



Università degli Studi di Bari Aldo Moro

Relazione tecnica Sustainability of RecSys

Emanuele Fontana

Tirocinio tesi triennale in Informatica
Anno accademico 2023/2024

Indice

1	Introduzione	3
2	Dataset del regressore	5
2.1	Descrizione delle feature di output	5
2.2	Descrizione delle feature di input	5
2.3	Pre-Processing	8
3	Modelli di regressione	9
3.1	Introduzione ai regressori	9
3.2	Regressori utilizzati	9
3.2.1	Support Vector Regression (SVR)	9
3.2.2	Decision Tree Regressor	10
3.2.3	Random Forest Regressor	11
3.2.4	AdaBoost Regressor	13
4	Risultati	14
4.1	Risultati ottenuti	14
4.2	Analisi dei risultati	15
4.3	Conclusioni	15
5	Statistiche dei dataset	16
5.1	Dataset LFM-1b_artist	16
5.1.1	Processing del dataset	16
5.2	Statistiche dei dataset MovieLens10M	19
5.2.1	Processing del dataset	19

6	Configurazione	21
6.1	Parametri di running	21
7	Emissioni	23
7.1	LFM-1b_artist	23
7.2	ml-10m	24
8	Trade-off	25
8.1	Introduzione	25
8.2	LFM-1b_artist_20U50I	25
8.3	LFM-1b_artist_20U50I_75strat	26
8.4	LFM-1b_artist_20U50I_50strat	27
8.5	LFM-1b_artist_20U50I_25strat	28
8.6	ml-10m_50U10I	29
8.7	Conclusioni	29
9	Regressore	31
9.1	Dataset originale	31
9.1.1	Dataset completo	31
9.1.2	Dataset Azure	40
9.2	Dataset ridotto	50
9.2.1	Dataset completo ridotto	50
9.2.2	Dataset AZURE	59
9.2.3	Confronto tra i quattro dataset	68
10	Errori delle classi	69
10.1	Dataset Completo	69
10.2	Dataset Completo Ridotto	69
10.3	Dataset Azure	70
10.4	Dataset Azure Ridotto	70

1 Introduzione

Tracciare le emissioni degli algoritmi di raccomandazione e cercare di prevederle è molto importante quando si parla di sviluppo sostenibile in campo RecSys. Ancora oggi si tende a trascurare l'impatto ambientale di un'attività e, in questo ambito, si è molto propensi nell'utilizzare dei modelli molto complessi e pesanti che richiedono molte risorse per essere addestrati ed eseguiti per ottenere delle buone performance. Spesso, però, modelli molto più leggeri e semplici riescono a ottenere delle performance molto simili (se non superiori) a modelli più complessi e il tutto con un impatto ambientale decisamente minore. Ad oggi il carbon dioxide equivalent (CO₂eq) è il principale indicatore utilizzato da governi e enti per misurare l'impatto ambientale di un'attività. Il CO₂eq è un'unità di misura che esprime l'equivalente in CO₂ di tutti i gas serra emessi da un'attività, in modo da poter confrontare l'impatto ambientale di attività diverse. Una strategia comune per calcolare il CO₂eq è quella di moltiplicare tra loro il **carbon intensity(CI)** e l'**energia consumata(PC)** dall'attività (nel nostro caso l'esecuzione di algoritmi).

$$emission = CI \cdot PC$$

In particolare i valori di CI dipendono dalle diverse fonti di energia utilizzate durante la computazione (es. energia solare, energia eolica, etc.). Se s è la fonte di energia, e_s sono le emissioni per KW/h di energia e p_s è la percentuale di energia prodotta dalla fonte s , allora il CI è dato da:

$$CI = \sum_{s \in S} e_s \cdot p_s$$

Lo scopo di questo lavoro è quello di valutare e prevedere l'impatto ambientale di un sistema di raccomandazione (RecSys) in base alla sua sostenibilità. Per quanto riguarda la valutazione, mediante la libreria CodeCarbon¹ è stato possibile misurare le emissioni prodotte dalla macchina durante l'addestramento con parametri di default per un dato modello dato un dataset. In questo ambito Spillo et al.[19] mostrano come spesso algoritmi più semplici riescono ad avere delle performance molto simili a modelli più complessi, ma con un impatto ambientale decisamente minore.

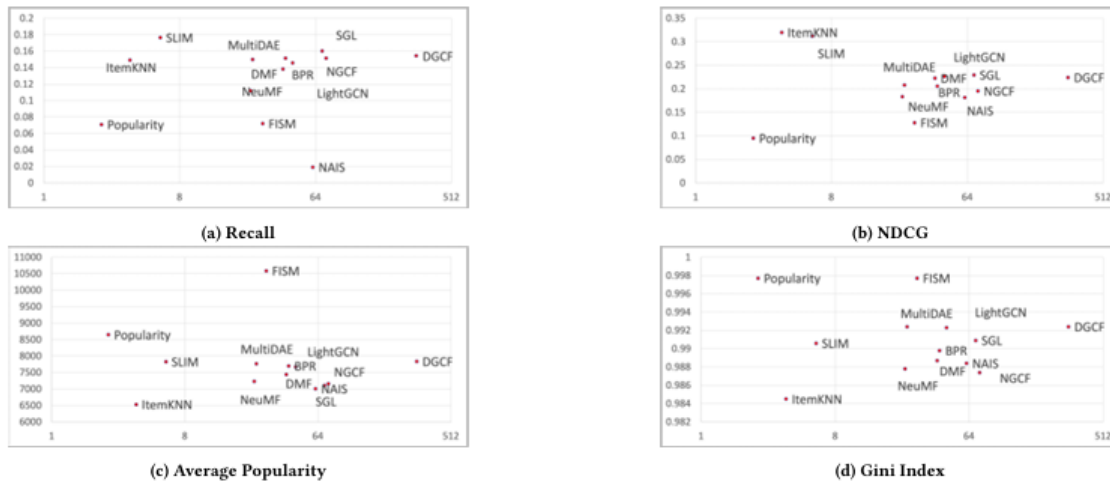


Figura 1: Trade-off tra emissioni e performance con dataset Mind

Per quanto riguarda la previsione questa prima parte del documento si propone di presentare lo stato attuale del lavoro svolto in questo ambito.

¹CodeCarbon

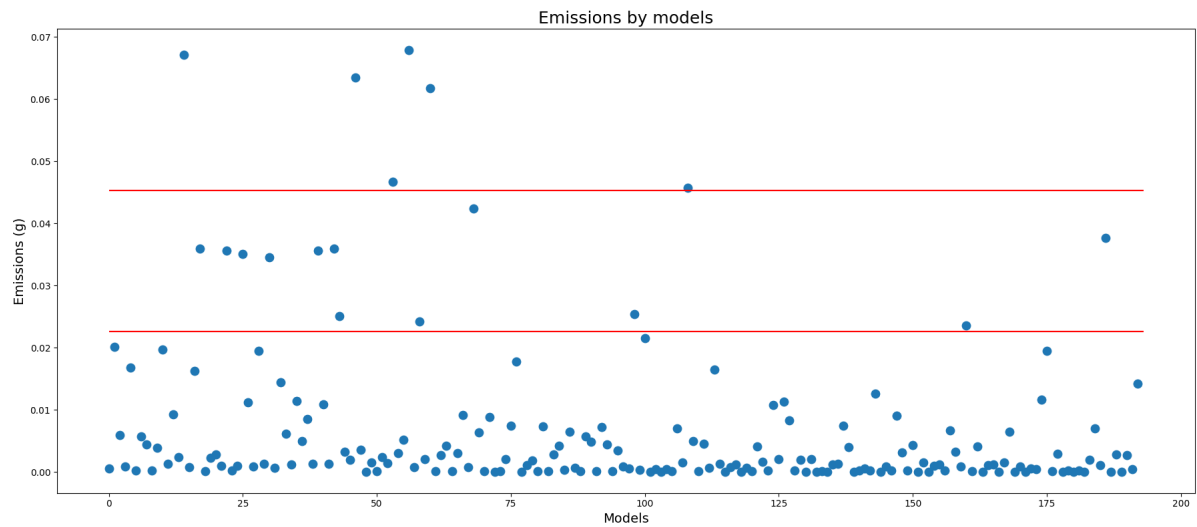


Figura 2: Emissioni prodotte dai vari modelli

L'esperimento è stato condotto su dataset presenti all'interno della libreria python RecBole², una libreria open-source che offre un'implementazione di modelli di raccomandazione. I dataset utilizzati per l'addestramento dei modelli sono:

- **MovieLens-1M**³
- **Amazon_Book_60core_kg**⁴
- **Mind**⁵

²RecBole

³Dataset MovieLens

⁴Dataset Amazon_Book_60core_kg

⁵Dataset MIND

2 Dataset del regressore

In questo capitolo si analizza il dataset utilizzato e come questo è stato trattato per l'addestramento dei modelli. Inoltre, si descrivono le feature di input e output del modello.

