SPOTIFY SONG ANALYSIS

A CLASSIFICATION PROBLEM OF SONG POPULARITY

- SULAIMON OYELEYE

OUTLINE

- Motivation
- About the dataset
- Insights and Visualizations
- Model used
- Results & Comparison
- Model Evaluation on generated sample data
- Conclusion

MOTIVATION

XYZ Music Production company understands that making popular song is very important for its business to stay competitive.

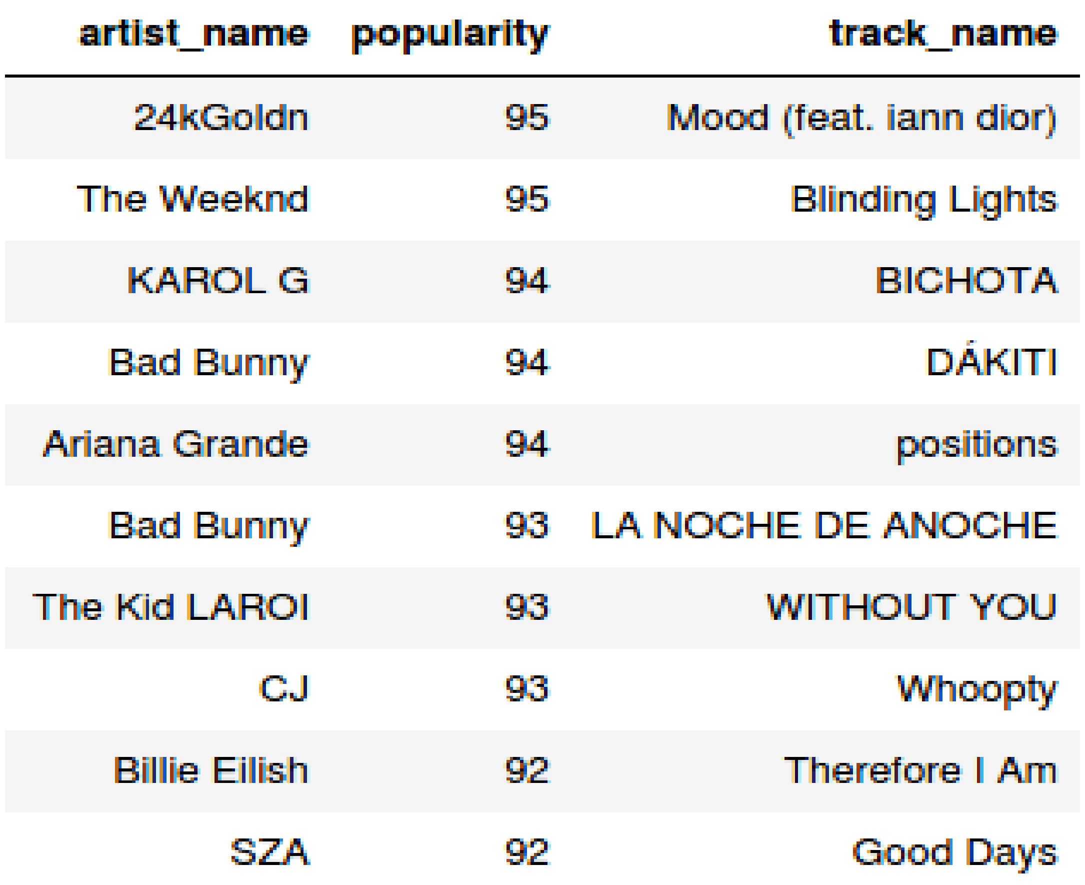
Hence, they have contacted the NLB Team to help in building a model to classify what kind of songs will be popular and figure out a way to make help the business stay competitive.

ABOUT THE DATASET

- Web scrap using spotipy api.
- Datasets gotten in 2 parts: the track IDs and audio features. I combined both to make a more complete dataset.
- Data cleaning and duplicates removal was done.
- Saved to HDFS and Local Disk for future usage.

