Università degli Studi di Bari Aldo Moro

Dataset LFM-1b_artist

Emanuele Fontana

Tirocinio tesi triennale in Informatica
Anno accademico 2023/2024

Indice

1 Statistiche dei dataset

LFM1b_artist è un dataset contenente informazioni riguardanti le interazioni tra utenti e artisti musicali.

1.1 Dataset originale

Descrizione del dataset

Feature	Descrizione
n_users	120322
n_items	3123496
n_inter	65133026
sparsity	0.9998266933373666
avg_inter_user	541.3226675088513

Tabella 1: Informazioni sul dataset LFM1b_artist

Descrizione del knowledge graph

n_ent_head	823213
n_ent_tail	353607
n_rel	8
n_triple	2114049

Tabella 2: Informazioni sul knowledge graph del dataset LFM1b_artist

I nomi delle relazioni presenti nel knowledge graph sono i seguenti:

- music.recording.artist
- music.recording.releases
- music.recording.producer
- music.recording.engineer
- music.recording.featured_artists
- music.featured_artist.recordings
- music.release.artist
- music.artist.release

1.2 Processing del dataset

1.2.1 Descrizione procedimento

Il dataset originale risultava essere troppo grande per le risorse a nostra disposizione, dunque è stato opportunamente processato. In paritcolare sono state svolte le seguenti operazioni

- **Filtraggio:** il dataset è stato filtrato eliminando tutte le interazioni in cui erano coinvolti utenti e/o item con meno di 5 interazioni

- **Sampling:** dopo la fase di filtraggio, è stato effettuato un sampling casuale il cui scopo era quello di ridurre il numero di utenti e di item presenti. In particolare sono stati selezionati casualmente 20000 utenti e 50000 item e sono state mantenute solo le interazioni in cui erano coinvolti utenti e item selezionati

In questo modo è stato ottenuto un dataset più piccolo e più facilmente gestibile rispetto a quello originale. Per poter lavorare su più dataset si è deciso di effettuare un ulteriore processing del dataset, andando a creare dei sampling con una strategia di stratificazione: [1]

- **75%:** Per ogni utente sono state mantenute il 75% delle interazioni originali
- **50%:** Dal dataset al 75% sono state mantenute circa il 66.67% delle interazioni di ogni utente, in modo tale da avere il 50% delle interazioni originali
- **25%:** Dal dataset al 50% sono state mantenute il 50% delle interazioni di ogni utente, in modo tale da avere il 25% delle interazioni originali

1.2.2 Dataset core5

Descrizione del dataset

Feature	Descrizione
n_users	120175
n_items	585095
n_inter	61534450
sparsity	0.9991248594539152
avg_inter_user	512.0403578115248

Tabella 3: Informazioni sul dataset LFM1b_artist_core5

Descrizione del knowledge graph

n_ent_head	823213
n_ent_tail	353607
n_rel	8
n_triple	2114049

Tabella 4: Informazioni sul knowledge graph del dataset LFM1b_artist_core5

I nomi delle relazioni presenti nel knowledge graph sono i seguenti:

- music.recording.artist
- music.recording.releases
- music.recording.producer
- music.recording.engineer
- music.recording.featured_artists
- music.featured_artist.recordings
- music.release.artist
- music.artist.release

[1] Mantenendo il numero di utenti inalterato per ognuno di essi sono stati campionati casualmente un determinato numero di interazioni cercando di mantenere inalterati i "rapporti originali" tra i diversi utenti

1.2.3 Dataset 20.000 users, 50.000 items

Descrizione del dataset

Feature	Descrizione
n_users	19841
n_items	42457
n_inter	900212
sparsity	0.9989313587429705
avg_inter_user	45.371301849705155

Tabella 5: Informazioni sul dataset LFM1b_artist_20U50I

Descrizione del knowledge graph

n_ent_head	15509
n_ent_tail	35156
n_rel	5
n_triple	46827

Tabella 6: Informazioni sul knowledge graph del dataset LFM1b_artist_20U50I

I nomi delle relazioni presenti nel knowledge graph sono i seguenti:

- music.recording.artist
- music.recording.releases
- music.recording.producer
- music.recording.engineer
- music.recording.featured_artists