Il dataset nella sua totalità è composto da 13 feature di input e una feature di output. Le feature di input possiamo suddividerle in 4 categorie:

- **Feature relative al dataset**, quali *n_users*, *n_items*, *n_inter*, *sparsity*
- **Feature relative al knowledge graph**, quali *kg_entities*, *kg_relations*, *kg_triples*, *kg_items*
- **Feature relative all'hardware utilizzato per l'addestramento**, quali *cpu_cores*, *ram_size*, *is_gpu*
- **Feature relative al modello**, quali *model_name*, *model_type*

Nel dataset sono presenti 201 righe (dunque 201 esperimenti distinti).

2.1 Descrizione delle feature di output

La feature di output *emissions* rappresenta le emissioni di CO₂eq prodotte dalla macchina durante l'addestramento del modello.

2.2 Descrizione delle feature di input

Feature	Descrizione
<i>n_users</i>	Numero di utenti presenti nel dataset
<i>n_items</i>	Numero di items presenti nel dataset
<i>n_inter</i>	Numero di interazioni nel dataset. Per interazione si intendono le varie interazioni (valutazioni) tra gli utenti nel dataset e gli item nel dataset
<i>sparsity</i>	Sparsità del dataset. La sparsità indica la percentuale di valori mancanti nel dataset (quindi mancanza di interazione tra utenti e item)
<i>kg_entities</i>	Numero di entità nel knowledge graph. Un'entità è un oggetto distintivo o un concetto unico all'interno del Knowledge Graph
<i>kg_relations</i>	Numero di relazioni nel knowledge graph. Le relazioni rappresentano i legami o collegamenti tra le entità all'interno del Knowledge Graph. Sono spesso definite dai predicati nelle triple
<i>kg_triples</i>	Numero di triple nel knowledge graph. Una triple è una struttura dati fondamentale nel Knowledge Graph che consiste in tre parti: soggetto, predicato e oggetto. Queste triple rappresentano le relazioni tra le entità
<i>kg_items</i>	Numero di items nel knowledge graph. Gli "Items" nel contesto del Knowledge Graph sono gli oggetti specifici o le entità che sono inclusi nel grafo
<i>cpu_cores</i>	Numero di core della CPU
<i>ram_size</i>	Dimensione della RAM
<i>is_gpu</i>	Booleano che indica se la macchina ha usato una GPU per l'addestramento
<i>model_name</i>	Nome del modello
<i>model_type</i>	Tipo del modello

Tabella 1: Descrizione delle feature di input

Per quanto riguarda la feature *model_type* abbiamo i seguenti valori:

- **General**: Modelli che si basano su tecniche tradizionali
- **Knowledge**: Modelli che incorporano conoscenza esterna (knowledge graph) per migliorare le raccomandazioni

Per quanto riguarda la feature *model_name* abbiamo i seguenti valori:

- **BPR** [17]: General
- **CDAE** [35]: General
- **CFKG** [1]: Knowledge
- **CKE** [37]: Knowledge
- **DGCF** [32]: Knowledge
- **DMF** [36]: General
- **DiffRec** [28]: General
- **ENMF** [5]: General
- **FISM** [10]: General
- **GCMC** [23]: General
- **ItemKNN** [2]: General
- **KGCN** [27]: Knowledge
- **KGIN**: [31] Knowledge
- **KGNNLS** [25]: Knowledge
- **KTUP** [4]: Knowledge
- **LDiffRec** [29]: General
- **LINE** [22]: General
- **LightGCN** [6]: General
- **MKR** [26]: Knowledge
- **MacridVAE** [14]: General
- **MultiDAE** [11]: General
- **MultiVAE** [12]: General
- **NCEPLRec** [33]: General
- **NCL** [13]: General
- **NGCF** [30]: General

- **NeuMF** [9]: General
- **Pop**: General
- **Random**: General
- **RecVAE** [18]: General
- **RippleNet** [24]: Knowledge
- **SGL** [34]: General
- **SLIMElastic** [16]: General
- **SimpleX** [15]: General
- **SpectralCF** [38]: General
- **EASE** [20]: General
- **NAIS** [8]: General
- **ADMMSLIM** [21]: General
- **ConvNCF** [7]: General
- **NNCF** [3]: General

Altri valori unici presenti per le varie feature di input sono:

- **n_users**: [22155, 23679, 6040]
- **n_items**: [54458, 4414, 3706]
- **n_inter**: [1465871, 1048575, 1000209]
- **sparsity**: [0.99878504, 0.98996762, 0.95531637]
- **kg_entities**: [26315, 0, 79347]
- **kg_relations**: [16, 0, 49]
- **kg_triples**: [96476, 0, 385923]
- **kg_items**: [11446, 0, 3655]
- **cpu_cores**: [12, 4]
- **ram_size**: [64, 16, 27.40581512]
- **is_gpu**: [1, 0]

2.3 Pre-Processing

Per poter sfruttare le feature di input per l'addestramento del modello, è stato necessario effettuare un pre-processing. Le feature *model_name* e *model_type* sono state trasformate in variabili numeriche. In particolare il valore *general* è stato trasformato in 0 e il valore *knowledge* è stato trasformato in 1. Per quanto riguarda la feature *model_name*, ogni valore è stato trasformato in un numero intero univoco. In questo modo, il modello può sfruttare queste feature per l'addestramento. Prima di cominciare con l'addestramento dei modelli la feature di output è stata separata dalle feature di input. I dati sono poi stati suddivisi rispettivamente in training set e test set. In particolare il 70% dei dati è stato usato per l'addestramento, mentre il 30% è stato usato per la valutazione. Inoltre, mediante il *random_state=2*, è garantita la riproducibilità dell'esperimento.

3 Modelli di regressione

3.1 Introduzione ai regressori

I regressori sono dei modelli di machine learning il cui obiettivo è quello di prevedere un valore numerico continuo (in questo caso il valore delle emissioni). I regressori vengono valutati sulla base di metriche quali, per esempio, l'errore quadratico medio (MSE), l'errore assoluto medio (MAE), e l'errore quadrato logaritmico medio (MSLE).

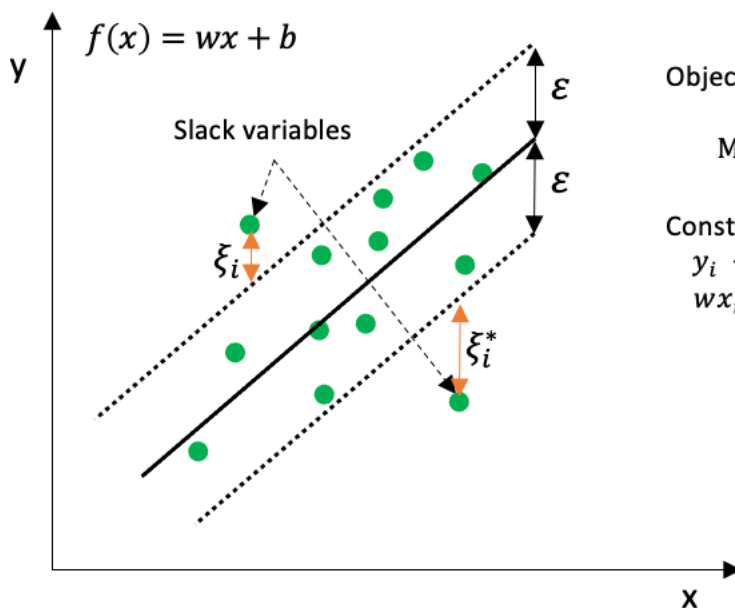
3.2 Regressori utilizzati

In questo lavoro sono stati utilizzati i seguenti regressori:

- **Support Vector Regression (SVR)**⁶
- **Decision Tree Regressor**⁷
- **Random Forest Regressor**⁸
- **AdaBoost Regressor**⁹

3.2.1 Support Vector Regression (SVR)

La SVR è un algoritmo che estende il concetto di Support Vector Machine (SVM)¹⁰ al caso della regressione.



