity (heatmap)

Weak correlation coefficients for popularity

EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION

Relationship Visualisation between variables and popularity (box plot)

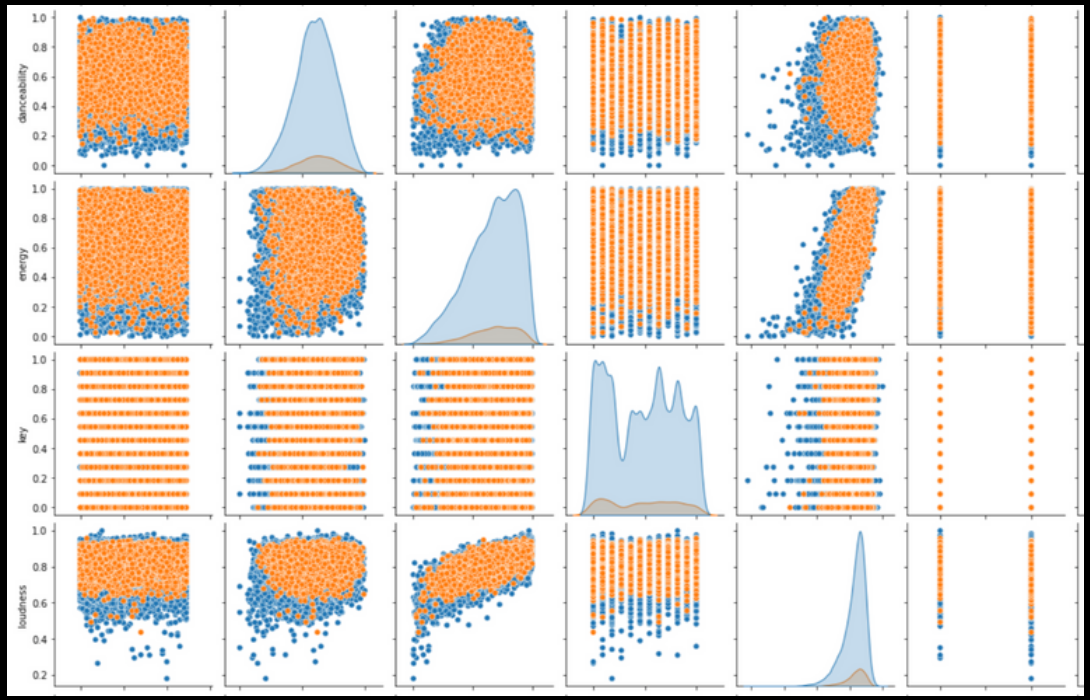


Hard to distinguish between the 2 classes

EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION

Relationship Visualisation between variables and popularity (pair plot)