1.2.4 Dataset 75%

Descrizione del dataset

Feature	Descrizione
n_users	19841
n_items	38932
n_inter	667850
sparsity	0.9991354130849345
avg_inter_user	33.660097777329774

Tabella 7: Informazioni sul dataset LFM1b_artist_20U50I_75strat

Descrizione del knowledge graph

n_ent_head	14327
n_ent_tail	32981
n_rel	5
n_triple	43559

Tabella 8: Informazioni sul knowledge graph del dataset LFM1b_artist_20U50I_75strat

I nomi delle relazioni presenti nel knowledge graph sono i seguenti:

- music.recording.artist
- music.recording.releases
- music.recording.producer
- music.recording.engineer
- music.recording.featured_artists

1.2.5 Dataset 50%

Descrizione del dataset

Feature	Descrizione
n_users	19841
n_items	33653
n_inter	440620
sparsity	0.9993401019218887
avg_inter_user	22.20755002268031

Tabella 9: Informazioni sul dataset LFM1b_artist_20U50I_50strat

Descrizione del knowledge graph

n_ent_head	12522
n_ent_tail	29509
n_rel	5
n_triple	38491

Tabella 10: Informazioni sul knowledge graph del dataset LFM1b_artist_20U50I_50strat

I nomi delle relazioni presenti nel knowledge graph sono i seguenti:

- music.recording.artist
- music.recording.releases
- music.recording.producer
- music.recording.engineer
- music.recording.featured_artists

1.2.6 Dataset 25%

Descrizione del dataset

Feature	Descrizione
n_users	19841
n_items	24878
n_inter	218457
sparsity	0.9995574249320202
avg_inter_user	11.01038254120256

Tabella 11: Informazioni sul dataset LFM1b_artist_20U50I_25strat

Descrizione del knowledge graph

n_ent_head	9444
n_ent_tail	23463
n_rel	5
n_triple	29822

Tabella 12: Informazioni sul knowledge graph del dataset LFM1b_artist_20U50I_25strat

I nomi delle relazioni presenti nel knowledge graph sono i seguenti:

- music.recording.artist
- music.recording.releases
- music.recording.producer
- music.recording.engineer
- music.recording.featured_artists

2 Configurazione

2.1 Parametri di running

Qui di seguito vengono riportati i parametri di running dei vari esperimenti effettuati.

Parametri di environment

I parametri di environment servono per configurare l'ambiente di esecuzione.

- **gpu_id**: 0
- **worker**: 0
- **use_gpu**: True
- **seed**: 2020
- **state**: INFO
- **encoding**: utf-8
- **reproducibility**: True
- **shuffle**: True

Parametri di training

I parametri di training servono per l'addestramento dei modelli.

- **epochs**: 200
- **train_batch_size**: 2048
- **learner**: adam
- **learning_rate**: .001
- **train_neg_sample_args**:
  - **distribution**: uniform
  - **sample_num**: 1
  - **dynamic**: False
  - **candidate_num**: 0
- **eval_step**: 1
- **stopping_step**: 10
- **clip_grad_norm**: None
- **loss_decimal_place**: 4
- **weight_decay**: .0
- **require_pow**: False
- **enable_amp**: False
- **enable_scaler**: False

Parametri di evaluation

I parametri di evaluation servono per valutare i modelli.

- **eval_args**:
-
  - **group_by**: user
  - **order**: RO
  - **split**: RS : [0.8, 0.1, 0.1]
  - **mode**: full
- **repeatable**: False
- **metrics**: ['Recall', 'MRR', 'NDCG', 'Hit', 'MAP', 'Precision', 'GAUC', 'ItemCoverage', 'AveragePopularity', 'GiniIndex', 'ShannonEntropy', 'TailPercentage']
- **topk**: 10
- **valid_metric**: MRR@10
- **eval_batch_size**: 4096
- **metric_decimal_place**: 4

Iper parametri dei modelli

Gli iper parametri dei modelli sono un insieme di parametri che vengono utilizzati per configurare i modelli. La loro configurazione può influenzare il risultato finale. Esistono delle tecniche di HyperTuning che permettono di trovare i migliori iper parametri per un determinato modello e dataset. In questo caso si è scelto di utilizzare gli iper parametri di default

3 Emissioni

Qui di seguito vengono riportate le emissioni di CO2 per ogni esperimento effettuato.