Objective:

$$\text{Minimize: } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*)$$

Constraints:

$$\begin{aligned} y_i - wx_i - b &\leq \varepsilon + \xi_i \\ wx_i + b - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned}$$

Univariate linear SVR (allowing for errors)

Esempio di SVR

⁶SVR

⁷Decision Tree Regressor

⁸Random Forest Regressor

⁹AdaBoost Regressor

¹⁰Algoritmi di classificazione che cercano di trovare un iperpiano ottimale per ottenere separabilità lineare

In questo lavoro il modello è stato costruito usando gli iperparametri di default. Questi ultimi sono:

- **kernel=rbf**: E' il tipo di Kernel¹¹ utilizzato. Questo kernel valuta quanto due punti siano simili sulla base della loro distanza
- **degree=3**: grado del polinomio utilizzato per la trasformazione dei dati nello spazio
- **gamma=scale**: coefficiente del kernel
- **coef0=0.0**: termine indipendente nel kernel
- **tol=0.001**: tolleranza per il criterio di arresto
- **C=1.0**: parametro di regolarizzazione
- **epsilon=0.1**: controlla la tolleranza del modello nei confronti degli errori nella predizione
- **shrinking=True**: se abilitare o meno la riduzione del set di supporto
- **cache_size=200**: dimensione della cache in MB
- **max_iter=-1**: numero massimo di iterazioni. -1 indica nessun limite

3.2.2 Decision Tree Regressor

I Decision Tree Regressors sono modelli di regressione basati su alberi decisionali. Questi algoritmi suddividono ripetutamente il set di dati in base alle caratteristiche (features) per creare una struttura ad albero che rappresenta la relazione di decisione. Ogni nodo interno dell'albero rappresenta una decisione basata su una caratteristica, e le foglie dell'albero contengono i valori di output previsti.

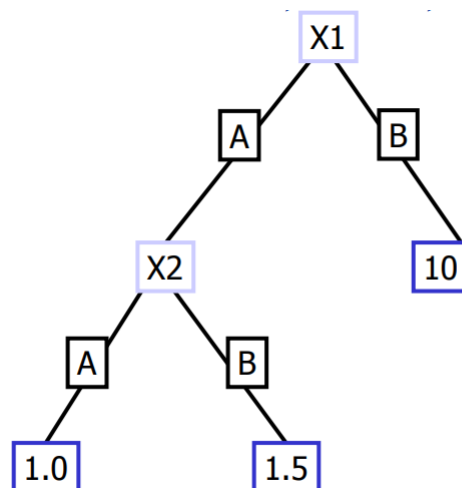


Figura 3: Esempio di albero decisionale di regressione

In questo lavoro due iperparametri sono stati decisi all'atto della costruzione del modello:

- **max_depth=5**: profondità massima dell'albero

¹¹Funzione matematica usata per mappare i dati originali in uno spazio di dimensione superiore

- **random_state=3**: controlla la casualità nell'addestramento del modello. Quando viene fissato a un numero intero specifico l'addestramento del modello sarà deterministico, cioè produrrà gli stessi risultati in ogni esecuzione

mentre gli altri iperparametri sono stati lasciati ai valori di default:

- **criterion=squared_error**: Criterio per stabilire la qualità di una suddivisione
- **splitter=best**: Strategia per scegliere la suddivisione in ogni nodo
- **min_samples_split=2**: Numero minimo di campioni richiesti per suddividere un nodo interno
- **min_samples_leaf=1**: Numero minimo di campioni richiesti per essere in una foglia
- **min_weight_fraction_leaf=0.0**: La frazione minima dei campioni totali (pesati) necessaria affinché si verifichi un nodo foglia
- **max_features=None**: Numero di features da considerare quando si cerca la migliore suddivisione. None indica tutte
- **max_leaf_nodes=None**: Numero massimo di foglie. None indica nessun limite
- **min_impurity_decrease=0.0**: Un nodo verrà suddiviso se questa suddivisione induce una diminuzione dell'impurità maggiore o uguale a questo valore
- **ccp_alpha=0.0**: Parametro di complessità usato per la potatura
- **monotonic_constraints=None**: Vincoli monotoni sui valori delle features

3.2.3 Random Forest Regressor

I Random Forest Regressors sono modelli di regressione basati su alberi decisionali. Vengono costruiti su più alberi decisionali che combinano le loro previsioni per ottenere una previsione più accurata e stabile.

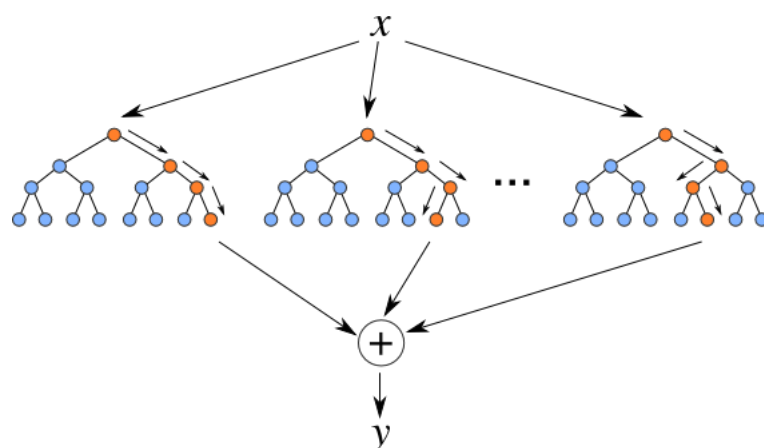


Figura 4: Esempio di Random Forest Regressor

In questo lavoro tre iperparametri sono stati decisi all'atto della costruzione del modello:

- **n_estimators=500**: numero di alberi nella foresta
- **max_depth=5**: profondità massima dell'albero

- **random_state=3**: controlla la casualità nell'addestramento del modello. Quando viene fissato a un numero intero specifico l'addestramento del modello sarà deterministico, cioè produrrà gli stessi risultati in ogni esecuzione

mentre gli altri iperparametri sono stati lasciati ai valori di default:

- **criterion=squared_error**: Criterio per stabilire la qualità di una suddivisione
- **min_samples_split=2**: Numero minimo di campioni richiesti per suddividere un nodo interno
- **min_samples_leaf=1**: Numero minimo di campioni richiesti per essere in una foglia
- **min_weight_fraction_leaf=0.0**: La frazione minima dei campioni totali (pesati) necessaria affinché si verifichi un nodo foglia
- **max_features=1**: Numero di features da considerare quando si cerca la migliore suddivisione
- **max_leaf_nodes=None**: Numero massimo di foglie. None indica nessun limite
- **min_impurity_decrease=0.0**: Un nodo verrà suddiviso se questa suddivisione induce una diminuzione dell'impurità maggiore o uguale a questo valore
- **ccp_alpha=0.0**: Parametro di complessità usato per la potatura
- **bootstrap=True**: Se utilizzare il bootstrap per la costruzione degli alberi
- **oob_score=False**: Se calcolare l'errore out-of-bag
- **n_jobs=None**: Numero di lavori da eseguire in parallelo. None indica 1
- **warm_start=False**: Se utilizzare la soluzione precedente come inizializzazione
- **max_samples=None**: Numero massimo di campioni da utilizzare per la costruzione di ciascun albero. None indica tutti
- **monotonic_constraints=None**: Vincoli monotoni sui valori delle features

3.2.4 AdaBoost Regressor

L'AdaBoost Regressor è un modello basato sull'algoritmo di boosting AdaBoost¹². Questo algoritmo costruisce un modello di previsione combinando più modelli di previsione più deboli.