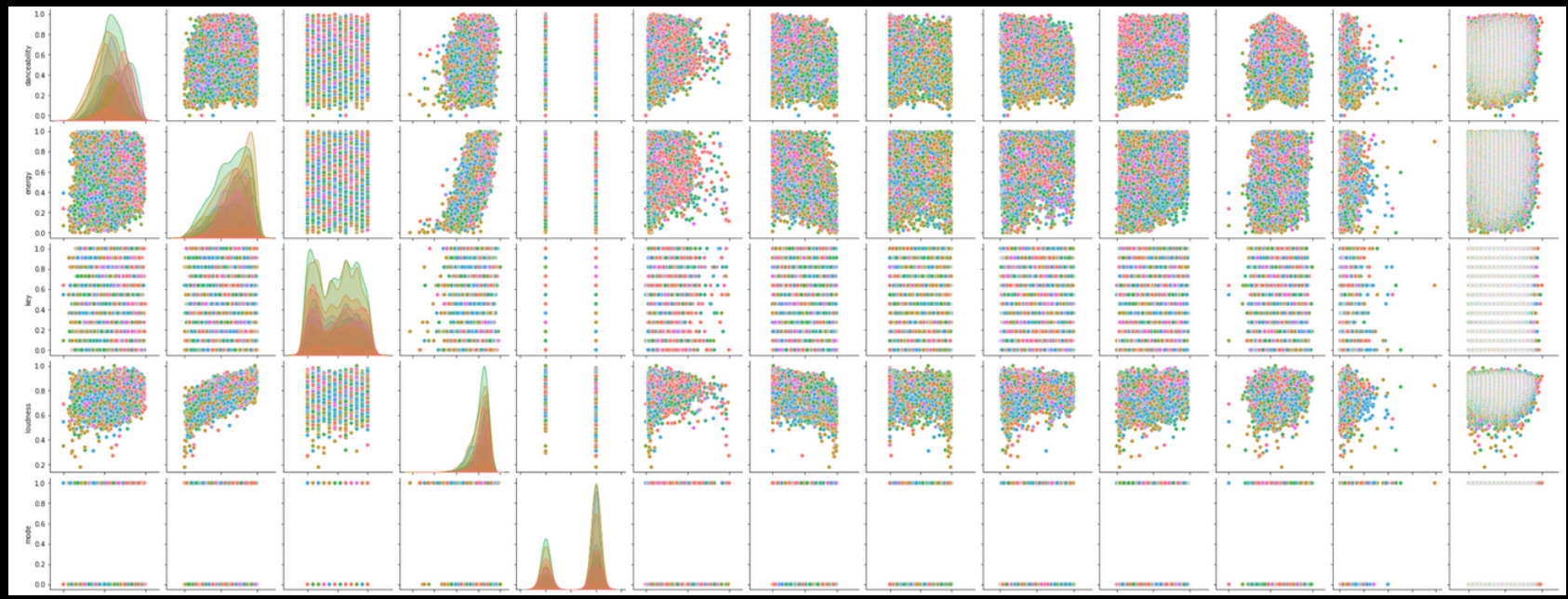
- Clusters are not clearly separated
- Relationships not clearly distinguished

EXPLORATORY ANALYSIS

STATISTICAL DESCRIPTION

Relationship Visualisation between genres and popularity (pair plot)

(snipet of pairplot)



messy

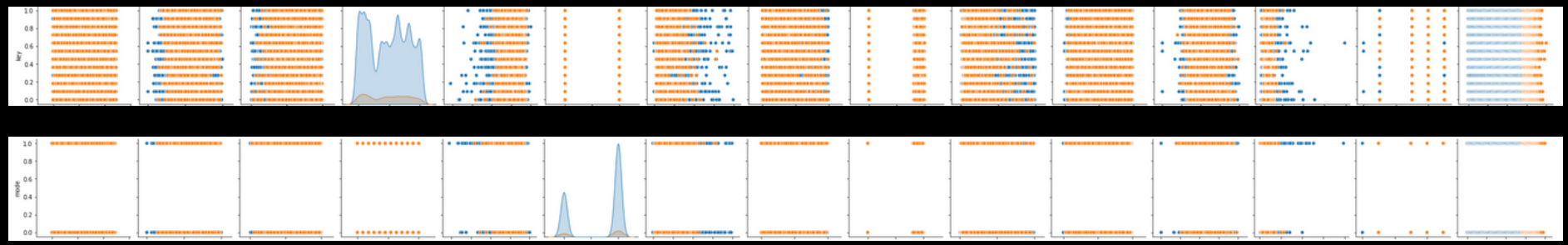
same as other variables

ANALYTIC VISUALISATION

PATTERN RECOGNITION

Relationship Visualisation between variables and popularity (pair plot)