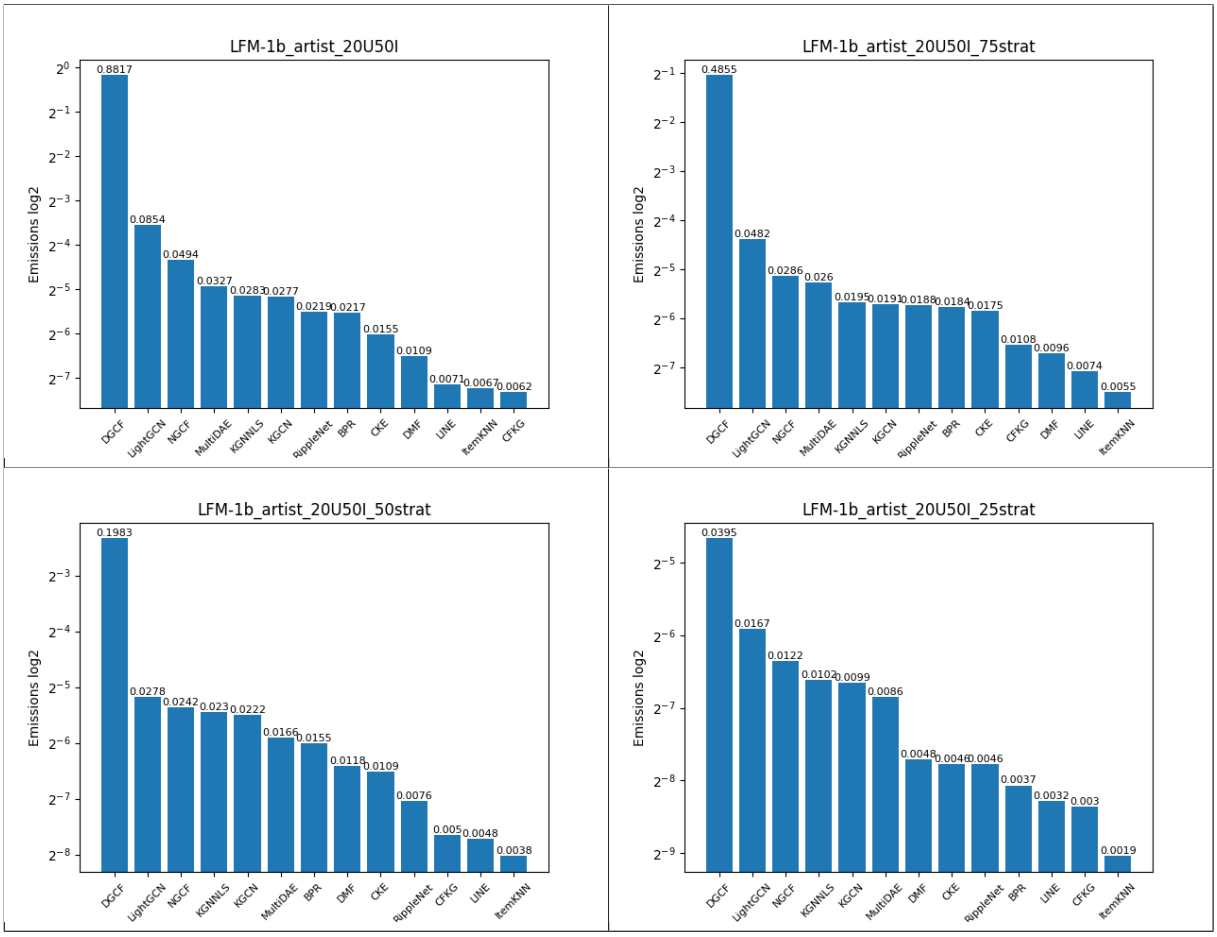

Tabella 13: Emissioni di CO2 per i vari dataset

Si può subito notare come DGCF è il modello che emette più CO2 in assoluto. In particolare con il dataset al 100% e al 75% DGCF emette circa 10 volte di più rispetto a LightGCN (il secondo per emissioni) mentre con il dataset al 50% emette circa 7 volte di più e con il dataset al 25% emette circa 2 volte di più (sempre rispetto a LightGCN). LightGCN e NCFG sono rispettivamente il secondo e il terzo modello che emettono più CO2. Questi due modelli sono invece di tipo general, ma nonostante ciò emettono di più rispetto ad altri di tipo knowledge-aware, come per esempio il KGCN. In generale possiamo vedere che ItemKNN,LINE e CFKG sono i modelli che emettono meno. Per LINE e ItemKNN questo era abbastanza prevedibile in quanto modelli di tipo General. Interessante invece notare come CFKG, di tipo knowledge-aware, emetta meno di altri modelli di tipo General

4 Trade-off

4.1 Introduzione

In questa sezione verranno analizzati i trade-off tra le varie metriche di valutazione e le emissioni di CO2 analizzando un dataset per volta.

Di seguito un elenco delle metriche con una piccola descrizione:

- Recall: è una metrica che misura la capacità di un modello di raccomandare gli item rilevanti per un utente
- NDCG: è una metrica che misura la qualità delle raccomandazioni.
- Gini Index: è una metrica che misura l'equità nella distribuzione delle raccomandazioni. Un valore più vicino a zero indica una distribuzione più equa
- Average Popularity: è una metrica che misura la popolarità media degli item raccomandati. Un valore alto indica che le raccomandazioni sono concentrate su item popolari.