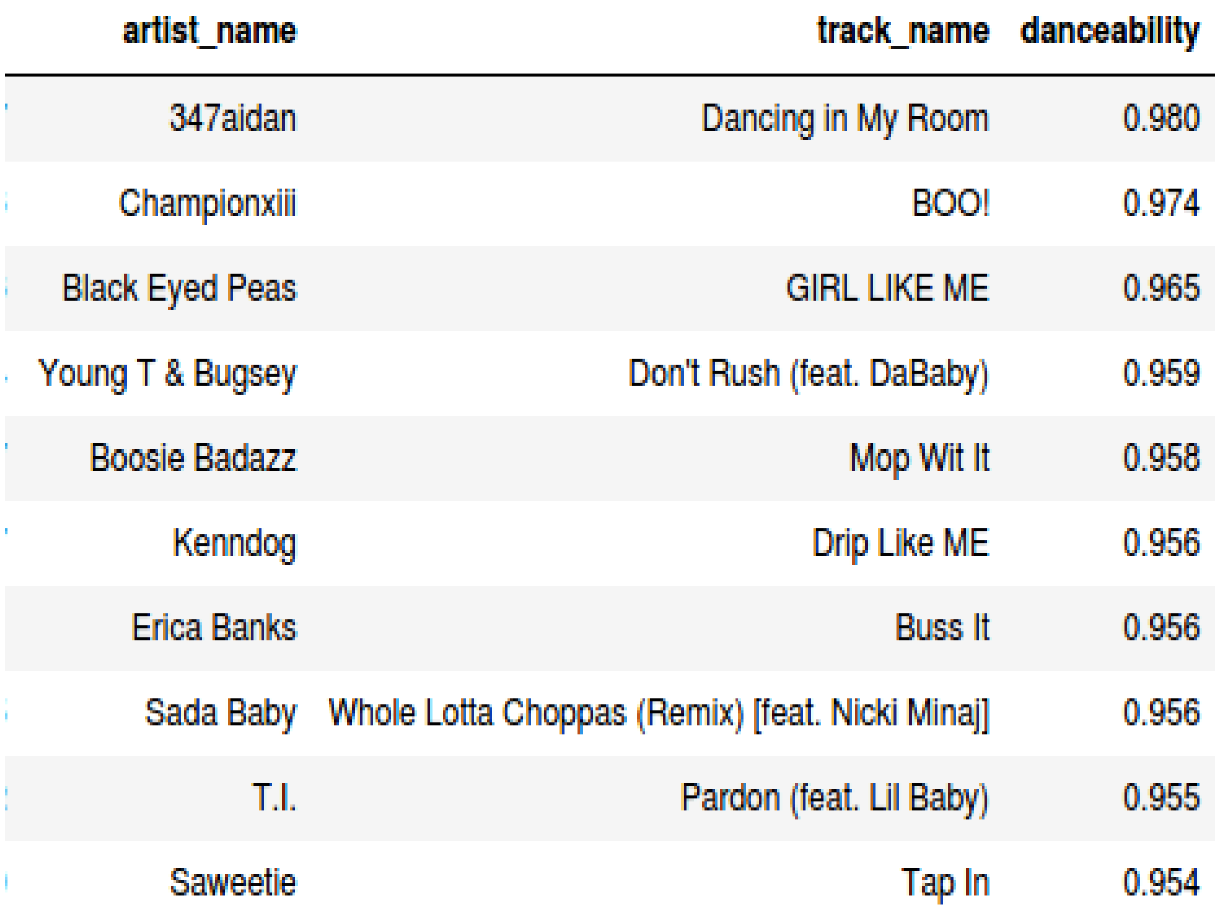
INSIGHTS AND VISUALIZATIONS

MOST POPULAR SONGS



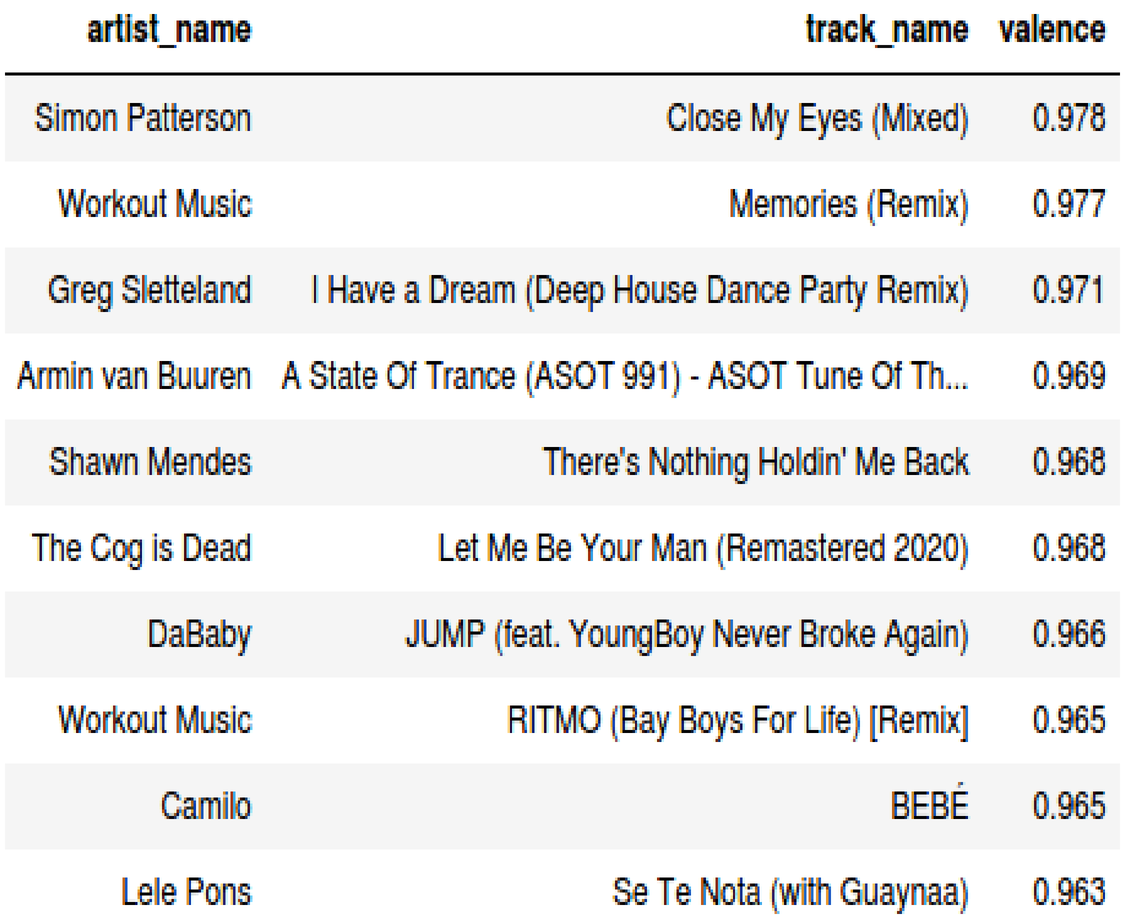
artist_name	popularity	track_name
24kGoldn	95	Mood (feat. iann dior)
The Weeknd	95	Blinding Lights
KAROL G	94	BICHOTA
Bad Bunny	94	DÁKITI
Ariana Grande	94	positions
Bad Bunny	93	LA NOCHE DE ANOCHE
The Kid LAROI	93	WITHOUT YOU
CJ	93	Whoopty
Billie Eilish	92	Therefore I Am
SZA	92	Good Days

