SPOTIFY HIT PREDICTOR

Spotify

INDEX

01 Introduction

This Project revolves around the application of data and machine learning to discern patterns in the realm of music popularity.

We used Spotify's extensive dataset and employing the Random Forest Classifier alongside Principal Component Analysis, we implemented a systematic exploration to decode the factors that elevate songs to chart-topping status. through this intersection of data science and music, we aim to uncover the empirical foundation of musical success.

02 Data description

The Spotify Hit Predictor Dataset (1960-2019)

- **Danceability:** how suitable a track is for dancing
- **Energy:** a perceptual measure of intensity and activity
- **Loudness:** the overall loudness of a track in decibels.
- **Mode:** indicates the modality (major or minor)
- **Speechiness:** detects the presence of spoken words in a track
- **Acousticness**: whether the track is acoustic
- **Liveness:** presence of an audience in the recording
- **Valence:** the musical positiveness conveyed by a track

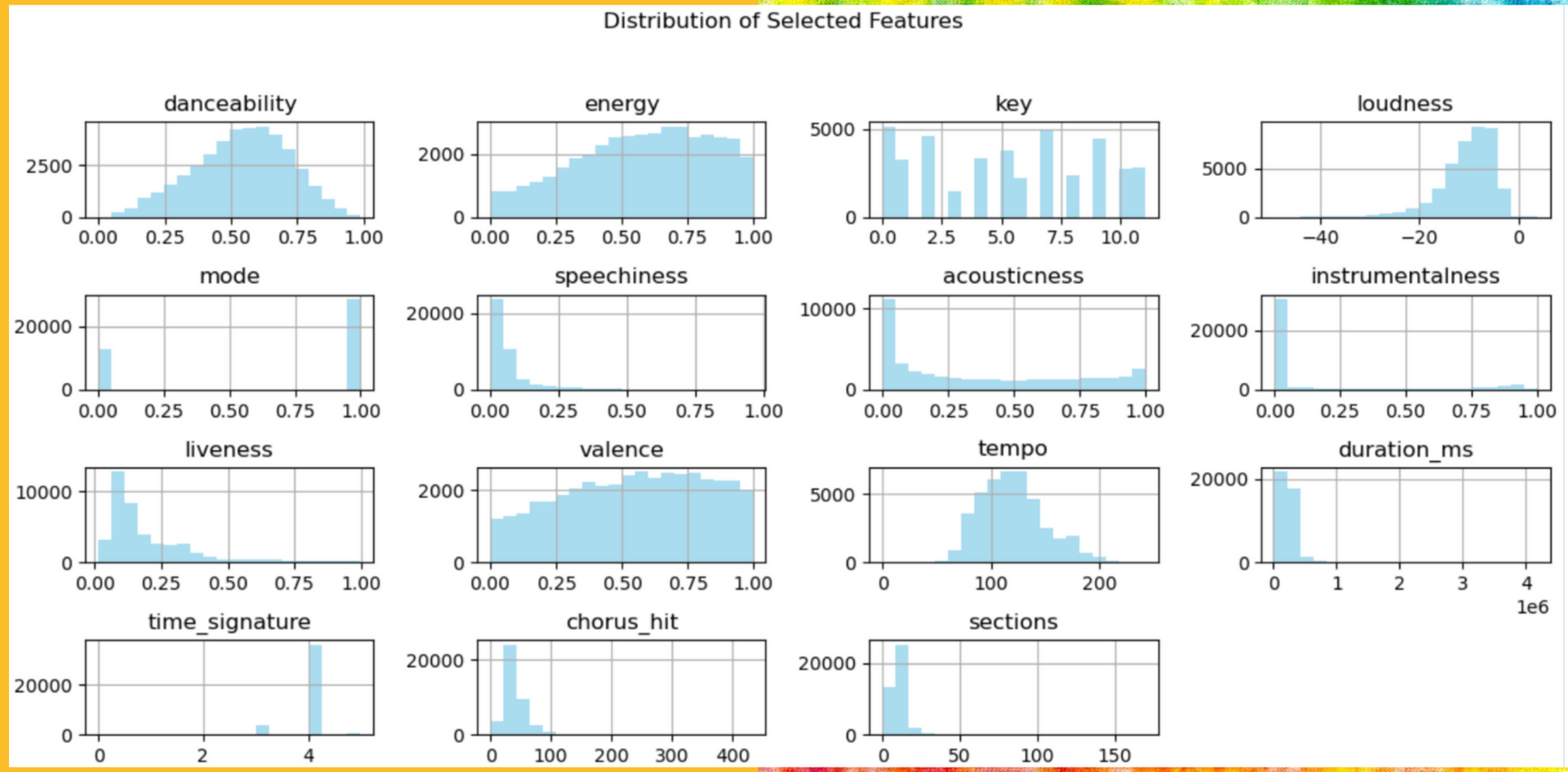
This dataset consists of features for tracks fetched using Spotify's Web API. The tracks are labeled '1' or '0' ('Hit' or 'Flop'). It has 20 columns and 41,160 rows.