4.2 LFM-1b_artist_20U50I

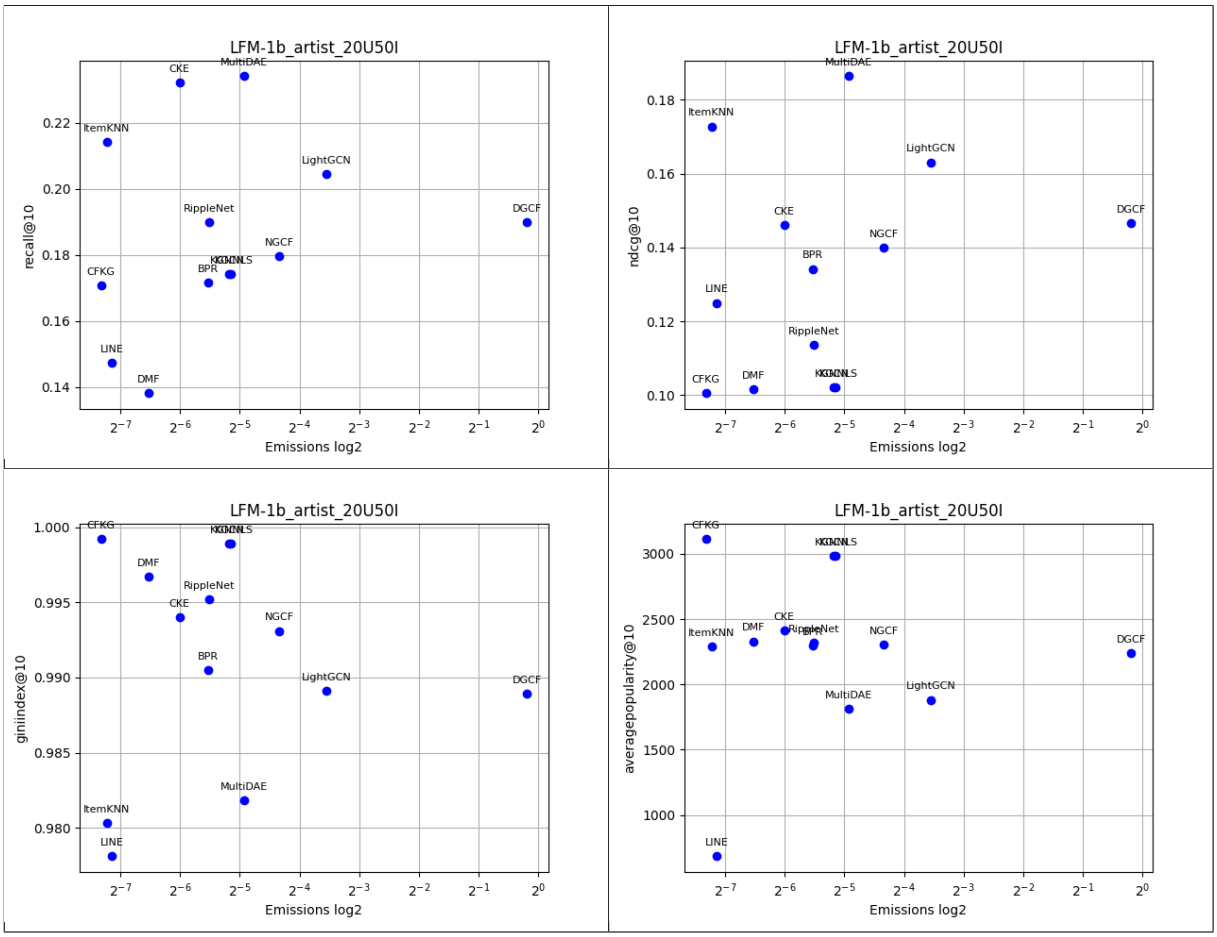

Tabella 14: Trade-off con il dataset LFM-1b_artist_20U50I

Come già visto precedentemente, DGCF è il modello che emette di più. Nonostante ciò possiamo notare che per la recall e l'ndcg le sue performance risultano peggiori rispetto ad

algoritmi più semplici come l'ItemKNN che risulta essere uno degli algoritmi che emette meno e performa meglio in queste metriche. Per quanto riguarda il Gini Index possiamo notare che DGCF si comporta meglio di molti altri modelli ma l'ItemKNN e LINE risultano essere migliori di quest'ultimo. LINE è il miglior algoritmo. Infine, per quanto riguarda l'Average Popularity, anche in questo caso possiamo notare anche che DGCF performa meglio di altri modelli, ma LINE risulta il miglior in assoluto ed è uno degli algoritmi che emette meno.