SONGS MOST SUITED TO DANCING



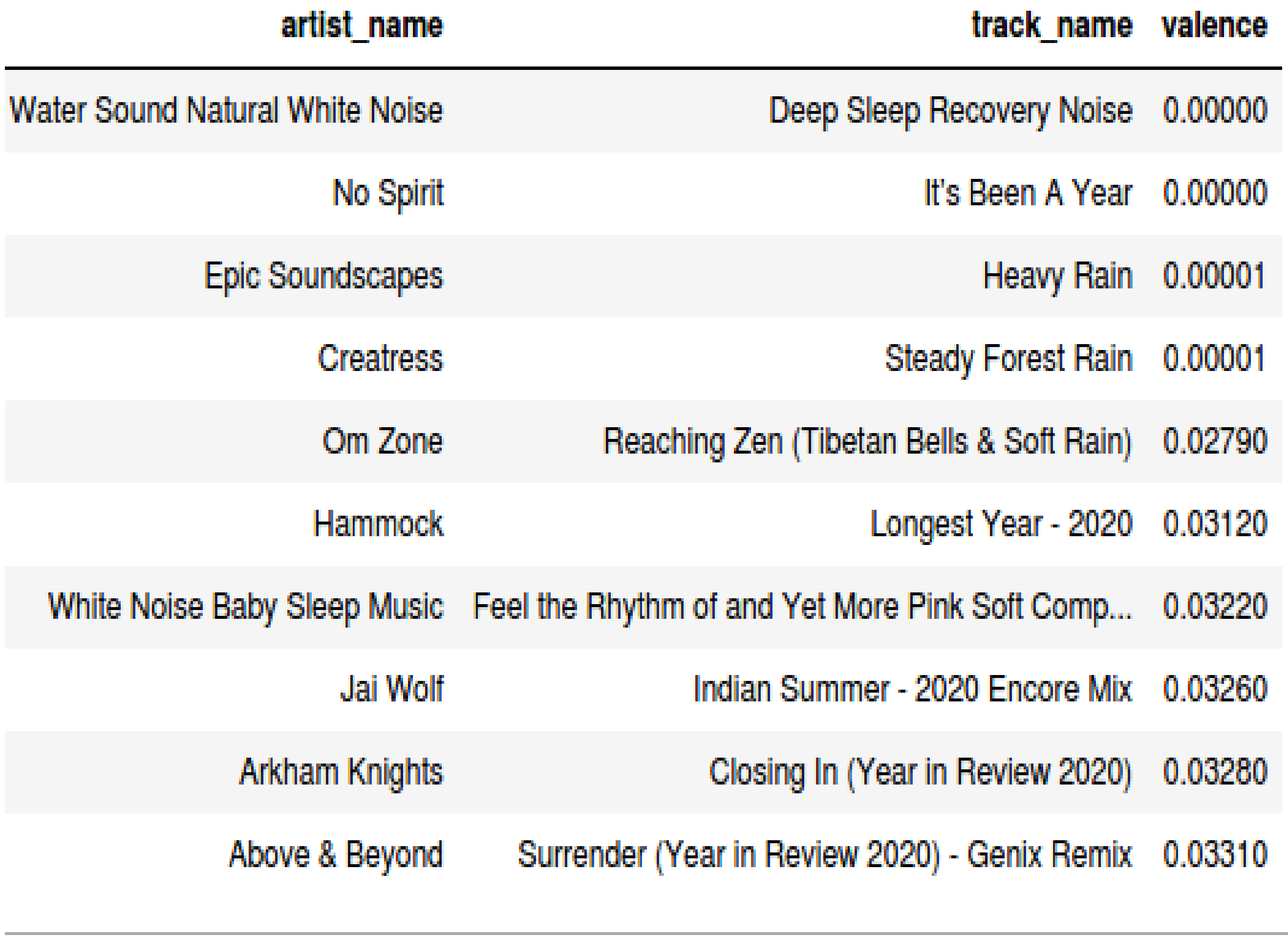
artist_name	track_name	danceability
347aidan	Dancing in My Room	0.980
Championxiii	BOO!	0.974
Black Eyed Peas	GIRL LIKE ME	0.965
Young T & Bugsey	Don't Rush (feat. DaBaby)	0.959
Boosie Badazz	Mop Wit It	0.958
Kenndog	Drip Like ME	0.956
Erica Banks	Buss It	0.956
Sada Baby	Whole Lotta Choppas (Remix) [feat. Nicki Minaj]	0.956
T.I.	Pardon (feat. Lil Baby)	0.955
Saweetie	Tap In	0.954