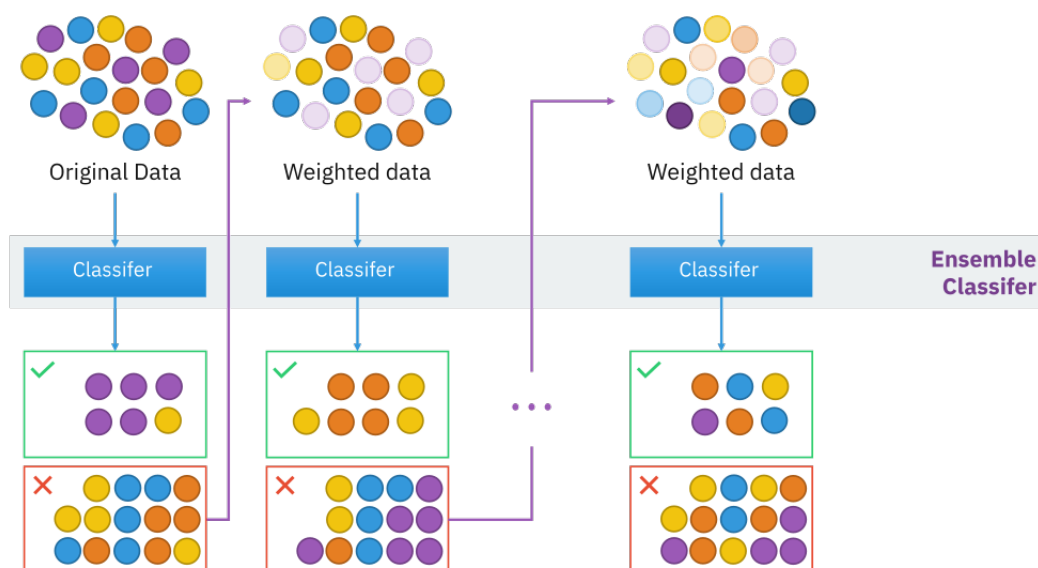


Figura 5: Struttura di modello basato su AdaBoost

In questo lavoro due iperparametri sono stati decisi all'atto della costruzione del modello:

- **n_estimators=500**: numero di stimatori
- **random_state=3**: controlla la casualità nell'addestramento del modello. Quando viene fissato a un numero intero specifico l'addestramento del modello sarà deterministico, cioè produrrà gli stessi risultati in ogni esecuzione

mentre gli altri iperparametri sono stati lasciati ai valori di default:

- **estimator=None**: stimatore base. Se None, viene utilizzato DecisionTreeRegressor
- **learning_rate=1.0**: tasso di apprendimento
- **loss=linear**: funzione di perdita

¹²Crea un modello forte combinando modelli più deboli, ognuno allenato su diverse parti del dataset

4 Risultati

In questo capitolo si analizzano i risultati degli esperimenti che sono stati condotti. I regressori utilizzati sono stati valutati utilizzando le seguenti metriche per la regressione:

- **Mean Absolute Error (MAE):** è la media della differenza assoluta tra le previsioni e i valori reali.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

- **Root Mean Squared Error (RMSE):** è la radice quadrata della media della differenza tra le previsioni e i valori reali al quadrato.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- **Mean Logarithmic Squared Error (MSLE):** è la media del logaritmo dei quadrati degli errori.

$$MSLE = \frac{1}{n} \sum_{i=1}^n (\log(y_i + 1) - \log(\hat{y}_i + 1))^2$$

4.1 Risultati ottenuti

Regressor	MAE	RMSE	MSLE
SVR	0.0288215	0.0008862	0.0008537
Decision Tree	0.0048531	0.0000969	0.0000918
Random Forest	0.0054369	0.0001088	0.0001026
AdaBoost	0.0071778	0.0001113	0.0001059

Tabella 2: Risultati ottenuti

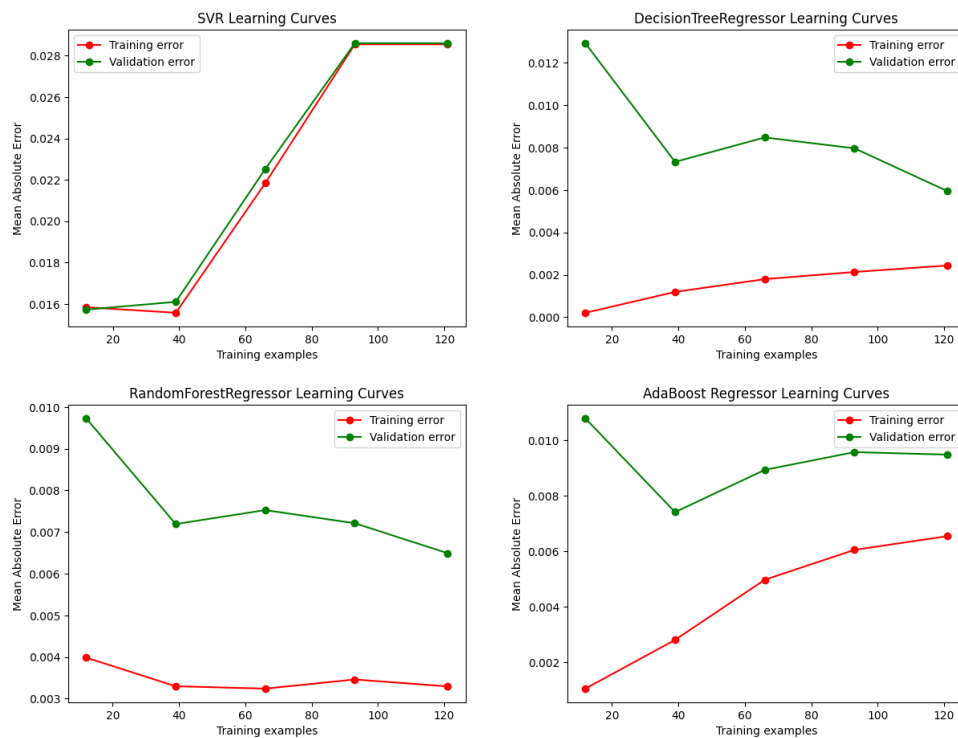


Tabella 3: Learning curve dei regressori

4.2 Analisi dei risultati

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello è in overfitting, infatti gli errori aumentano all'aumentare del numero di istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i migliori risultati. La learning curve mostra una curva di training che aumenta all'aumentare del numero di istanze, mentre la curva di validation diminuisce. Le due curve si avvicinano, ma non si sovrappongono. Sembra esserci un leggero underfitting.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il secondo miglior modello. La learning curves di train e di validation set diminuiscono entrambe all'aumentare del numero di istanze, ma non si sovrappongono. Anche qui sembra esserci underfitting.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Random Forest Regressor ma migliori rispetto al SVR. La learning curve mostrano che il modello non è riuscito a generalizzare bene

4.3 Conclusioni

Il dataset sembra essere troppo piccolo per poter generalizzare bene i modelli. Inoltre, il dataset è molto sbilanciato, con pochi valori per ogni feature. Questo potrebbe essere un motivo per cui i modelli non generalizzano bene. Inoltre, i modelli non sono stati ottimizzati, quindi potrebbero essere migliorati mediante ricerca di iperparametri. Per i risultati attuali, il Decision Tree Regressor è il modello migliore.

5 Statistiche dei dataset

5.1 Dataset LFM-1b_artist

LFM1b_artist¹³ è un dataset contenente informazioni riguardanti le interazioni tra utenti e artisti musicali.