4.3 LFM-1b_artist_20U50I_75strat

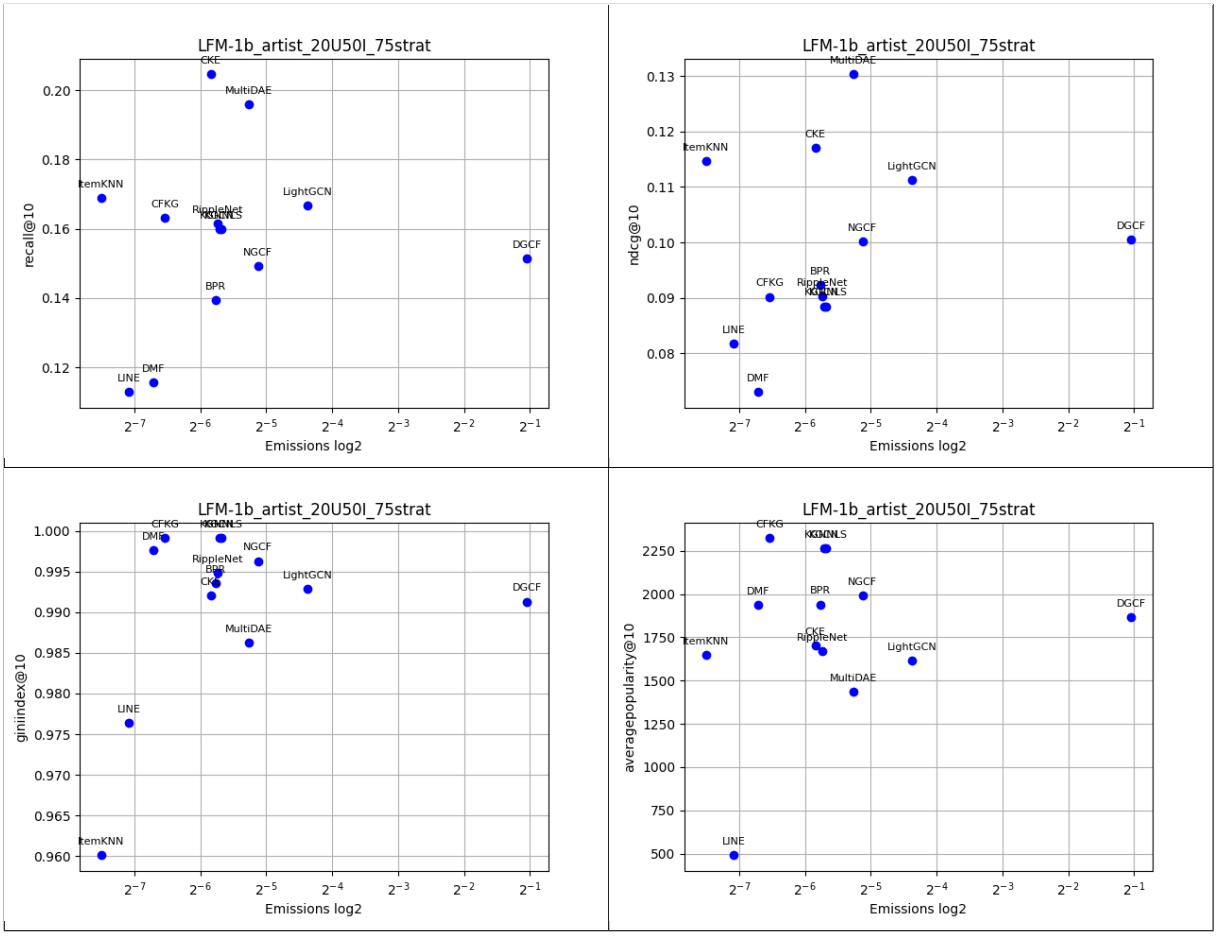

Tabella 15: Trade-off con il dataset LFM-1b_artist_20U50I

Come già visto precedentemente, DGCF è il modello che emette di più. Nonostante ciò possiamo notare che per la recall e l'ndcg le sue performance risultano peggiori rispetto ad algoritmi più semplici come l'ItemKNN che risulta essere uno degli algoritmi che emette meno e performa meglio in queste metriche. Per quanto riguarda il Gini Index possiamo notare che DGCF si comporta meglio di molti altri modelli ma l'ItemKNN e LINE risultano essere migliori di quest'ultimo. ItemKNN è il miglior algoritmo. Infine, per quanto riguarda l'Average Popularity, anche in questo caso possiamo notare anche che DGCF performa meglio di altri modelli, ma LINE risulta il miglior in assoluto ed è uno degli algoritmi che emette meno.

