WAS MACHT EINEN SONG AUF SPOTIFY BELIEBT?

MIT AUDIO-DATEN DIE BELIEBTHEIT VORHERSAGEN

ADRIANO ELIA, 09.05.2025

PROJEKTKONTEXT & ZIELSETZUNG

Datengrundlage

- Öffentlicher Datensatz von kaggle.com
- Ca. 90.000 Songs und deren Metadaten sowie Audio-Eigenschaften

Datenqualität

- Strukturiert und sauber, wenig Aufbereitung nötig

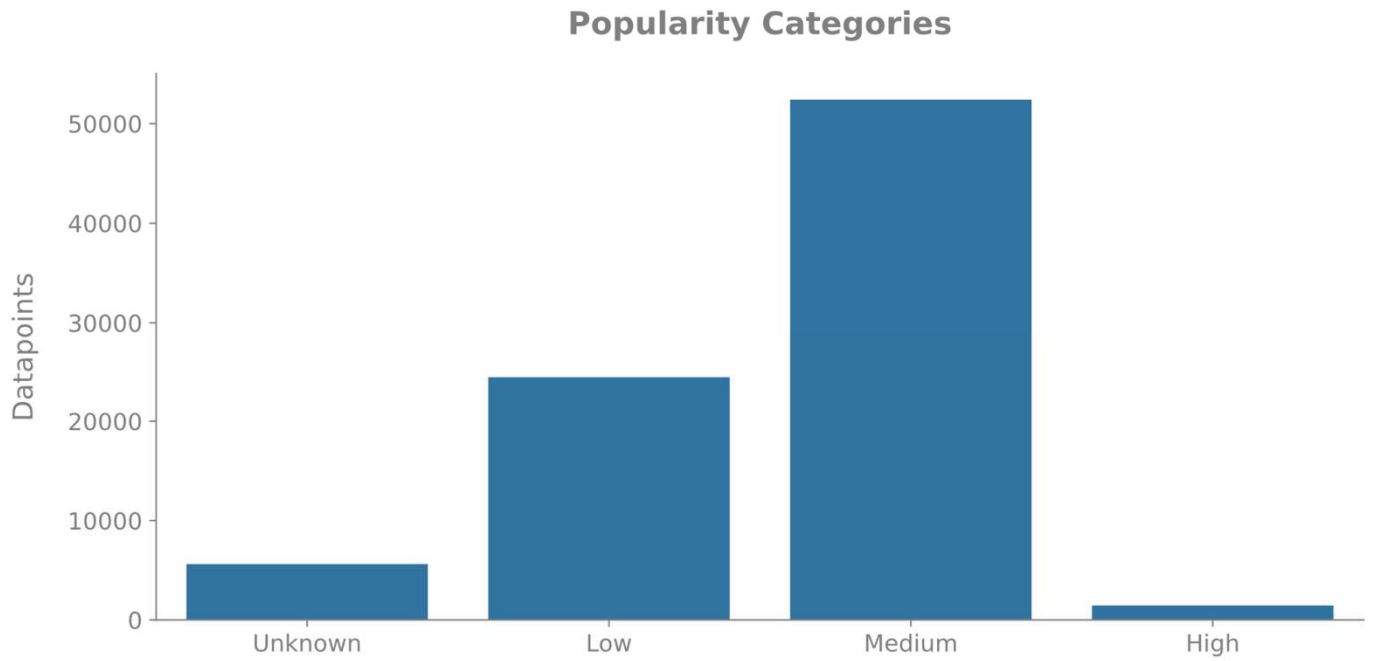
Ziel

- Die **Beliebtheit** eines Songs auf Spotify anhand von reinen Audio-Daten **vorhersagen**