'key' and 'mode' have extremely poor pairplots



ANALYTIC VISUALISATION

PATTERN RECOGNITION

Features Selected

```
['danceability',
 'energy',
 'loudness',
 'speechiness',
 'acousticness',
 'instrumentalness',
 'liveness',
 'valence',
 'tempo',
 'duration_ms',
 'CLASSIC ROCK',
 'POST-TEEN POP',
 'ELECTRO HOUSE',
 'ELECTROPOP',
 'STOMP AND HOLLER',
 'POP DANCE',
 'URBAN CONTEMPORARY',
 'ALTERNATIVE ROCK',
 'INDIE ROCK',
 'GANGSTER RAP',
 'COUNTRY ROAD',
 'R&B',
 'CONTEMPORARY COUNTRY',
 'SOUTHERN HIP HOP',
 'COUNTRY',
 'TRAP',
 'EDM',
 'MODERN ROCK',
 'POP ROCK',
 'HIP HOP',
 'POP RAP',
 'RAP',
 'ROCK',
 'DANCE POP',
 'POP']
```

Danceability
Energy
Loudness
Speechiness
Acousticness
Instrumentalness
Liveness
Valence
Tempo
Duration_ms
+
top 25 genres

ANALYTIC
VISUALISATION

PATTERN
RECOGNITION

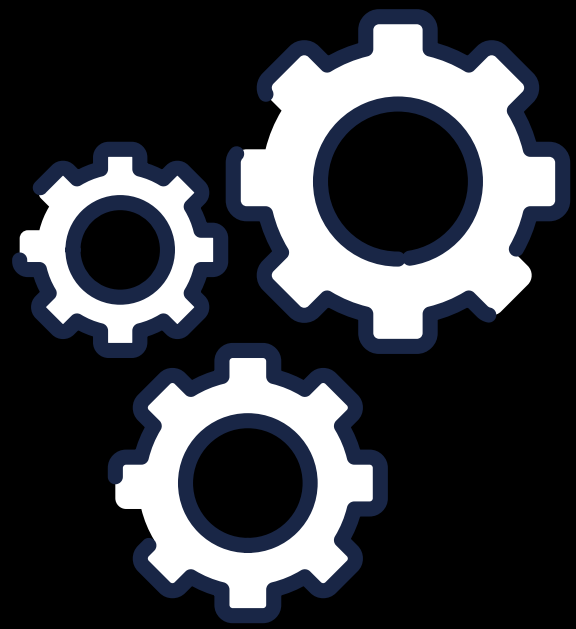
ONE HOT ENCODING

Preprocessing categorical data for Machine Learning

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	...	POP RAP	POP ROCK	POST-TEEN POP	R&B	RAP	ROCK	SOUTHERN HIP HOP
22122	0.920202	0.332	0.090909	0.827574	1.0	0.090021	0.075000	0.003697	0.105528	0.1020	...	0	0	0	0	0	0	0
22123	0.548485	0.766	0.090909	0.813395	1.0	0.168399	0.233936	0.001190	0.465327	0.1940	...	0	0	0	0	0	0	0
22127	0.518182	0.832	1.000000	0.895017	0.0	0.057069	0.000107	0.001378	0.154774	0.9080	...	0	0	0	0	0	0	0
22130	0.563636	0.326	0.363636	0.712922	0.0	0.031185	0.799197	0.085947	0.125628	0.0342	...	0	0	0	0	0	0	0
22129	0.839394	0.331	0.727273	0.645154	1.0	0.152807	0.098193	0.125213	0.117588	0.4970	...	0	0	0	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
63027	0.880808	0.391	0.000000	0.786987	0.0	0.251559	0.470884	0.002037	0.298492	0.4370	...	0	0	0	0	0	0	0
63028	0.532323	0.835	0.545455	0.865115	1.0	0.045010	0.016667	0.000000	0.250251	0.6540	...	0	0	0	0	0	0	0
63030	0.609091	0.783	0.545455	0.891015	1.0	0.064449	0.450803	0.002828	0.119598	0.7750	...	0	0	0	0	0	0	0
63029	0.958586	0.661	0.454545	0.886526	0.0	0.059459	0.030321	0.000000	0.045628	0.7600	...	0	0	0	0	0	0	0
63031	0.917172	0.669	0.636364	0.915311	1.0	0.076715	0.002871	0.000000	0.238191	0.6620	...	0	0	0	0	0	0	0