Descrizione del dataset

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

Tabella 4: Informazioni sul dataset LFM1b_artist

Descrizione del knowledge graph

n_ent_head	823213
n_ent_tail	353607
n_rel	8
n_triple	2114049

Tabella 5: Informazioni sul knowledge graph del dataset LFM1b_artist

5.1.1 Processing del dataset

Descrizione procedimento Il dataset originale risultava essere troppo grande per le risorse a nostra disposizione, dunque è stato opportunamente processato. In particolare 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: ¹⁴

- **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

¹³Dataset LFM

¹⁴Mantenendo 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

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 6: 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 7: Informazioni sul knowledge graph del dataset LFM1b_artist_core5

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 8: 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 9: Informazioni sul knowledge graph del dataset LFM1b_artist_20U50I

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 10: 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 11: Informazioni sul knowledge graph del dataset LFM1b_artist_20U50L75strat

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 12: Informazioni sul dataset LFM1b_artist_20U50L50strat

Descrizione del knowledge graph

n_ent_head	12522
n_ent_tail	29509
n_rel	5
n_triple	38491

Tabella 13: Informazioni sul knowledge graph del dataset LFM1b_artist_20U50L50strat

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 14: Informazioni sul dataset LFM1b_artist_20U50L25strat

Descrizione del knowledge graph

n_ent_head	9444
n_ent_tail	23463
n_rel	5
n_triple	29822

Tabella 15: Informazioni sul knowledge graph del dataset LFM1b_artist_20U50L25strat

5.2 Statistiche dei dataset MovieLens10M

MovieLens10M è la versione a 10 milioni di interazioni del dataset MovieLens. In questo caso è stato aggiunto anche un KG con un hop pari a 1

Descrizione del dataset

Feature	Descrizione
n_users	69878
n_items	10677
n_inter	10000054
sparsity	0.9865966722939162
avg_inter_user	143.10732991785684

Tabella 16: Informazioni sul dataset ml-10m

Descrizione del knowledge graph

n_ent_head	179775
n_ent_tail	181868
n_rel	49
n_triple	1051385

Tabella 17: Informazioni sul knowledge graph del dataset ml-10m

5.2.1 Processing del dataset

Descrizione procedimento Il dataset originale risultava essere troppo grande per le risorse a nostra disposizione, dunque è stato opportunamente processato. In particolare 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 50000 utenti e 10000 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.

Dataset core5

Descrizione del dataset

Feature	Descrizione
n_users	69878
n_items	10196
n_inter	9998816
sparsity	0.9859661030476576
avg_inter_user	143.0896133260826

Tabella 18: Informazioni sul dataset ml-10m_core5

Descrizione del knowledge graph

n_ent_head	10132
n_ent_tail	169130
n_rel	31
n_triple	531224

Tabella 19: Informazioni sul knowledge graph del dataset ml-10m_core5

Dataset con 50.000 utenti e 10.000 items

Descrizione del dataset

Feature	Descrizione
n_users	50000
n_items	10000
n_inter	7053774
sparsity	0.985892452
avg_inter_user	141.07548

Tabella 20: Informazioni sul dataset ml-10m_50U10I

Descrizione del knowledge graph

n_ent_head	9937
n_ent_tail	167042
n_rel	31
n_triple	521125

Tabella 21: Informazioni sul knowledge graph del dataset ml-10m_50U10I

6 Configurazione

6.1 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

7 Emissioni

Qui di seguito vengono riportate le emissioni di CO2 per ogni esperimento effettuato. Si può notare come il numero dei modelli è stato ridotto. Per questioni di risorse e tempo si sono scelti modelli rappresentativi di determinate "classi" di modelli (es. Generali, Deep Learning, Knowledge-aware, etc.)

7.1 LFM-1b_artist

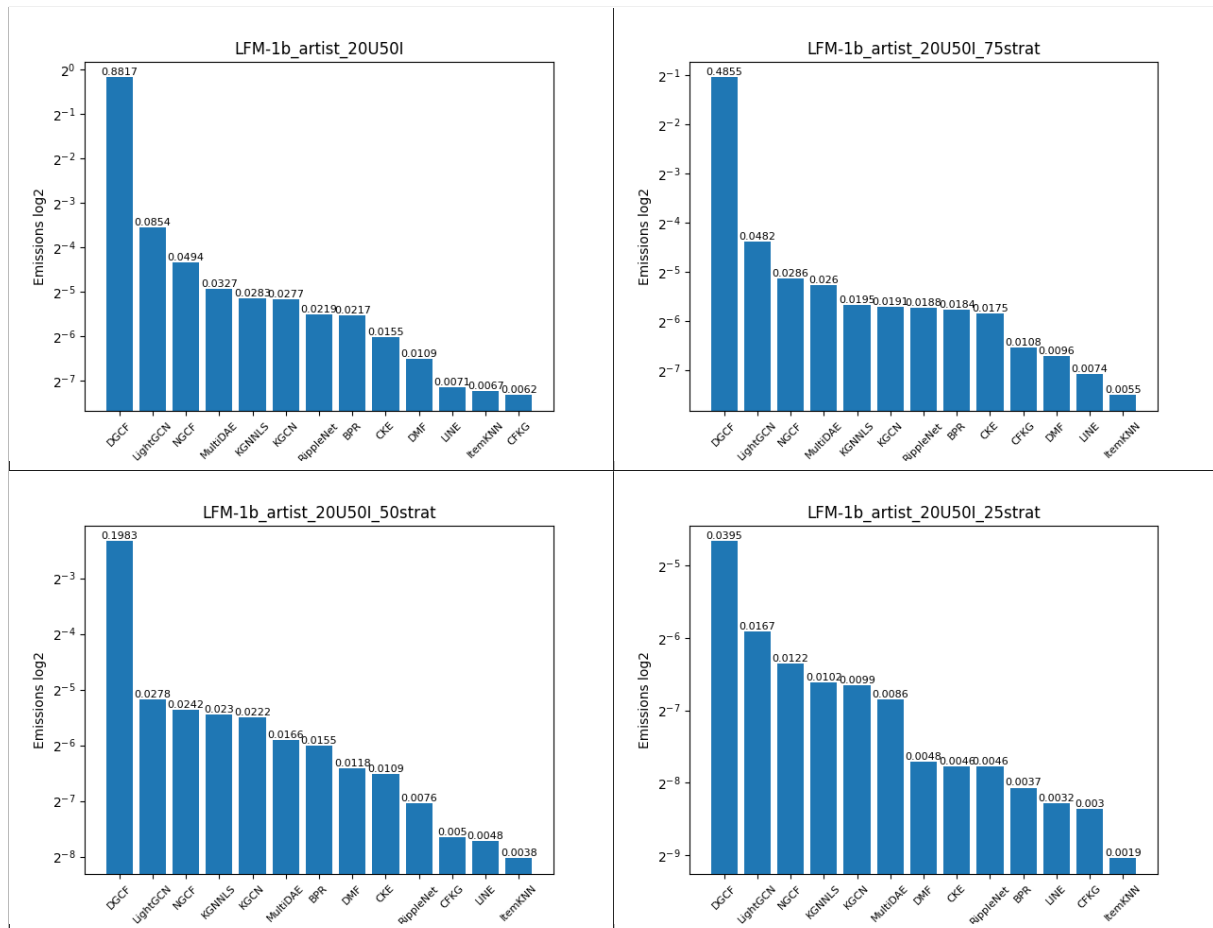


Tabella 22: Emissioni di CO2 per i vari dataset LFM-1b

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

7.2 ml-10m

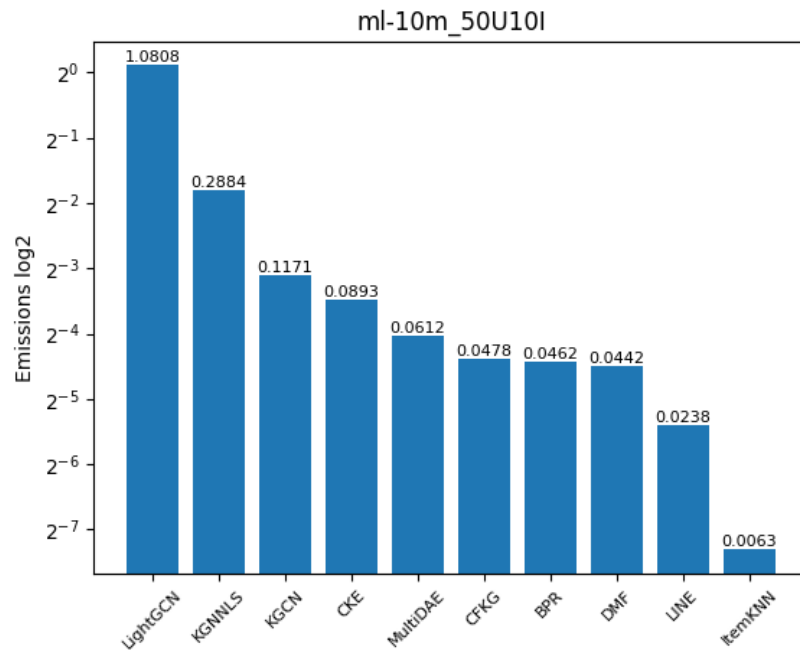


Figura 6: Emissioni di CO2 per il dataset ml-10m

Per quanto riguarda il dataset ml-10m, alcuni modelli non sono riusciti a completare l'addestramento per mancanza di risorse. Tuttavia è possibile notare come molti modelli rispetto al dataset LFM-1b_artist_20U50I abbiano emissioni di CO2 molto più alte (a volte anche 10 volte tanto) con eccezione di ItemKNN che si mantiene a un valore simile a LFM-1b_artist_20U50I. Ciò sicuramente è dovuto al fatto che il numero delle interazioni presenti in questo dataset sono circa 10 volte quelle del dataset LFM-1b_artist_20U50I. E' possibile comunque notare come i modelli che le emissioni dei modelli seguono circa lo stesso ordine dei dataset precedenti, confermando dunque che molti modelli emettono più CO2 rispetto ad altri anche con dataset molto diversi.