ZIELVARIABLE & KLASSIFIZIERUNG

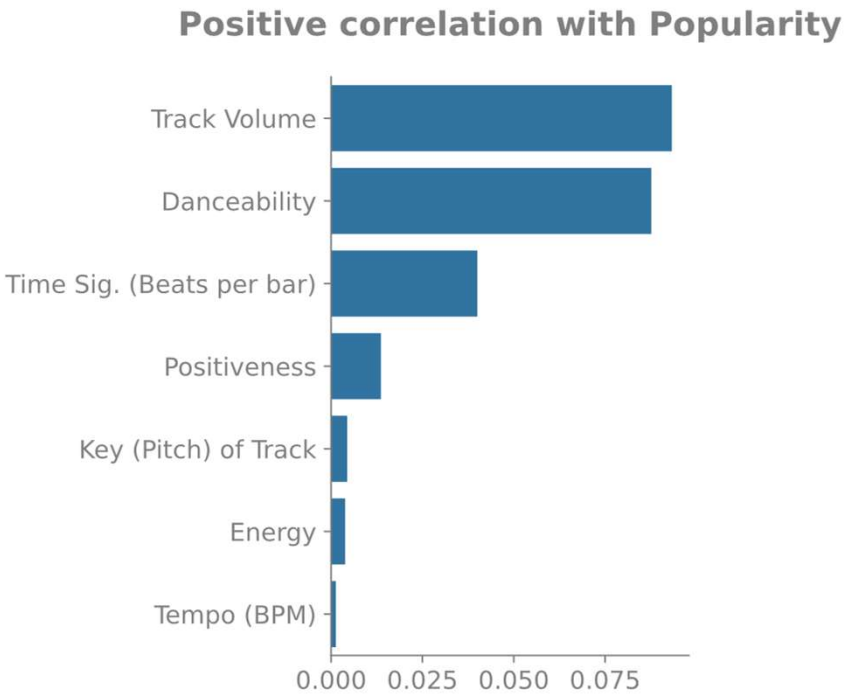
4 Klassen für Beliebtheitswert

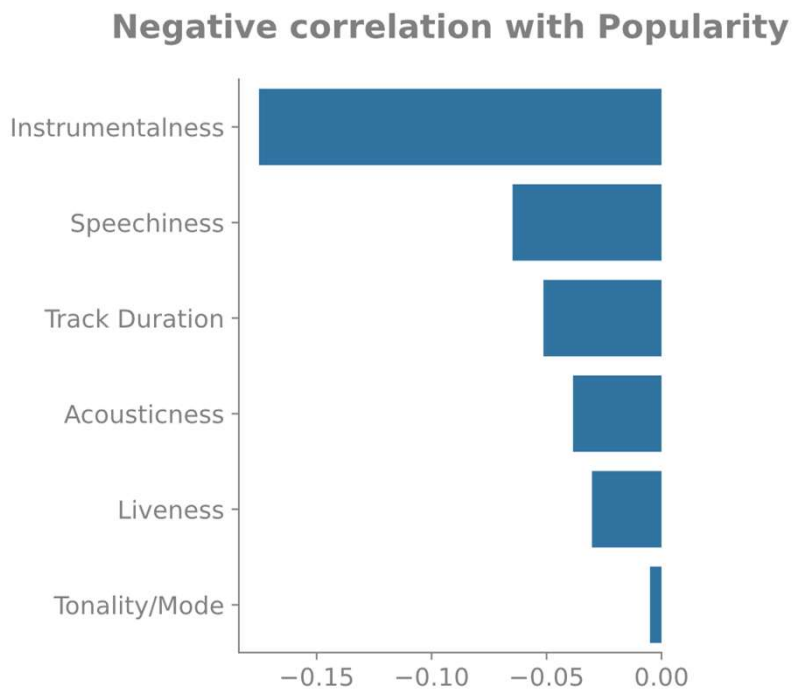
- Unknown (0), **Low** (<25), **Medium** (25-75) und **High** (75>)



MERKMALE & ZUSAMMENHÄNGE

- Lautstärke, „Tanzbarkeit“, 4/4 Beat, positive Vibes => **eher höhere Beliebtheit**
- Instrumentalität, Acapella/Podcasts, Dauer, Live sowie Akustik => **eher niedrigere Beliebtheit**