MOST CHEERFUL SONGS

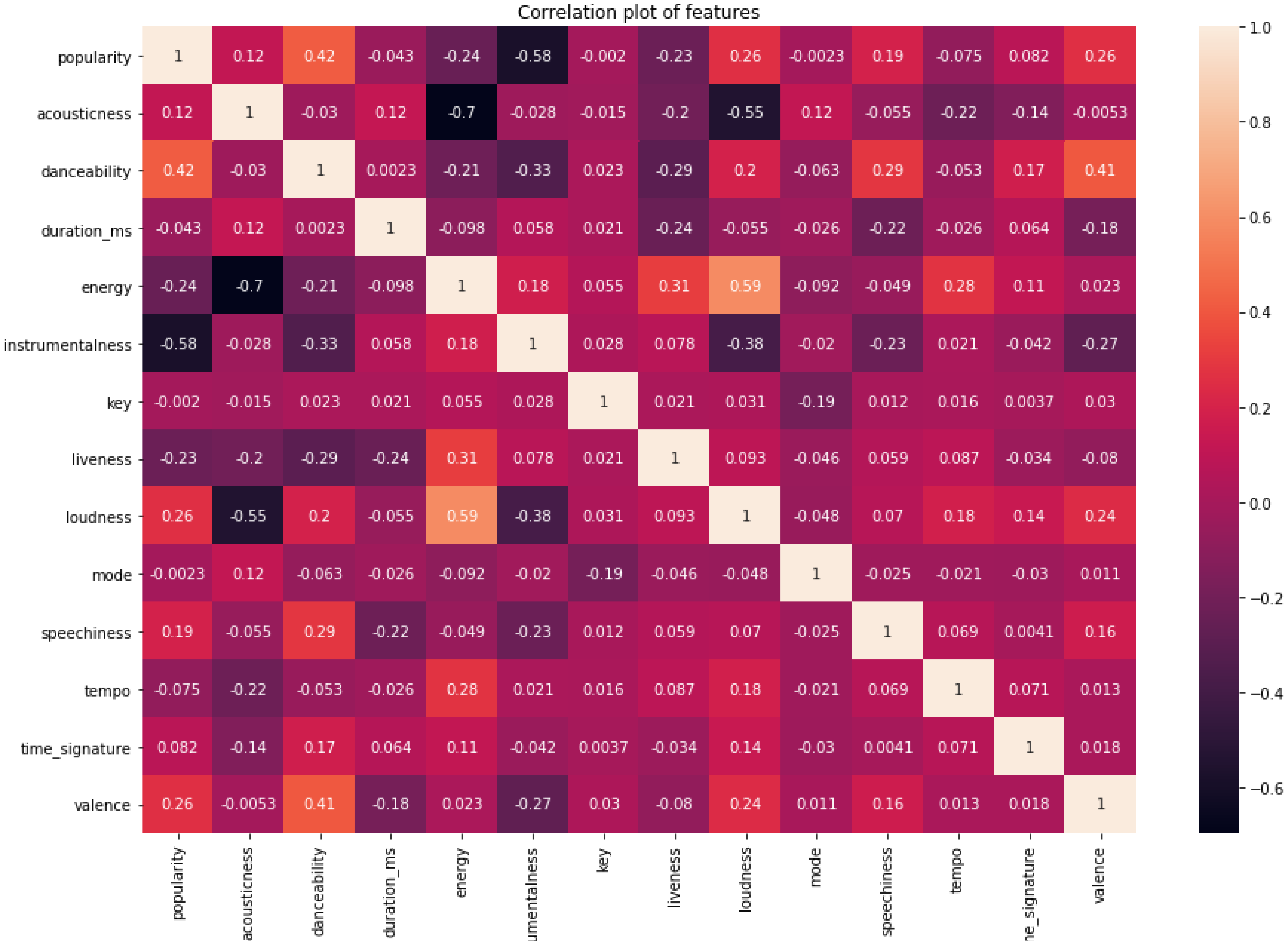


artist_name	track_name	valence
Simon Patterson	Close My Eyes (Mixed)	0.978
Workout Music	Memories (Remix)	0.977
Greg Sletteland	I Have a Dream (Deep House Dance Party Remix)	0.971
Armin van Buuren	A State Of Trance (ASOT 991) - ASOT Tune Of Th...	0.969
Shawn Mendes	There's Nothing Holdin' Me Back	0.968
The Cog is Dead	Let Me Be Your Man (Remastered 2020)	0.968
DaBaby	JUMP (feat. YoungBoy Never Broke Again)	0.966
Workout Music	RITMO (Bay Boys For Life) [Remix]	0.965
Camilo	BEBÉ	0.965
Lele Pons	Se Te Nota (with Guaynaa)	0.963