8 Trade-off

8.1 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.

8.2 LFM-1b_artist_20U50I

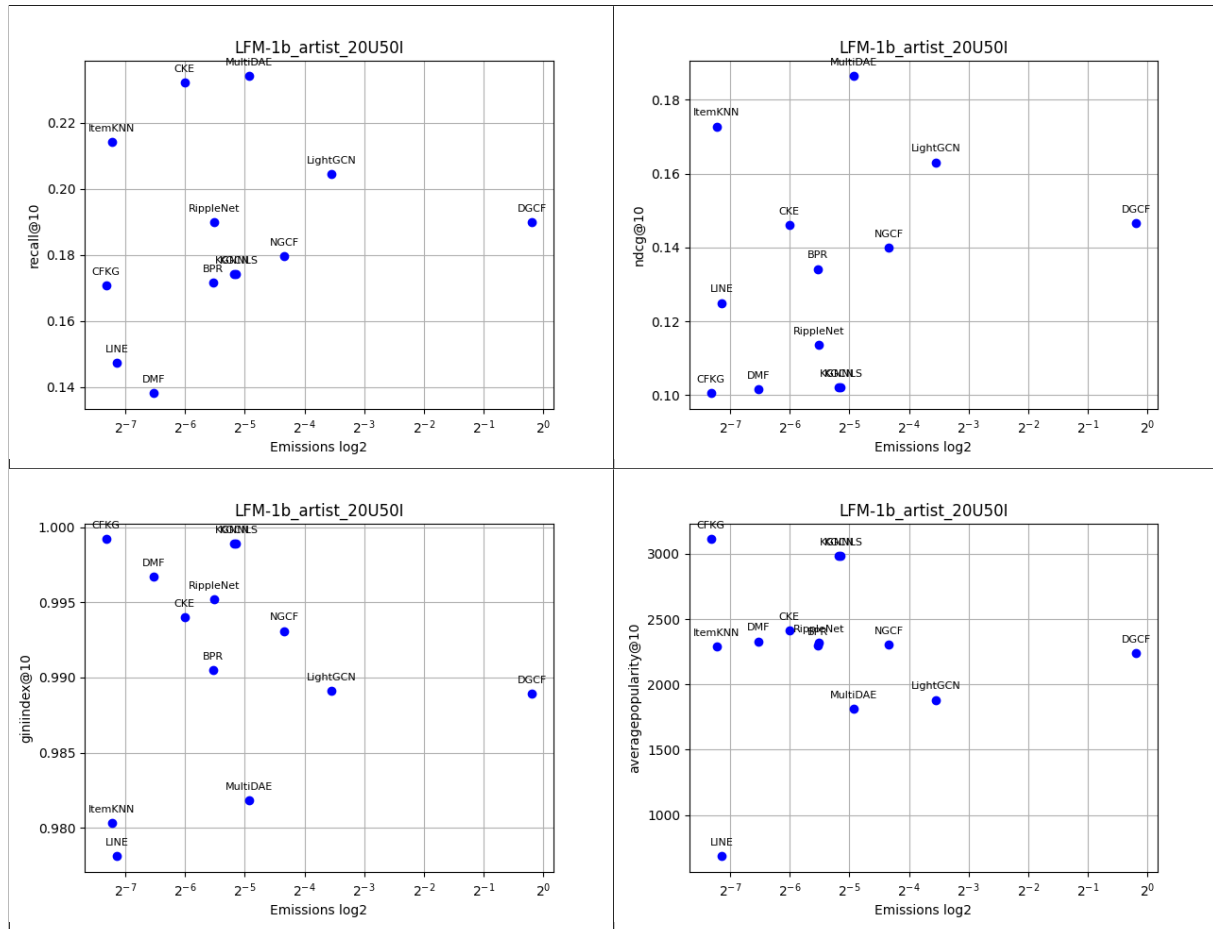


Tabella 23: Trade-off con il dataset LFM-1b_artist_20U50I

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.

8.3 LFM-1b_artist_20U50I_75strat

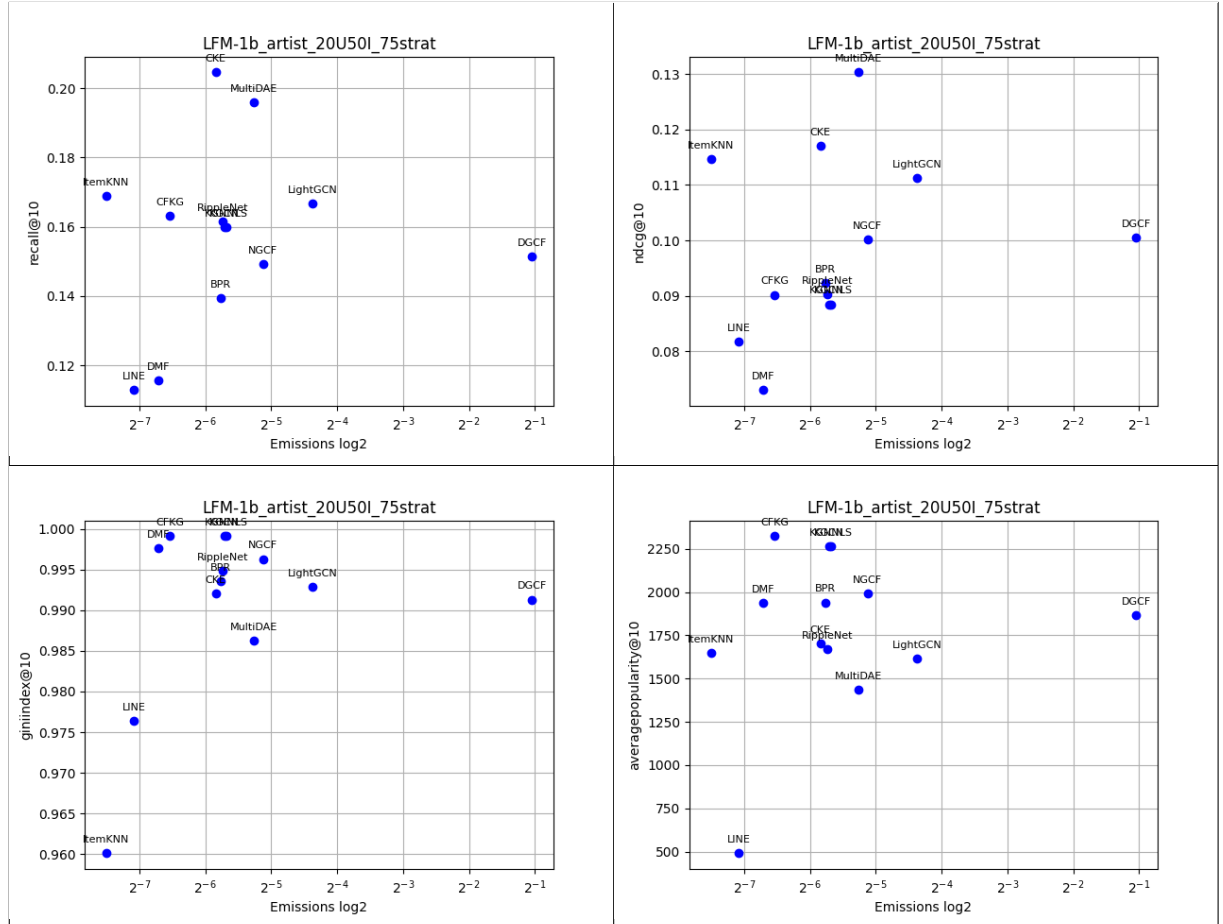


Tabella 24: Trade-off con il dataset LFM-1b_artist_20U50I

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.