4.4 LFM-1b_artist_20U50I_50strat

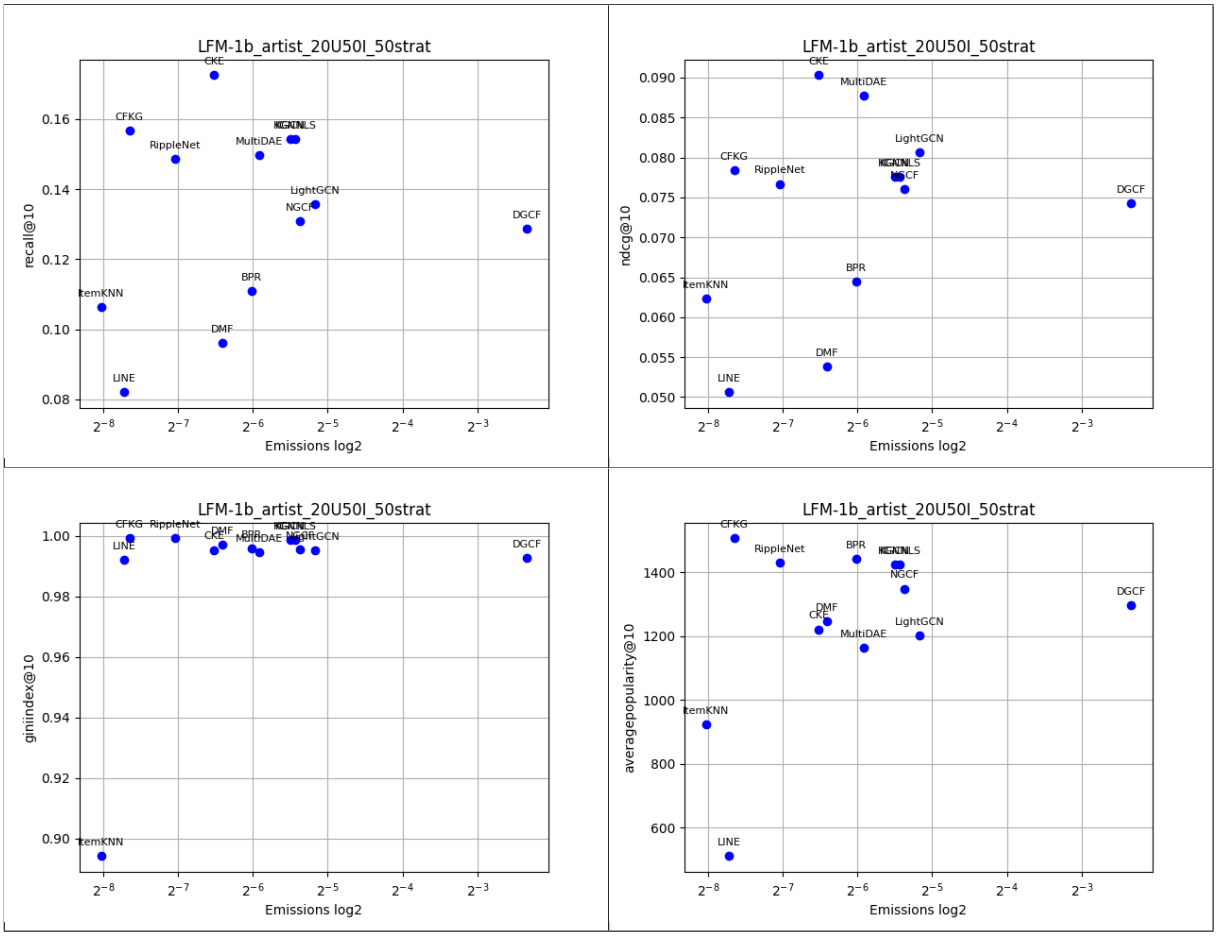

Tabella 16: Trade-off con il dataset LFM-1b_artist_20U50I

Come già visto precedentemente, DGCF è il modello che emette di più. Nonostante ciò possiamo notare che per la recall e l'ndcg le sue performance risultano peggiori rispetto ad altri algoritmi che emettono meno come CKE e CKFG(anch'essi di tipo Knowledge-Aware).. Per quanto riguarda il Gini Index possiamo notare che DGCF si comporta meglio di molti altri modelli ma l'ItemKNN risulta essere migliore di quest'ultimo ed il migliore in assoluto. Infine, per quanto riguarda l'Average Popularity, anche in questo caso possiamo notare anche che DGCF performa meglio di altri modelli, ma LINE risulta il miglior in assoluto ed è uno degli algoritmi che emette meno.