VISIT TABLEAU

03 Data Distribution

This set of histograms displays the distribution of selected feature in the dataset.

These histograms provide insights into the distribution of individual features, helping you understand their characteristics



04 Data preprocessing

We merged 6 csv files with information from the 60's until 2010´s

- Rename columns
- Add a column with the decade
- Check the data types
- Look for N/A values

We used Mongo DB for the database

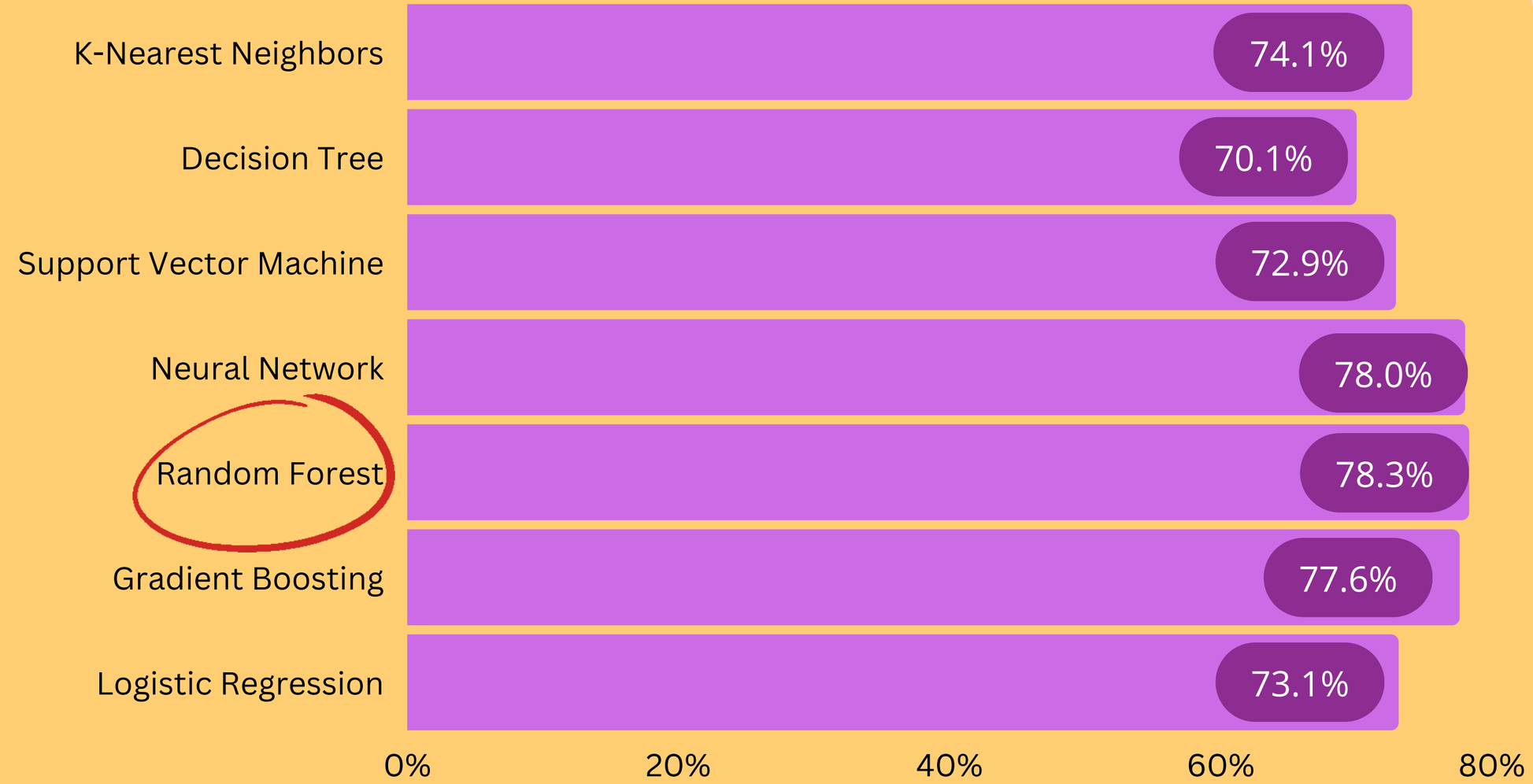
```
mongo= MongoClient(port=27017)
db = mongo.SpotifySongs

SpotifySongs = df
SpotifySongs = SpotifySongs.to_dict(orient='records')

db.SpotifySongs.insert_many(SpotifySongs)
```

2.2s

	key	loudness	mo
	3	-7.727	1
98	3	-12.475	1
657	5	-13.392	1
90	7	-12.058	0
	11	-3.515	0
	...	...	

