جامعة الملك سعود
King Saud University

كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

King Saud University
College of Computer and Information Sciences
Department of Information Technology

Spotify

IT 362 Course Project
Semester-1, 1447H

Student Name	Student ID
Dana Alsalami	443200945
Norah Aljayan	444200832
Aljohara Alsultan	444203635
Layan Alhowaimel	444200969
Alya Almuqren	444200610

Course Instructor: **Dr. Abeer Aldayel**

November 6

Logbook Entry – Model Selection & Evaluation

Date: 2025-11-18

Steps Taken:

- Selected a set of algorithms capable of addressing two complementary goals: establishing interpretable baselines and capturing more complex patterns in the data.
- Chose Linear Regression as the initial benchmark to quantify the linear contribution of each feature to artist popularity.
- Introduced Random Forest to model nonlinear relationships and interactions between genres, and to evaluate feature importance robustly.
- Added **XGBoost** because of its strong performance on structured tabular data and ability to capture subtle interactions between features.
- Prepared the dataset for modelling by using numeric features (genre_count, followers) and one-hot encoded genre features, limited to the top 50 genres to prevent high dimensionality issues.

Reason:

- Linear Regression provides transparency for baseline interpretation.
- Random Forest captures nonlinear patterns, interactions, and robustness to noise.
- XGBoost allows the detection of complex feature interactions and typically performs well on tabular datasets.

Tools:

- Python, Pandas, Scikit-learn, XGBoost

Challenges:

- Inconsistent genre labels across entries required a custom parser to standardize naming.
- High dimensionality risk due to one-hot encoding; mitigated by limiting to top 50 genres.
- Some genres appeared infrequently, so cross-validation folds were carefully chosen to maintain at least 20 samples per fold.

Decisions Made:

- Retained **genre_count** despite low standalone predictive power, as it improves ensemble model performance.
- Performed 5-fold cross-validation to avoid overfitting and ensure reliable generalization estimates.

Notes:

- Baseline Linear Regression using only genre_count showed low R^2, confirming that the number of genres alone is a weak predictor.
- After adding one-hot encoded genre features, Linear Regression R^2 increased substantially, showing that specific genres influence popularity differently.
- Random Forest and XGBoost consistently outperformed Linear Regression, indicating strong nonlinear interactions and the importance of particular genres (e.g., khaleeji negatively, egyptian pop positively).