8.4 LFM-1b_artist_20U50I_50strat

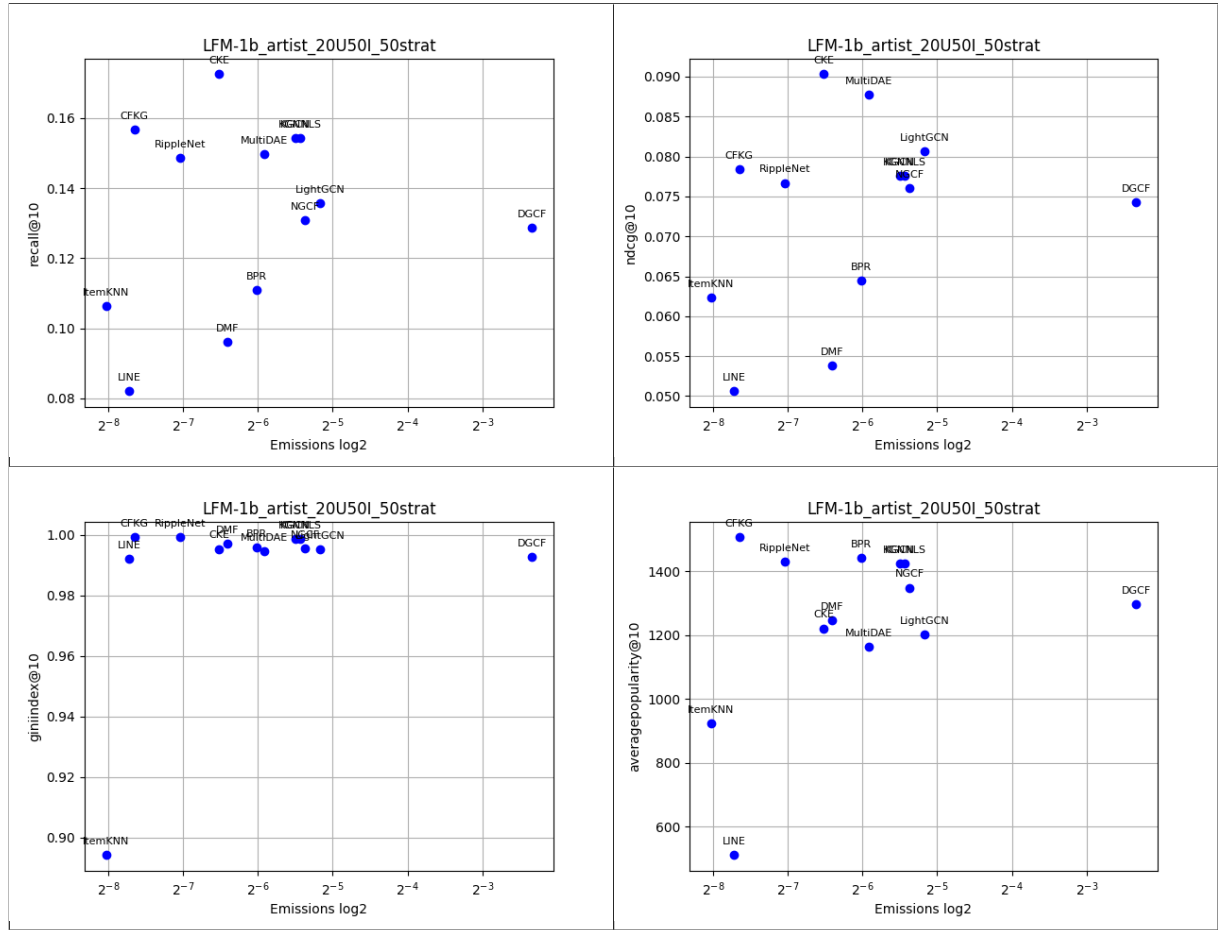


Tabella 25: Trade-off con il dataset LFM-1b_artist_20U50I

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.

8.5 LFM-1b_artist_20U50I_25strat

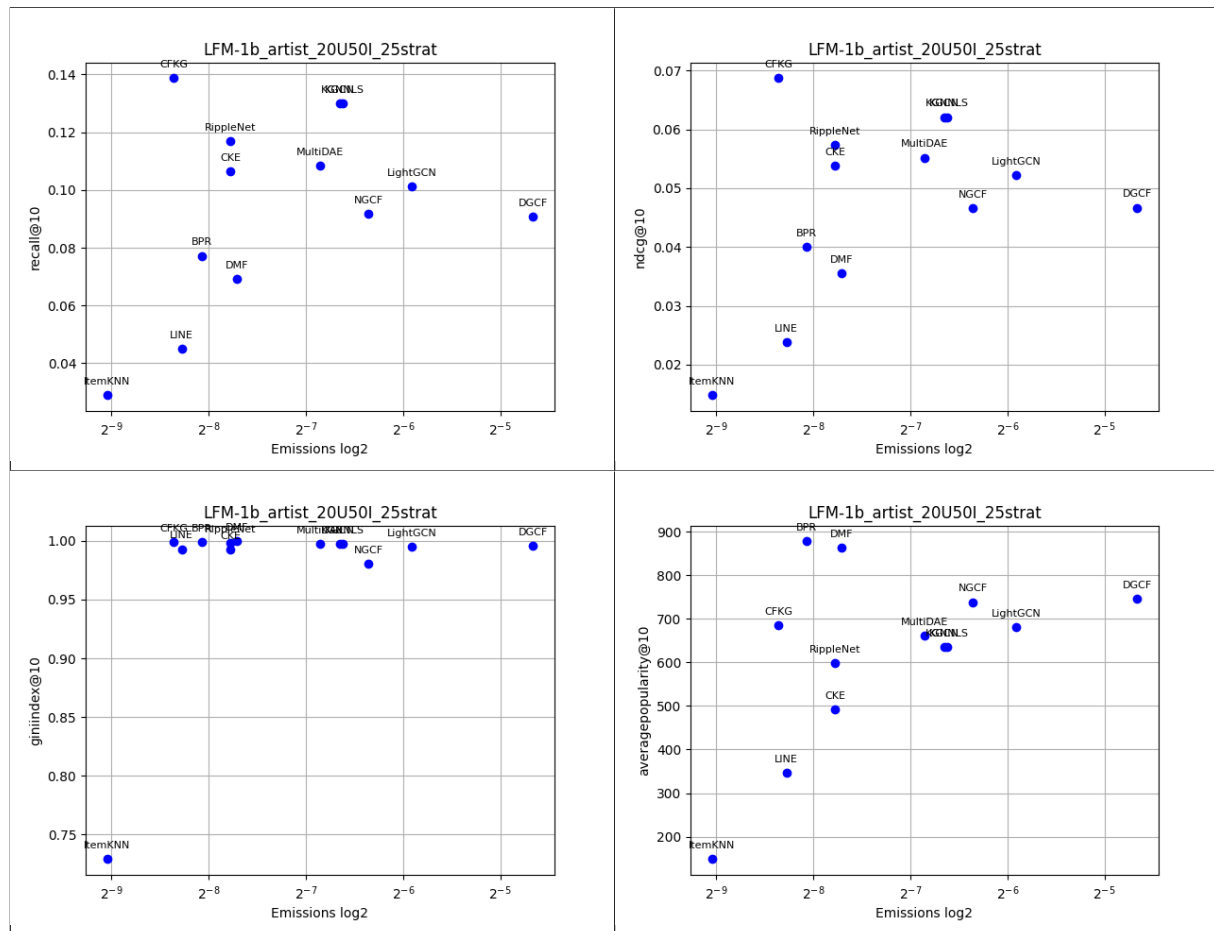


Tabella 26: Trade-off con il dataset LFM-1b_artist_20U50I

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 CKF 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.