05 Model selection

Random Forest and Neural Network where the models showing the greatest socre

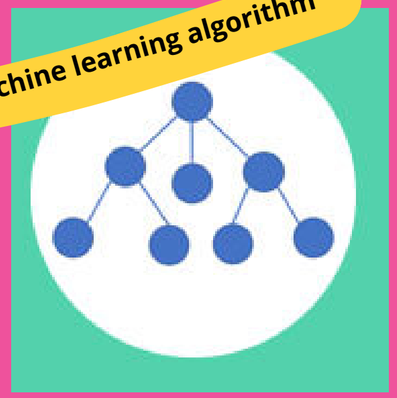


We ran several models to identify the ones with the highest scores

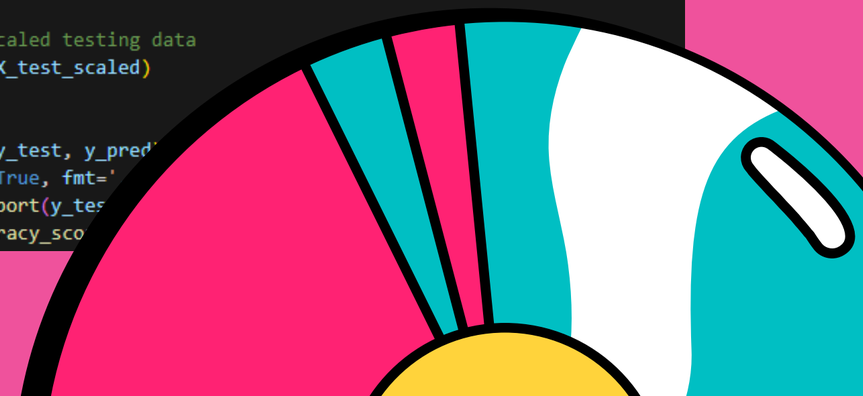
acknowledgment: GABRIEL ATKI

06 Random Forest

supervised machine learning algorithm



```python
# Split your dataset into features (X) and target variable (y)
X = df[['danceability', 'energy', 'key', 'loudness', 'mode', 'speechiness',
        'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo',
        'duration_ms', 'time_signature', 'chorus_hit', 'sections']]
y = df['hit']

# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=0)

# Initialize the StandardScaler
scaler = StandardScaler()

# Fit the scaler to the training data and transform both training and testing data
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Initialize the RandomForestClassifier
model = RandomForestClassifier(random_state=0)

# Train the model on scaled training data
model.fit(X_train_scaled, y_train)

# Make predictions on scaled testing data
y_pred = model.predict(X_test_scaled)

# Evaluate the model
cnf = confusion_matrix(y_test, y_pred
sns.heatmap(cnf, annot=True, fmt='
print(classification_report(y_tes
print("Accuracy:", accuracy_sco
```



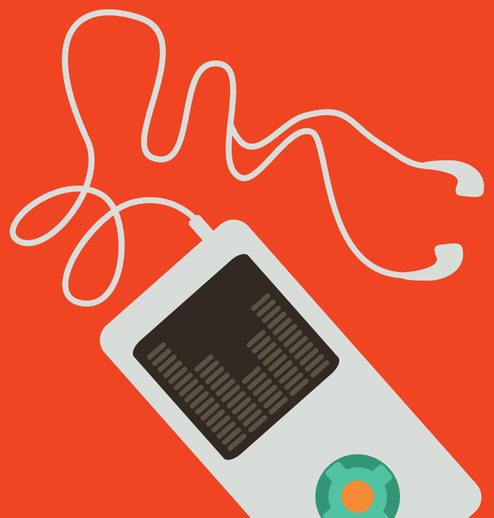
```python
import joblib

#save model to file
joblib.dump(model, 'model.pkl')
joblib.dump(scaler, 'scaler.pkl')
```

VISIT PREDICTOR

07 Conclusion

We observe that Hits are:

- Less Acoustic
- Have more energy
- Are more danceble
- Are not instrumetal



The best model was random forest due to the generation of n number of trees, an take votes for each prediction, is also a good model because is easy to explain and is very powerful

The classification model predicts whether a track would be a 'Hit' or not based on the premiss that pop music success can be the result of a recepie or formula. Which could hve big relevance in the music industry

The relevance of the variables increases for enegy and danceability towards 2010's.

By following the right convination of factors in a very precise way we are able to vislualize with our model the song that are hits. Therefore we ca conclude that there can be a formula or a recepie for a winning song but this combinaton must be done in a very specific combination that might require slight variations as time goes by

LARGA VIDA A LA MUSICA
THANK YOU!