MOST DEPRESSING SONGS

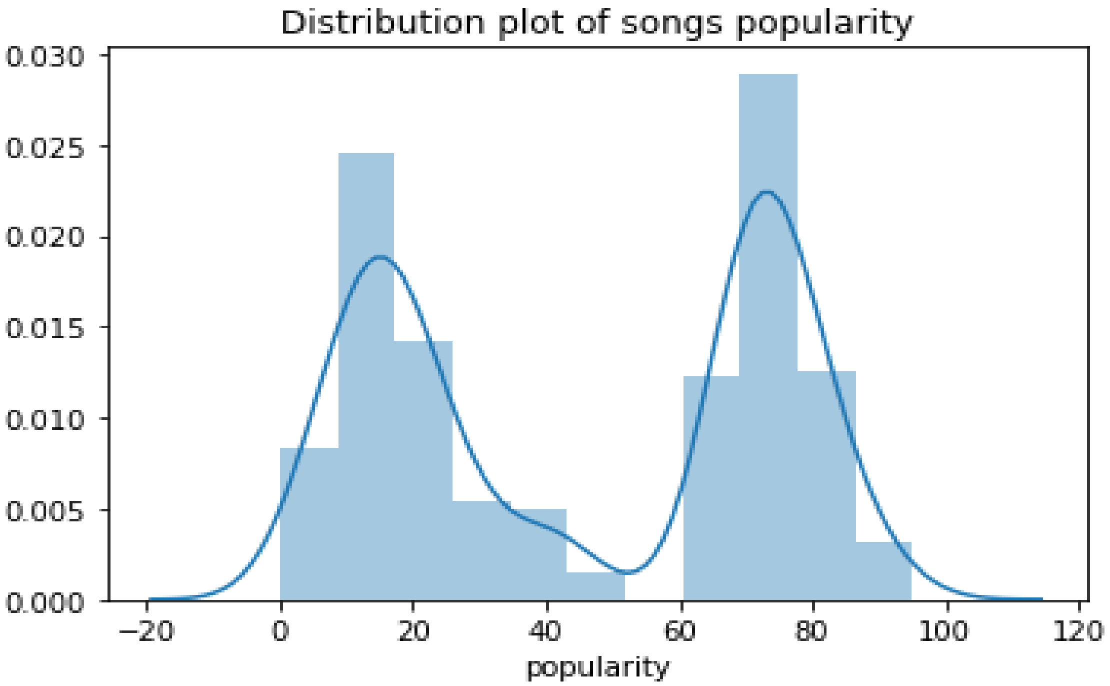


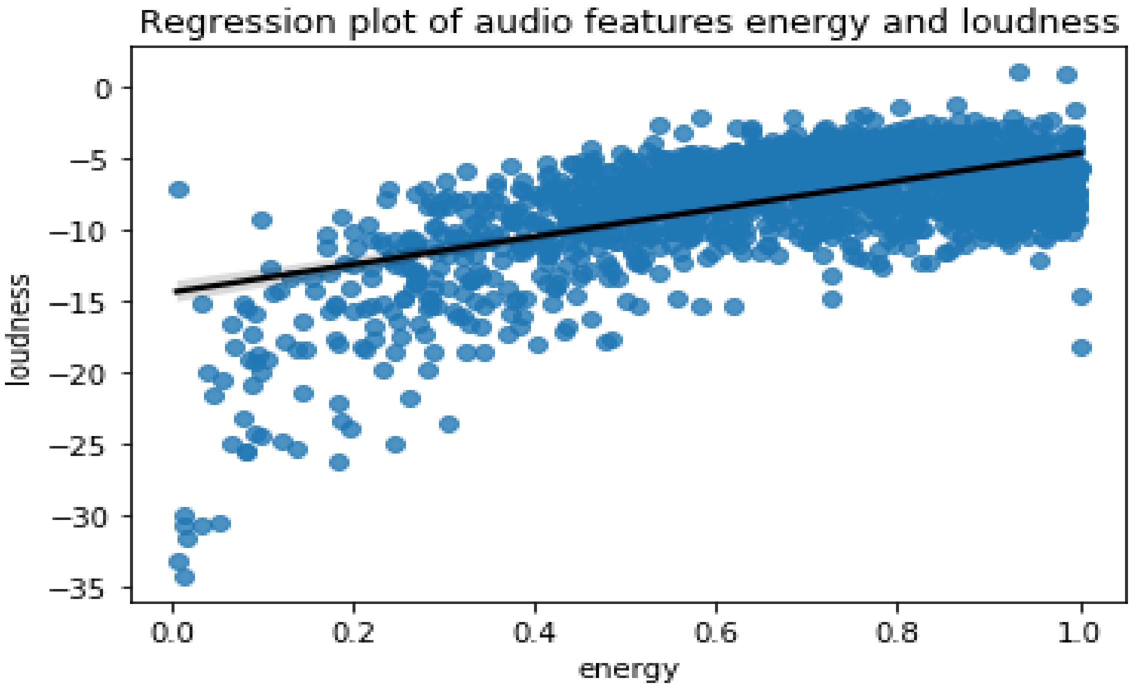
artist_name	track_name	valence
Water Sound Natural White Noise	Deep Sleep Recovery Noise	0.00000
No Spirit	It's Been A Year	0.00000
Epic Soundscapes	Heavy Rain	0.00001
Creatress	Steady Forest Rain	0.00001
Om Zone	Reaching Zen (Tibetan Bells & Soft Rain)	0.02790
Hammock	Longest Year - 2020	0.03120
White Noise Baby Sleep Music	Feel the Rhythm of and Yet More Pink Soft Comp...	0.03220
Jai Wolf	Indian Summer - 2020 Encore Mix	0.03260
Arkham Knights	Closing In (Year in Review 2020)	0.03280
Above & Beyond	Surrender (Year in Review 2020) - Genix Remix	0.03310



Correlation plot of features

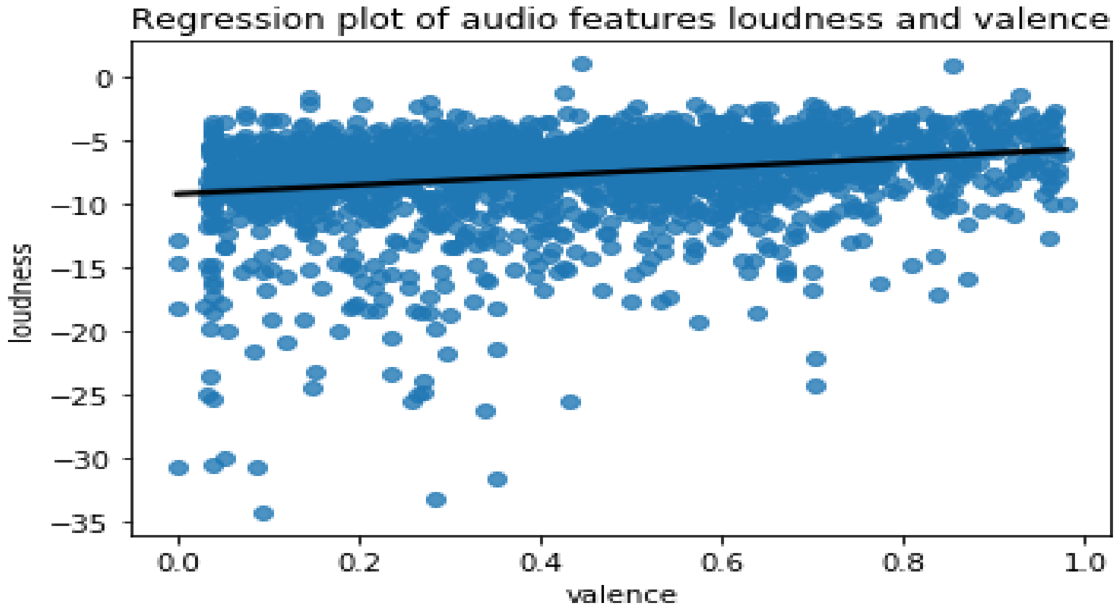
	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence
popularity	1	0.12	0.42	-0.043	-0.24	-0.58	-0.002	-0.23	0.26	-0.0023	0.19	-0.075	0.082	0.26
acousticness	0.12	1	-0.03	0.12	-0.7	-0.028	-0.015	-0.2	-0.55	0.12	-0.055	-0.22	-0.14	-0.0053
danceability	0.42	-0.03	1	0.0023	-0.21	-0.33	0.023	-0.29	0.2	-0.063	0.29	-0.053	0.17	0.41
duration_ms	-0.043	0.12	0.0023	1	-0.098	0.058	0.021	-0.24	-0.055	-0.026	-0.22	-0.026	0.064	-0.18
energy	-0.24	-0.7	-0.21	-0.098	1	0.18	0.055	0.31	0.59	-0.092	-0.049	0.28	0.11	0.023
instrumentalness	-0.58	-0.028	-0.33	0.058	0.18	1	0.028	0.078	-0.38	-0.02	-0.23	0.021	-0.042	-0.27
key	-0.002	-0.015	0.023	0.021	0.055	0.028	1	0.021	0.031	-0.19	0.012	0.016	0.0037	0.03
liveness	-0.23	-0.2	-0.29	-0.24	0.31	0.078	0.021	1	0.093	-0.046	0.059	0.087	-0.034	-0.08
loudness	0.26	-0.55	0.2	-0.055	0.59	-0.38	0.031	0.093	1	-0.048	0.07	0.18	0.14	0.24
mode	-0.0023	0.12	-0.063	-0.026	-0.092	-0.02	-0.19	-0.046	-0.048	1	-0.025	-0.021	-0.03	0.011
speechiness	0.19	-0.055	0.29	-0.22	-0.049	-0.23	0.012	0.059	0.07	-0.025	1	0.069	0.0041	0.16
tempo	-0.075	-0.22	-0.053	-0.026	0.28	0.021	0.016	0.087	0.18	-0.021	0.069	1	0.071	0.013
time_signature	0.082	-0.14	0.17	0.064	0.11	-0.042	0.0037	-0.034	0.14	-0.03	0.0041	0.071	1	0.018
valence	0.26	-0.0053	0.41	-0.18	0.023	-0.27	0.03	-0.08	0.24	0.011	0.16	0.013	0.018	1





