# Music Hit Prediction Summary

## Problem

Predict whether a song could be a hit song based on audio features (popularity >= 65), in order to help with commercial behaviors.

---

## Data

89,741 tracks from Kaggle Spotify Dataset and Spotify&Last.fm APIs.

Including:

- original audio features (danceability, loudness, speechiness, ...)
- metadata features (mode, time_signature, ...)
- engineered features (duration_min, loudness_normalized)

---

## Model

| Model | Accuracy | F1 Score |
|---|---|---|
| Decision Tree | 0.9328 | 0.0049 |
| Random Forest | 0.9357 | 0.0000 |
| XGBoost | 0.9353 | 0.0102 |
| **Tuned Decision Tree** | **0.8931** | **0.1095** |
| Tuned Random Forest | 0.9330 | 0.0491 |
| Tuned XGBoost | 0.9355 | 0.0119 |

Picked Tuned Decision Tree to be the best model.

---

## Results

Within 200 sample tracks

- Predicted hit but not hit: 6
- Predicted not hit but hit: 8

Accuracy: 186/200 (93.0%)

---

## Key Observations

- F1 Score is more important than Accuracy for imbalanced data.
- Top 3 most important features for prediction are tempo, liveness, duration_min

---

## Limitations

- Data is imbalanced: only 7% songs are hit songs.
- Data is kind of outdated: Spotify has restricted access to the Audio Features API for regular developer accounts recently.
- Audio features only: lyrics, languages, and cultural trends are not captured.
- F1 scores are low: the best f1 score is 0.1095, still needs improving.
- Static threshold: fixed threshold (65) may not reflect changing industry standards.

---

## Next Steps

- Fetch data from newly released songs in other ways.
- Add more features like artist follower count, languages, professional ratings.
- Try different thresholds like top 25% instead of fixed popularity.
- Combine multiple models to improve the accuracy of predictions.
- Train the model with neural networks, benefiting from deep learning.