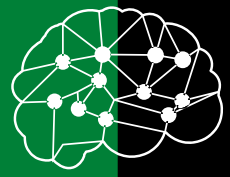
MACHINE LEARNING

1. DECISION TREE
2. RANDOM FOREST
3. K-NEAREST NEIGHBOUR
4. LOGISTIC REGRESSION*
5. GRADIENT BOOSTING*



* WILL NOT BE COVERED IN PRESENTATION

ALGORITHMIC OPTIMISATION



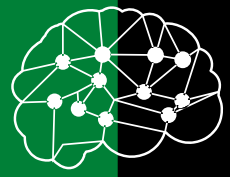
MACHINE LEARNING

SPLITTING DATASET

- Random split into train and test set

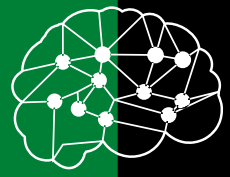
WHAT ARE WE LOOKING FOR?

- Balanced Accuracy Score
  - The arithmetic mean of sensitivity and specificity
  - Used to evaluate how good a classifier is



GRID SEARCH

Uses a different combination of all the specified hyperparameters and their values and calculates the performance for each combination and selects the best value for the hyperparameters

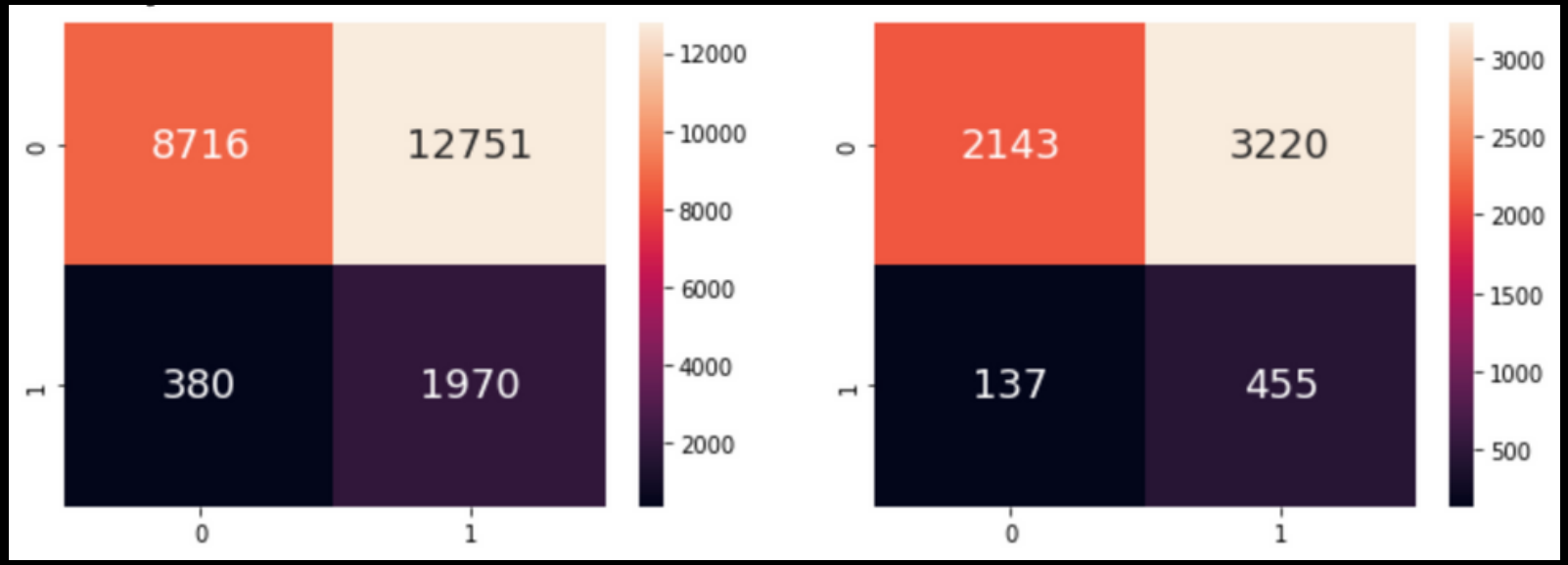


DECISION TREE

Supervised Machine Learning Algorithm where the data is continuously split according to a certain parameter (popularity)