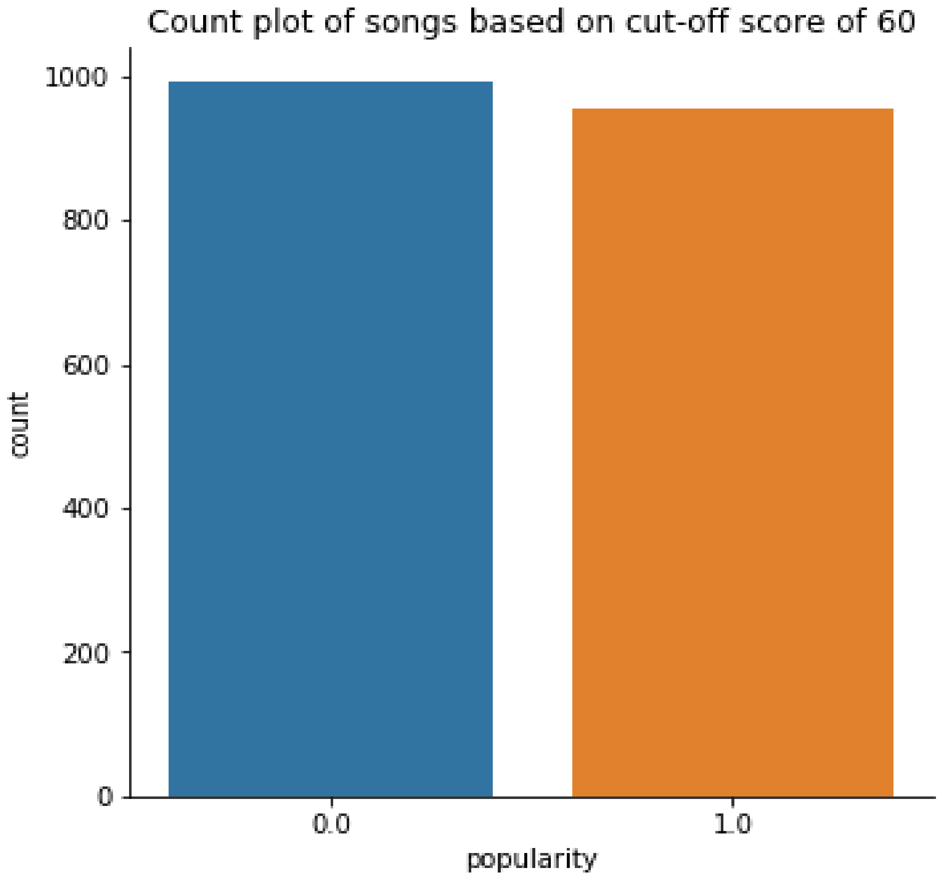
Shapiro-Wilk Normality Test confirms that popularity is not a Gaussian distribution.

Linear regression won’t be used.



MODEL USED

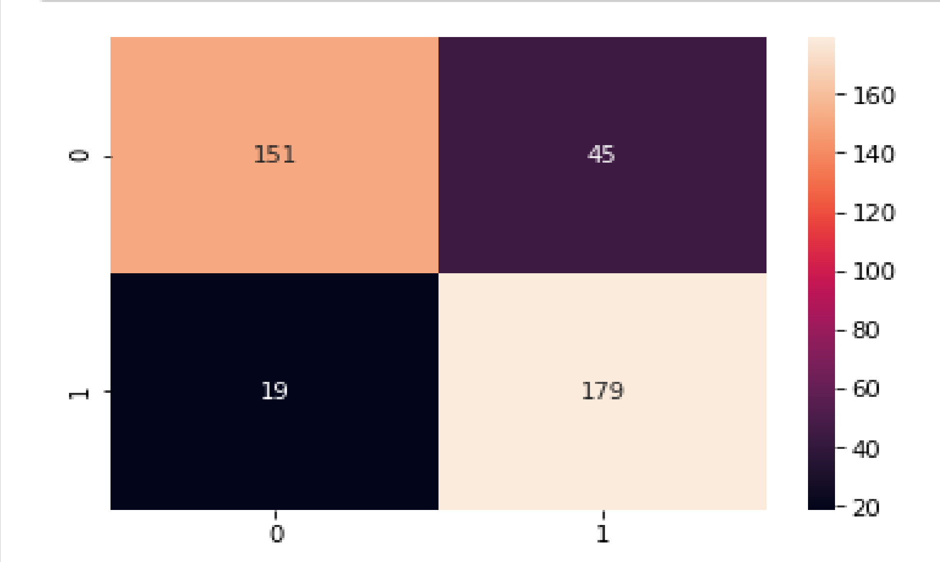
- ➢ Classification Model
  - o Logistic Regression Vs Decision Tree
- ➢ Three Categorical columns: Key, mode and time_signature
- ➢ Feature's engineering and scaling

