Numan Tok, Lily Djami,
Ilir Hajrullahu

Spotify Group Playlist ML Model Design

Overall Process:

1. Get data for the 100 (variable) most listened library saved tracks of each user
2. Apply Preprocessing
    1. numerical transformations
        1. Fill missing values with the mean (variable)
        2. Scale features (e.g. normalization)
    2. categorical transformations
        1. Fill missing values with "missing" (variable)
        2. turn into numbers: one-hot encoding
3. Apply a clustering algorithm on the whole dataset
4. Split the dataset into training, validation, and test sets with proportions 80%/10%/10% (variable) and a balanced distribution of cluster memberships
5. Train an Artificial Neural Network (ANN) (variable)
6. Validate/finetune the NN parameters with the validation set
7. Test the final model against the test set

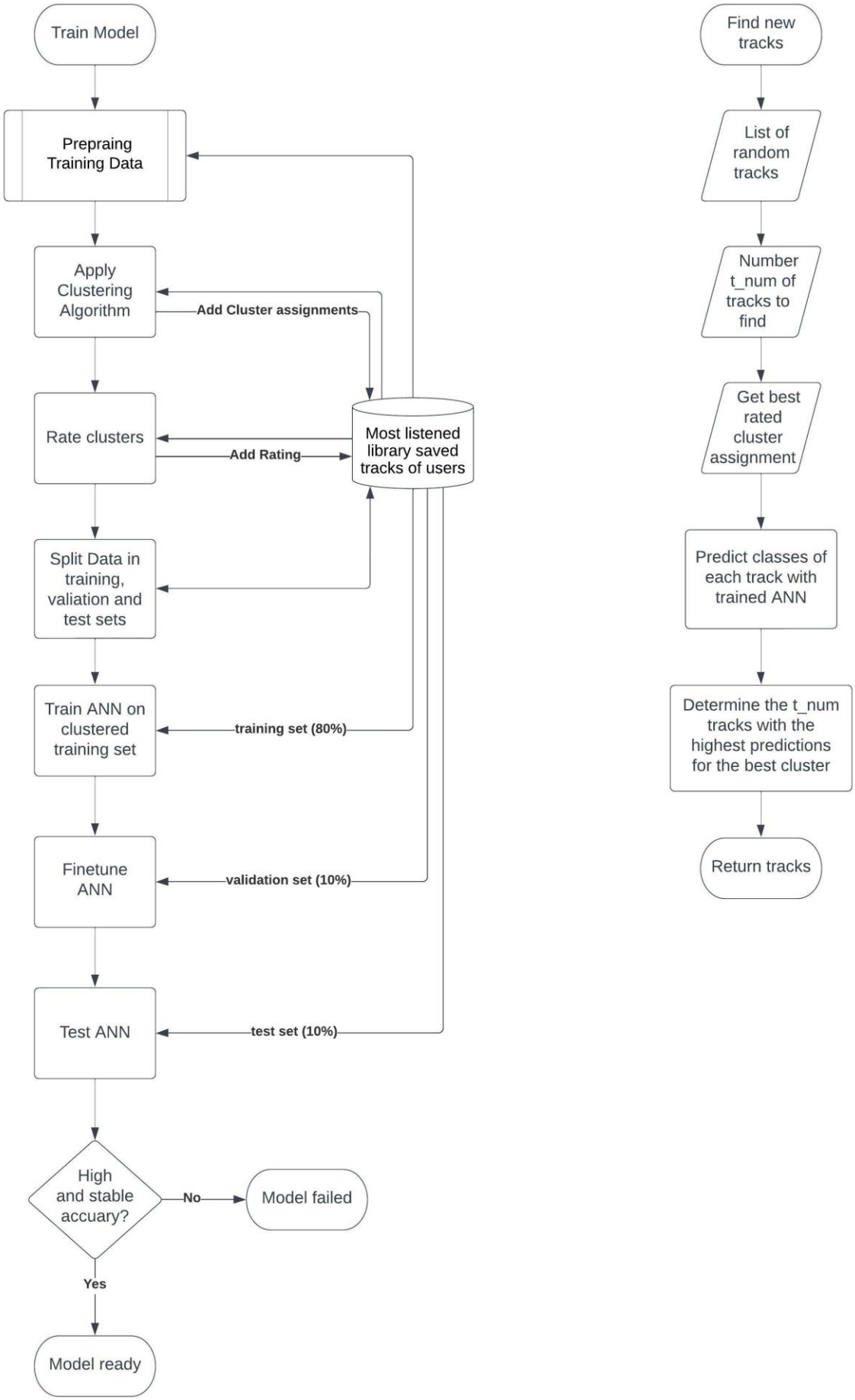
Hybrid Model Architecture:

Clustering:

- Try to find the most common similarities of the tracks by
    1. grouping the tracks into clusters based on the similarity of the track data
    2. finding the most promising cluster to use it as an ideal feature characteristics representative
        - Rate by factors number of tracks, listening count, user balancing by multiplying:
            - Percentage of covered tracks
            - Percentage of covered number of times the tracks were listened
            - Normalized standard deviation of

Classification / ANN

- Train a Classification model (here ANN) on with the track data as input
- Use the cluster assignments as class to learn
- Predict the class of new songs → The higher the predicted value for the most promising cluster, the more suitable it is for the playlist


Train Model
Prepraing Training Data
Apply Clustering Algorithm
Add Cluster assignments
Rate clusters
Add Rating
Most listened library saved tracks of users
Split Data in training, valiation and test sets
Train ANN on clustered training set
training set (80%)
Finetune ANN
validation set (10%)
Test ANN
test set (10%)
High and stable accuary?
No
Model failed
Yes
Model ready
Find new tracks
List of random tracks
Number t_num of tracks to find
Get best rated cluster assignment
Predict classes of each track with trained ANN
Determine the t_num tracks with the highest predictions for the best cluster
Return tracks

Requirements:

Data Requirements:

- At least 100 tracks from each user (with labeling from which users library the track is)
- There should be the same amount oft tracks from each user
- Track data should be from tracks that the users saved in their libraries and listened to the most
- Track data should include all metadata and analysis data available
    - Artist
    - Genre
    - Publishing date
    - Length
    - Bpm
    - Mood
    - …