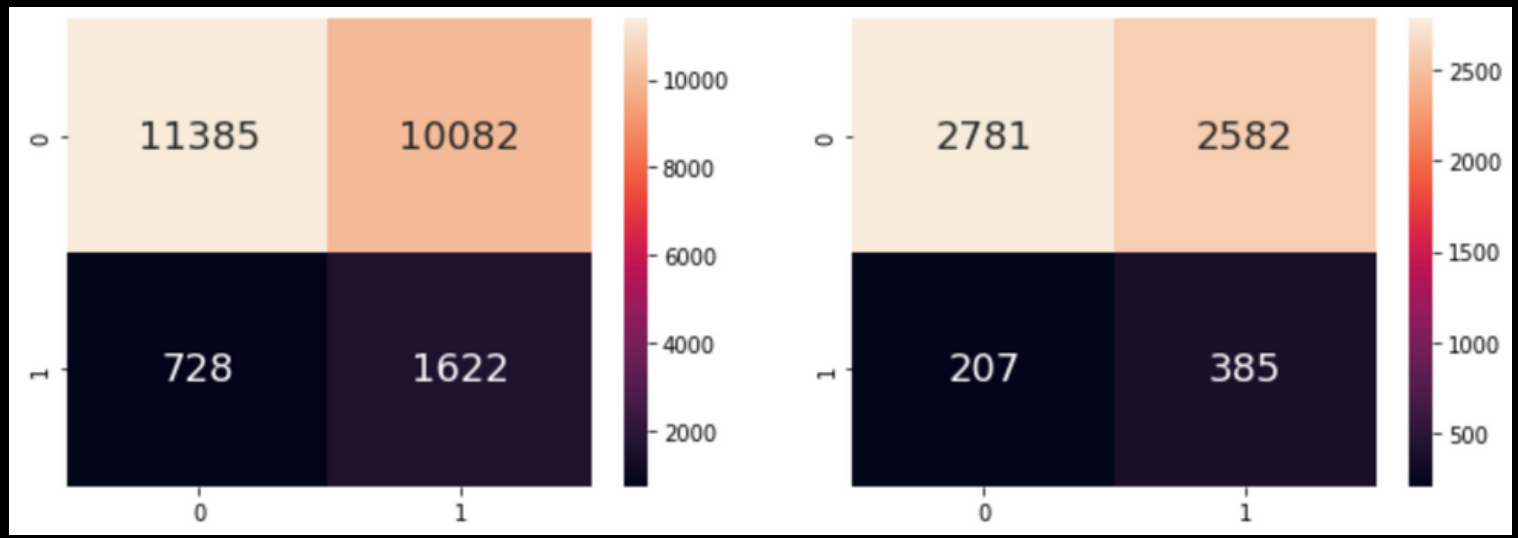
DECISION TREE

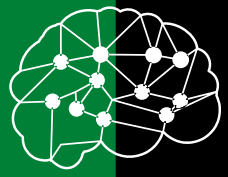
- Class_weight = "balanced" for imbalanced data
- Balanced accuracy score = 58.4%
- Decent TPR (76%) but poor TNR (39.9%)



DECISION TREE (POST FINE TUNING USING GRID SEARCH)