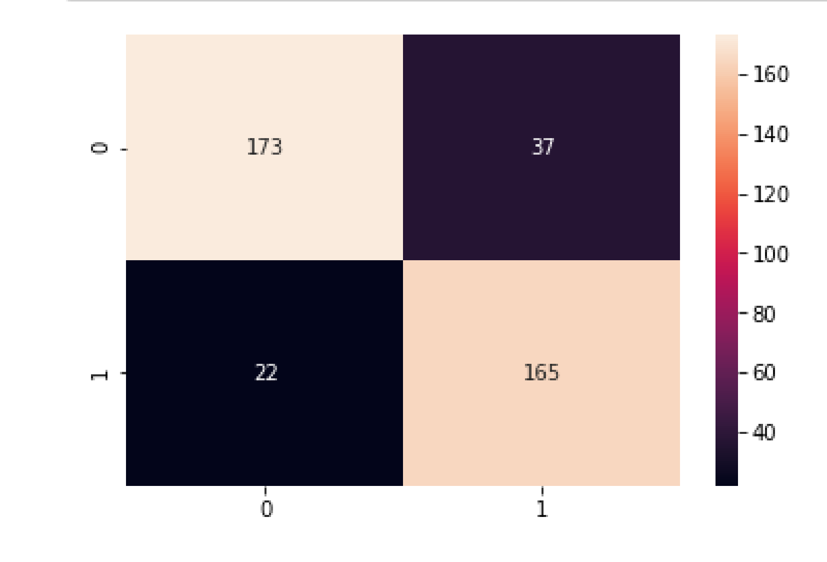


RESULTS & COMPARISON

MEASURE	LOGISTIC REGRESSION	DECISION TREE
Accuracy (Train, Test)	(86.8 %, 83.8 %)	(90%, 85.1 %)
Precision (Not Pop., Pop)	(0.89, 0.80)	(0.89, 0.82)
Recall (Not Pop., Pop)	(0.77, 0.9)	(0.82, 0.88)
F1-Score (Not Pop., Pop)	(0.83, 0.85)	(0.85, 0.85)