4.5 LFM-1b_artist_20U50I_25strat

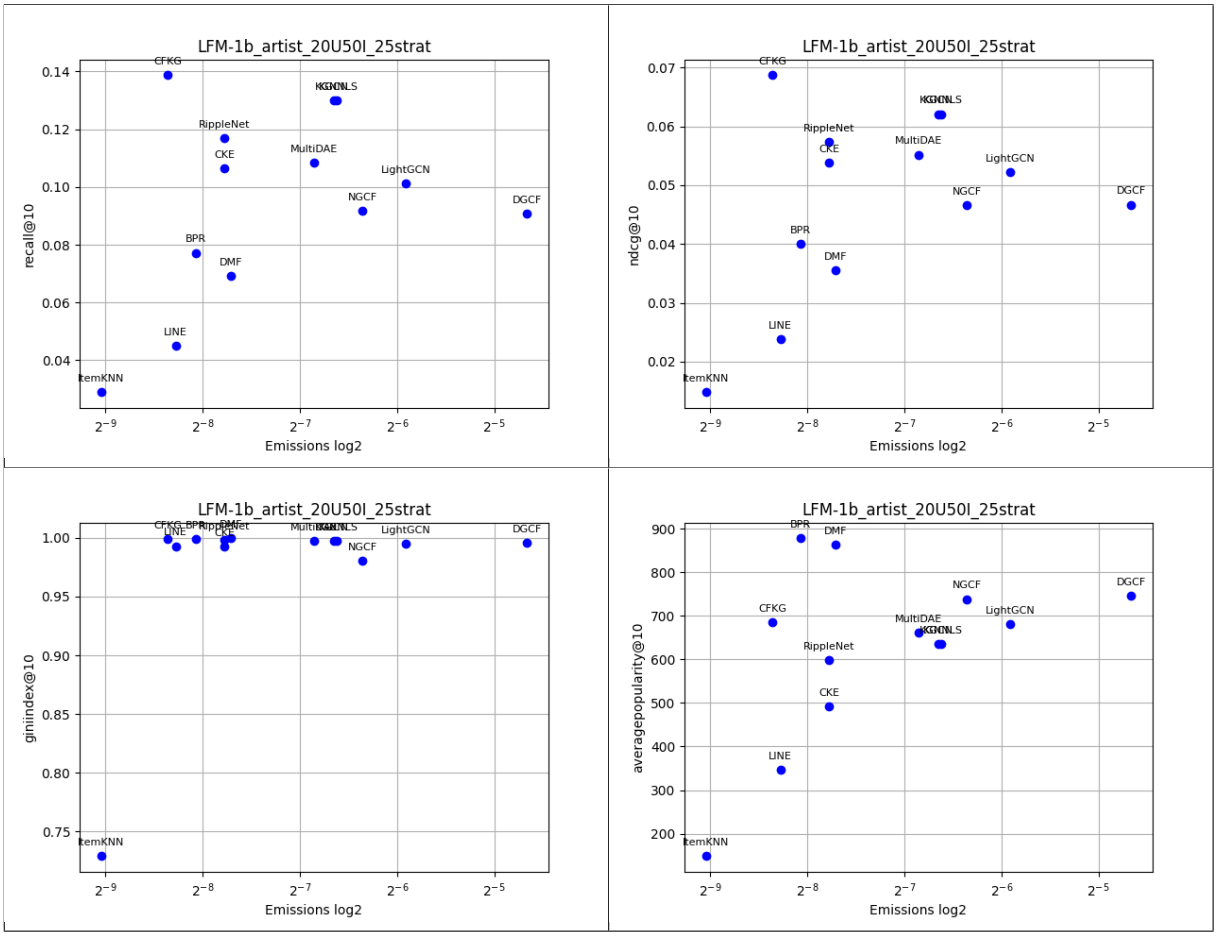

Tabella 17: Trade-off con il dataset LFM-1b_artist_20U50I

Come già visto precedentemente, DGCF è il modello che emette di più. Nonostante ciò possiamo notare che per la recall e l'ndcg le sue performance risultano peggiori rispetto ad altri algoritmi che emettono meno come CKE e CKFG (anch'essi di tipo Knowledge-Aware). Per quanto riguarda il Gini Index possiamo notare che DGCF si comporta meglio di molti altri modelli ma l'ItemKNN risulta essere di quest'ultimo migliore ed il migliore in assoluto. Infine, per quanto riguarda l'Average Popularity, in questo caso possiamo notare anche che DGCF è uno dei peggiori mentre ItemKNN risulta il miglior in assoluto ed è l'algoritmo che emette meno.

5 Conclusioni

Si può facilmente notare come il trade-off emissioni-performance sia decisamente a svantaggio dell'DGCF. Infatti, a fronte di emissioni molto elevate, le performance risultato spesso essere peggiori di modelli molto più semplici. Con i due dataset più grandi possiamo notare come in generale ItemKNN risulti essere uno degli algoritmi con il miglior trade-off emissioni-performance nelle metriche di ranking, mentre LINE risulta essere il migliore nelle metriche di popolarità e equità nelle distribuzioni. Al diminuire della dimensione del dataset DGCF comincia a comportarsi meglio nelle metriche di popolarità e equità, ma le sue emissioni rimangono sempre molto alte e non giustificano una possibile scelta di questo modello. ItemKNN comincia a non performare bene nelle metriche di ranking, mentre migliora nelle metriche di popolarità e equità, arrivando anche a risultare il migliore