- TNR improved by 12%
- Balanced Accuracy Score remains 58.4%



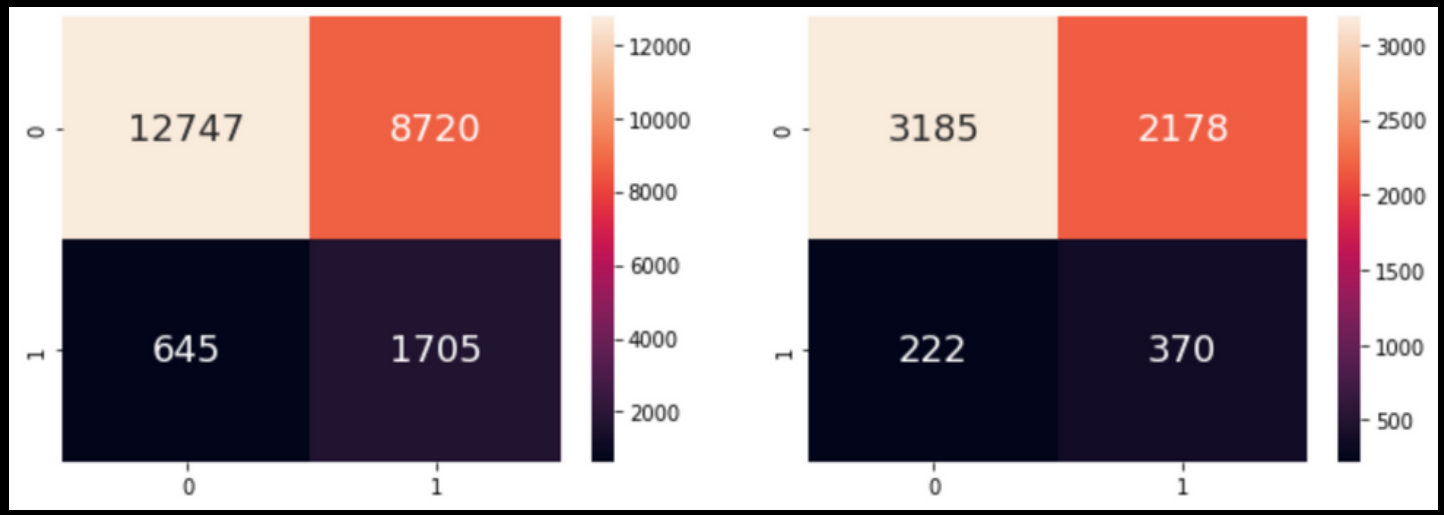


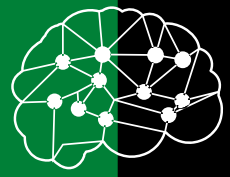
RANDOM FOREST

Supervised Machine Learning Algorithm that builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

RANDOM FOREST (POST FINE TUNING USING GRID SEARCH)

- Balanced Accuracy Score = 60.9% (slight improvement from decision tree but not very significant)