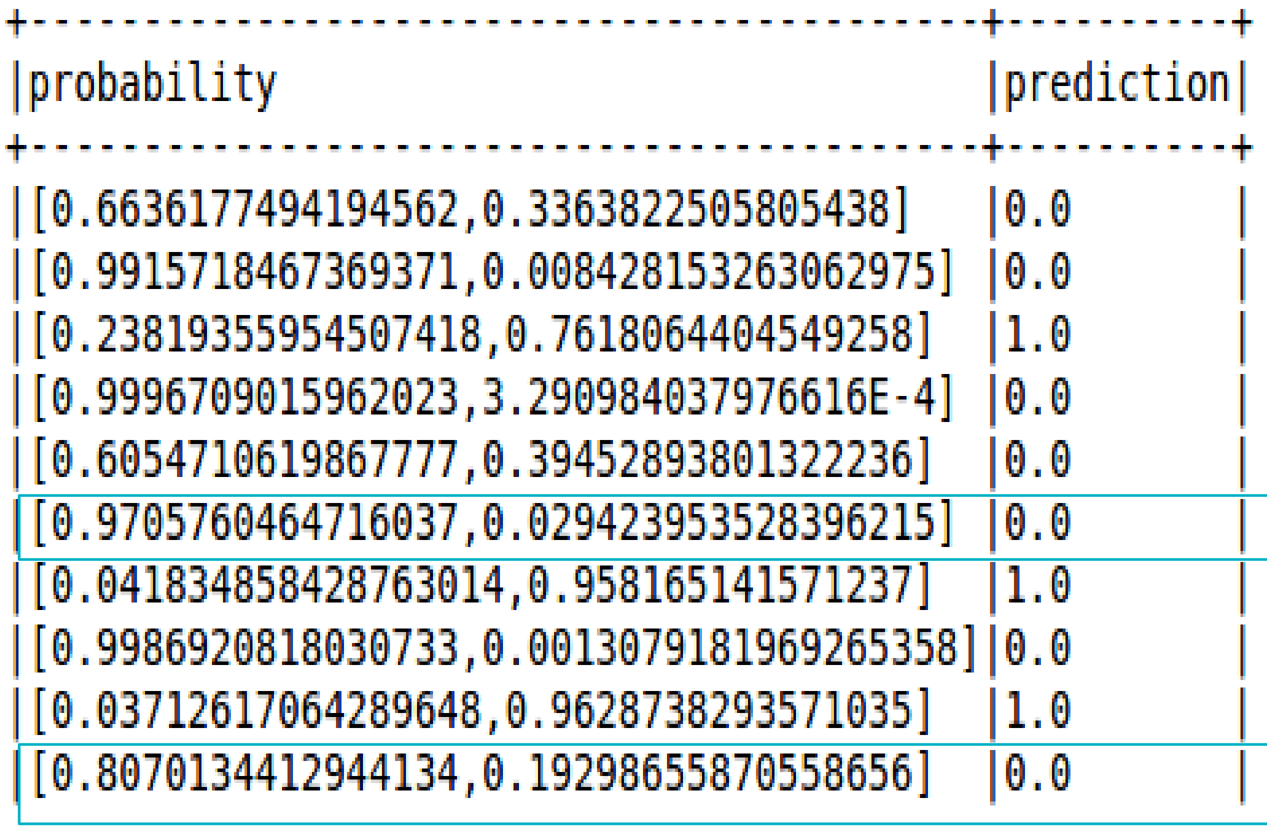
Confusion Matrix





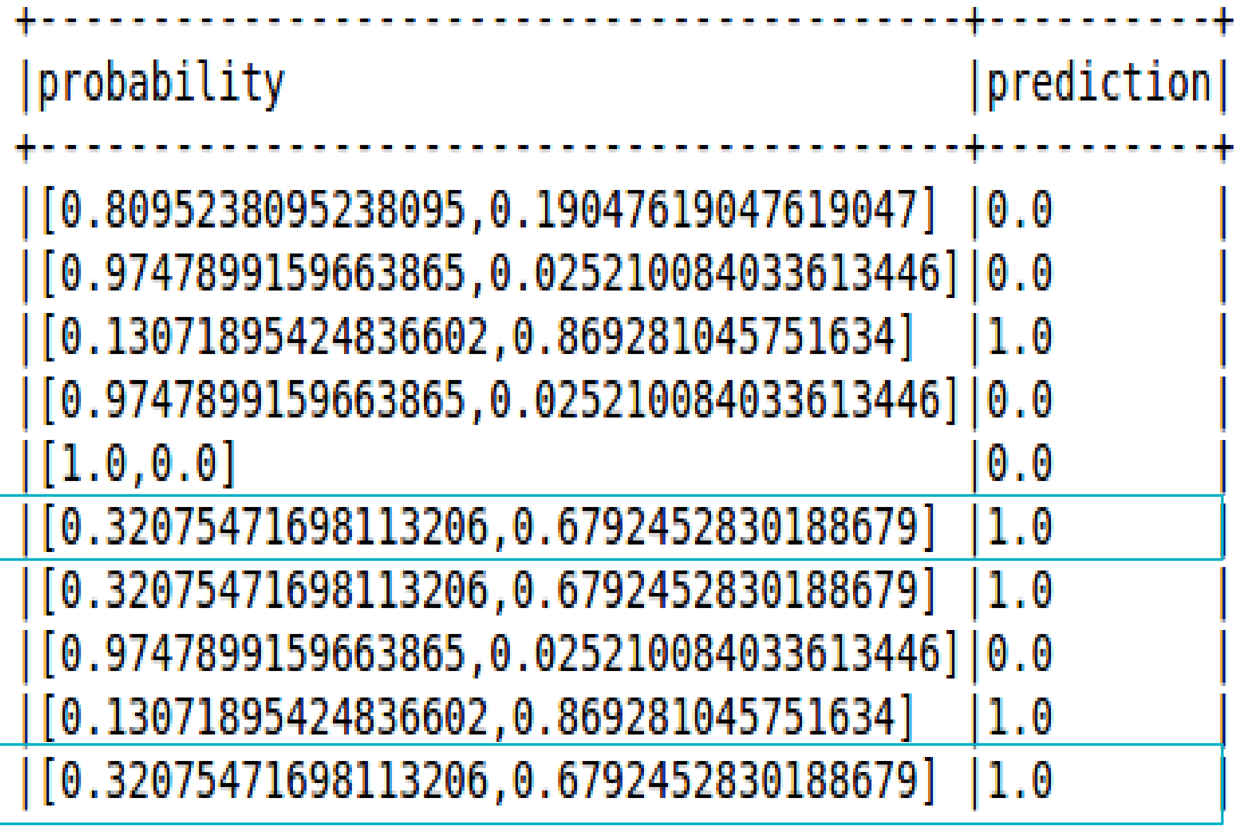
MODEL EVALUATION ON GENERATED SAMPLE DATA

LOGISTIC REGRESSION



```
+------------------------------------------+----------+
|probability                               |prediction|
+------------------------------------------+----------+
|[0.6636177494194562,0.3363822505805438]   |0.0       |
|[0.9915718467369371,0.008428153263062975] |0.0       |
|[0.23819355954507418,0.7618064404549258]  |1.0       |
|[0.9996709015962023,3.290984037976616E-4] |0.0       |
|[0.6054710619867777,0.39452893801322236]  |0.0       |
|[0.9705760464716037,0.029423953528396215] |0.0       |
|[0.041834858428763014,0.958165141571237]  |1.0       |
|[0.9986920818030733,0.0013079181969265358]|0.0       |
|[0.03712617064289648,0.9628738293571035]  |1.0       |
|[0.8070134412944134,0.19298655870558656]  |0.0       |
```

DECISION TREE



```
+-----------------------------------------+----------+
|probability                              |prediction|
+-----------------------------------------+----------+
|[0.8095238095238095,0.19047619047619047] |0.0       |
|[0.9747899159663865,0.025210084033613446]|0.0       |
|[0.13071895424836602,0.869281045751634]  |1.0       |
|[0.9747899159663865,0.025210084033613446]|0.0       |
|[1.0,0.0]                                |0.0       |
|[0.32075471698113206,0.6792452830188679] |1.0       |
|[0.32075471698113206,0.6792452830188679] |1.0       |
|[0.9747899159663865,0.025210084033613446]|0.0       |
|[0.13071895424836602,0.869281045751634]  |1.0       |
|[0.32075471698113206,0.6792452830188679] |1.0       |
```

CONCLUSION

- Model can be improved upon on both cases in the future
- Decision Tree is a better model for the most part
- Popular songs predicted were found to have high danceability value, high energy value, low speechiness value and above 0.5 loudness value.

THANKS