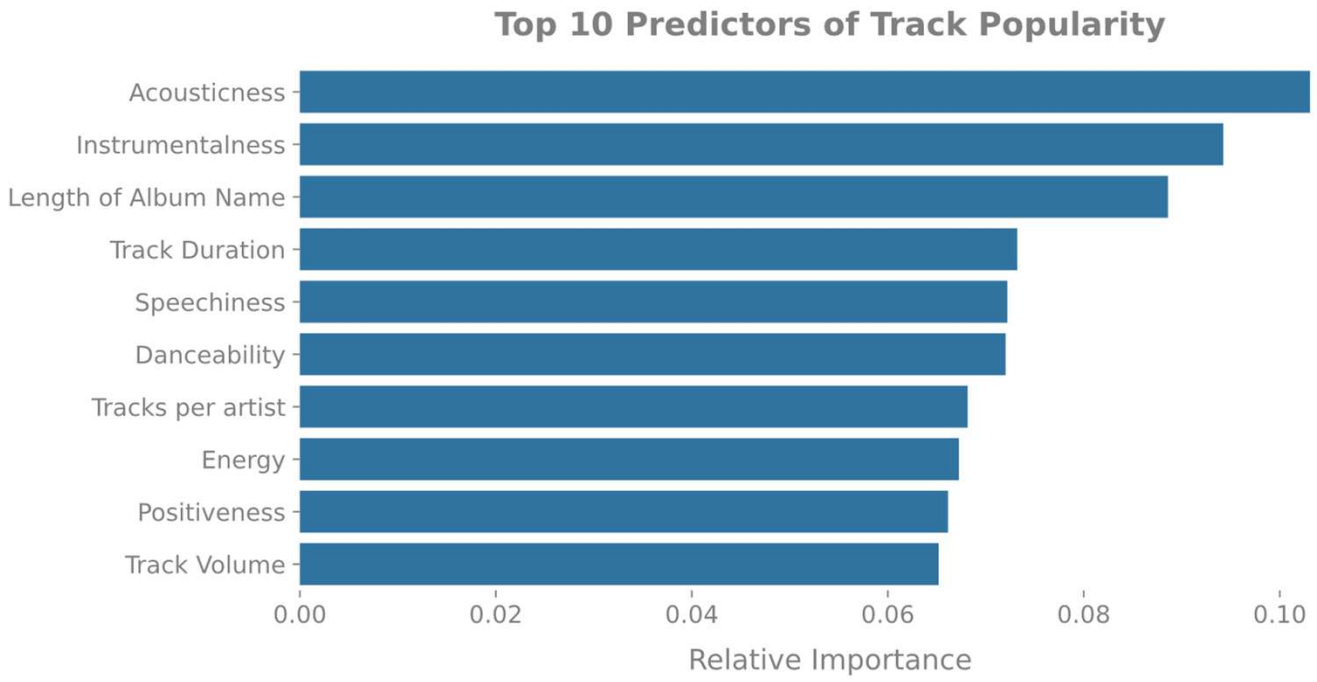




DIE EINFLUSSREICHSTEN MERKMALE

Akustik, Instrumentalität und die Länge des Album Namen?

- Im Machine-Learning-Modell als einflussreich auf die Vorhersage der Beliebtheit eingestuft



EVALUATION

Qualität des Vorhersagemodells

- Mittelmäßige Testwerte (0,63 Genauigkeit im Durchschnitt der 4 Kategorien)
- Ungleichgewicht der Kategorien der Zielvariable zeigt sich trotz Berücksichtigung im Modell deutlich
- „Medium“ = hohe Anzahl Daten = genaue Vorhersage; High = niedrige Anzahl Daten = ungenaue Vorhersage

AUSBLICK

Weitere Möglichkeiten

- Weitere Methoden zur Verringerung des Ungleichgewichts anwenden („Over-/Undersampling“)
- Weitere Merkmale erzeugen („Feature Engineering“)
- Datenlage erhöhen (weiterer Spotify-Datensatz hinzuziehen)
- Alternativen Modellansatz testen
  - Regressionsmodell statt Klassifizierung nutzen und Beliebtheit als Wert von 0-100 vorhersagen, statt in Kategorien einzuteilen

EVALUATIONSMETRIKEN & LEARNING CURVE

Nur falls von Interesse

- (links evaluation scores, rechts learning curve, beides vom finalen RandomForestClassifier)

	precision	recall	f1-score	support
High	0.32	0.59	0.42	242.0
Low	0.62	0.45	0.52	2747.0
Medium	0.78	0.74	0.76	6376.0
Unknown	0.41	0.71	0.52	1189.0
accuracy	0.66	0.66	0.66	0.66