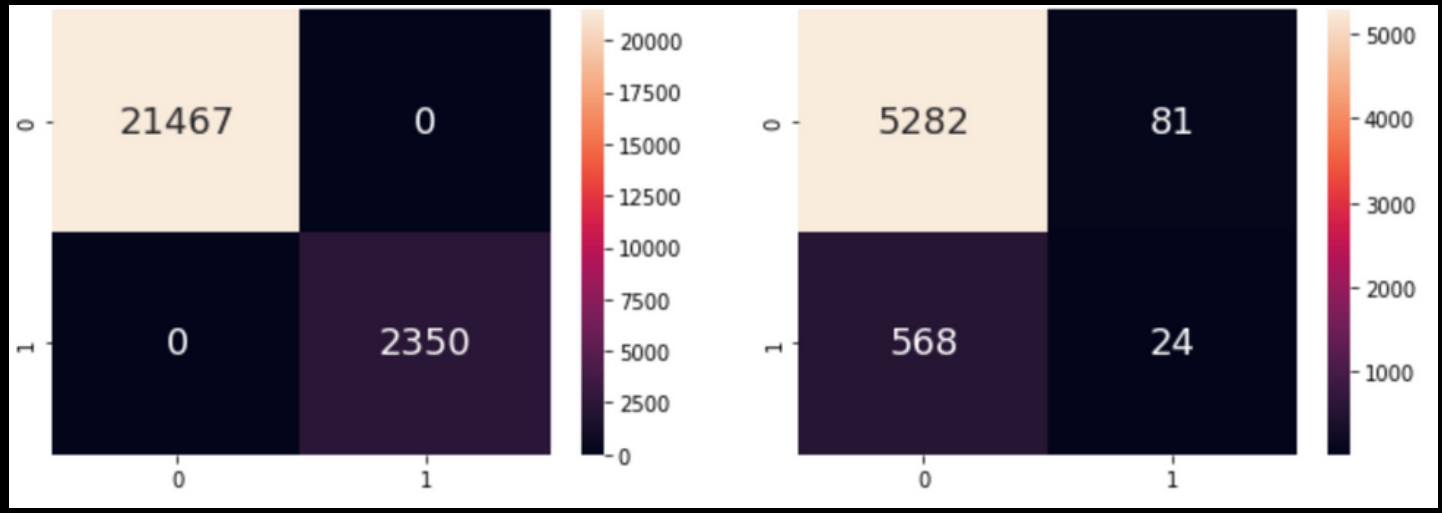


K-NEAREST NEIGHBOUR (KNN)

Supervised Machine Learning Algorithm that finds the distances between a query (popularity) and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label

K-NEAREST NEIGHBOUR (POST FINE TUNING USING GRID SEARCH)

- Balanced Accuracy Score = 51.2% (lowest of all the models)



INFORMATION PRESENTATION

STATISTICAL INFERENCE

WHY OUR MODELS WERE CHOSEN

- Appropriateness against our existing data set
- Accuracy and ability to achieve our goal (as per problem definition)

WHY BALANCED ACCURACY SCORE?

- Tackles the issue of imbalance by adjusting the weight of classes according to their frequency
- Appropriate to combat overfitting due to over-sampling

COMPARING RESULTS ACROSS MODELS

- Random Forest was the best model as it yielded the highest balance accuracy score of 60.9%

IMPROVEMENT OF RESULTS POST GRID SEARCH

- E.g. 1% increase for balance accuracy score of k-nearest neighbour

INSIGHTS

- OUTCOME
- SUBSEQUENT RECOMMENDATIONS
- TECHNIQUES LEARNT
- CONCLUSION

INFORMATION PRESENTATION

STATISTICAL INFERENCE

OUTCOME

- No strong correlation between popularity and any univariate variable
- Balance Accuracy Score remained < 60%
- Insufficient independent variables, weak dependent variable
- Large disproportion between number of popular songs and not popular songs

SUBSEQUENT RECOMMENDATIONS

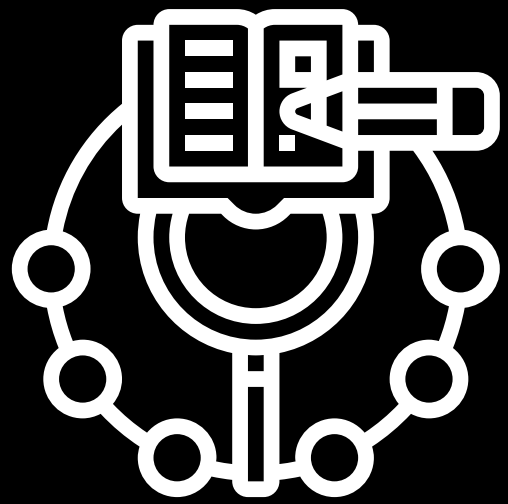
- Increase proportion of 'popular' songs to ensure dataset is balanced or adopt over-sampling / under-sampling
- Identify more possible variables
- Use songs' lifetime streams as popularity metric
- Break down popularity into more subsets, e.g. 'region', 'time period'

ETHICAL CONSIDERATION

INTELLIGENT DECISION

TECHNIQUES LEARNT

- One-hot encoding
- Random forest
- K-nearest neighbors (KNN)
- Gradient boosting



ETHICAL CONSIDERATION

INTELLIGENT DECISION

CONCLUSION / DATA DRIVEN INSIGHTS

Popularity can possibly be predicted. However, many additional variables need to be considered

- Existing trends in music
- Social media / Marketing
- Artist

Spotify

THANK YOU!