8.6 ml-10m_50U10I

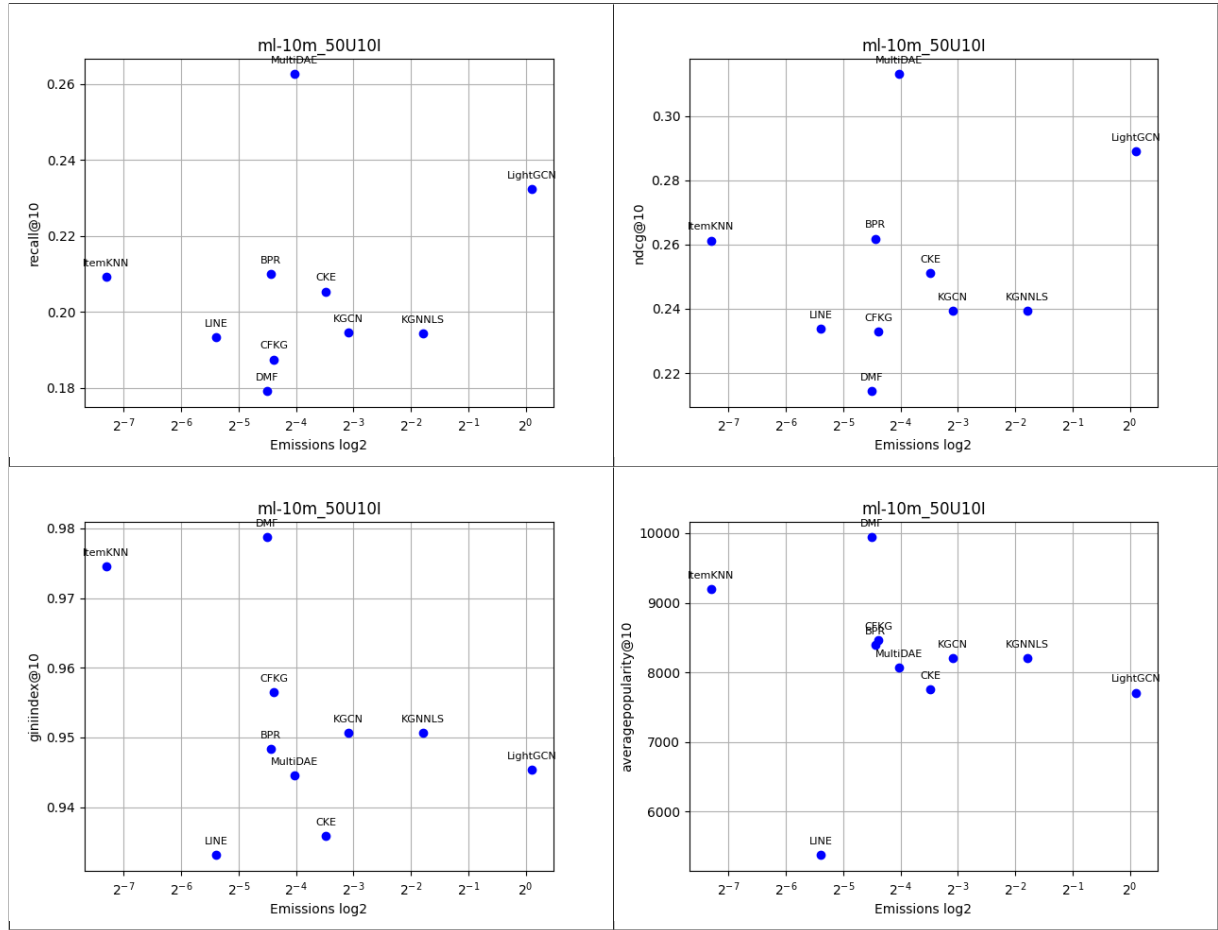


Tabella 27: Trade-off con il dataset ml-10m_50U10I

LightGCN è l'algoritmo che emette di più, osservando i valori da 5 a circa 180 volte rispetto agli altri modelli. Emissioni così alte non sono però giustificate. Nelle metriche di recall e ndcg LightGCN è il secondo migliore, ma la differenza rispetto al primo è molto bassa e quindi non giustifica emissioni così alte. ItemKNN si conferma uno dei migliori algoritmi per queste due metriche in quanto emette meno ed è uno dei più performanti. Per quanto riguarda le metriche di Gini Index e Average Popularity possiamo notare che LightGCN è uno di migliori, ma anche in questo caso la differenza rispetto ad altri modelli non giustifica emissioni così alte. ItemKNN si conferma con il secondo peggior modello come punteggi. Line risulta essere il migliore in quanto è il secondo per basse emissioni ed ottiene i punteggi migliori.

8.7 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

9 Regressore

9.1 Dataset originale

9.1.1 Dataset completo

Il nuovo dataset presenta i seguenti valori:

- **n_users**: [19841, 22155, 23679, 50000, 6040]
- **n_items**: [42457, 24878, 33653, 38932, 54458, 4414, 10000 3706]
- **n_inter**: [900212, 218457, 440620, 667850, 1465871, 1048575, 7053774, 1000209]
- **sparsity**: [0.99893136, 0.99955742, 0.9993401, 0.99913541, 0.99878504, 0.98996762, 0.98589245, 0.95531637]
- **kg_entities**: [50665, 32907, 42031, 47308, 26315, 0, 175646, 79347]
- **kg_relations**: [5, 16, 0, 31, 49]
- **kg_triples**: [46827, 29822, 38491, 43559, 96476, 0, 521125, 385923]
- **kg_items**: [816701, 11446, 0, 10601, 3655]
- **cpu_cores**: [12, 4]
- **ram_size**: [64, 16, 27.40581512]
- **is_gpu**: [1, 0]

Si può dunque notare un miglioramento per quanto riguarda la varietà dei valori rispetto al passato

I nuovi dati hanno portato alla seguente distribuzione dei dati (qui di seguito visualizzata):

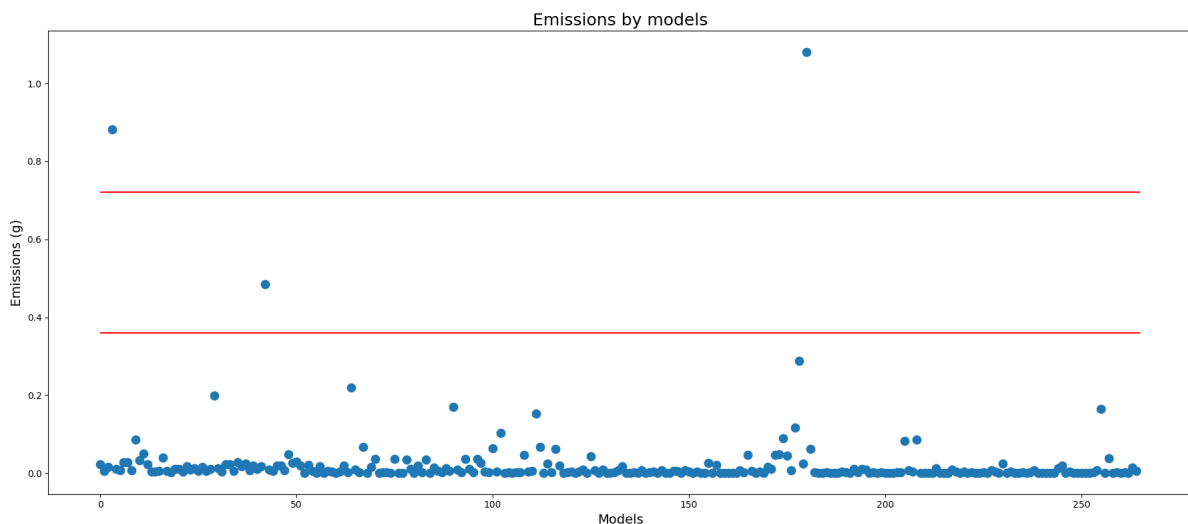


Figura 7: Distribuzione delle emissioni nel dataset completo

Come è possibile notare però, i nuovi esperimenti hanno portato a un'ulteriore sbilanciamento nel dataset, in quanto tutti gli esperimenti con DGCF e altri modelli sveltano sui risultati degli altri modelli in emissioni. Questo si riflette nei risultati ottenuti dai modelli di regressione. Per

quanto riguarda i modelli, sono stati eseguiti addestramenti con diversi split tra training e test set.

Split 50/50

Regressor	MAE	RMSE	MSLE
SVR	0.1053920	0.0207610	0.0131238
Decision Tree	0.0369020	0.0149577	0.0079782
Random Forest	0.0338687	0.0151556	0.0077735
AdaBoost	0.0397532	0.0148615	0.0079069

Tabella 28: Risultati ottenuti con il nuovo dataset split 50/50

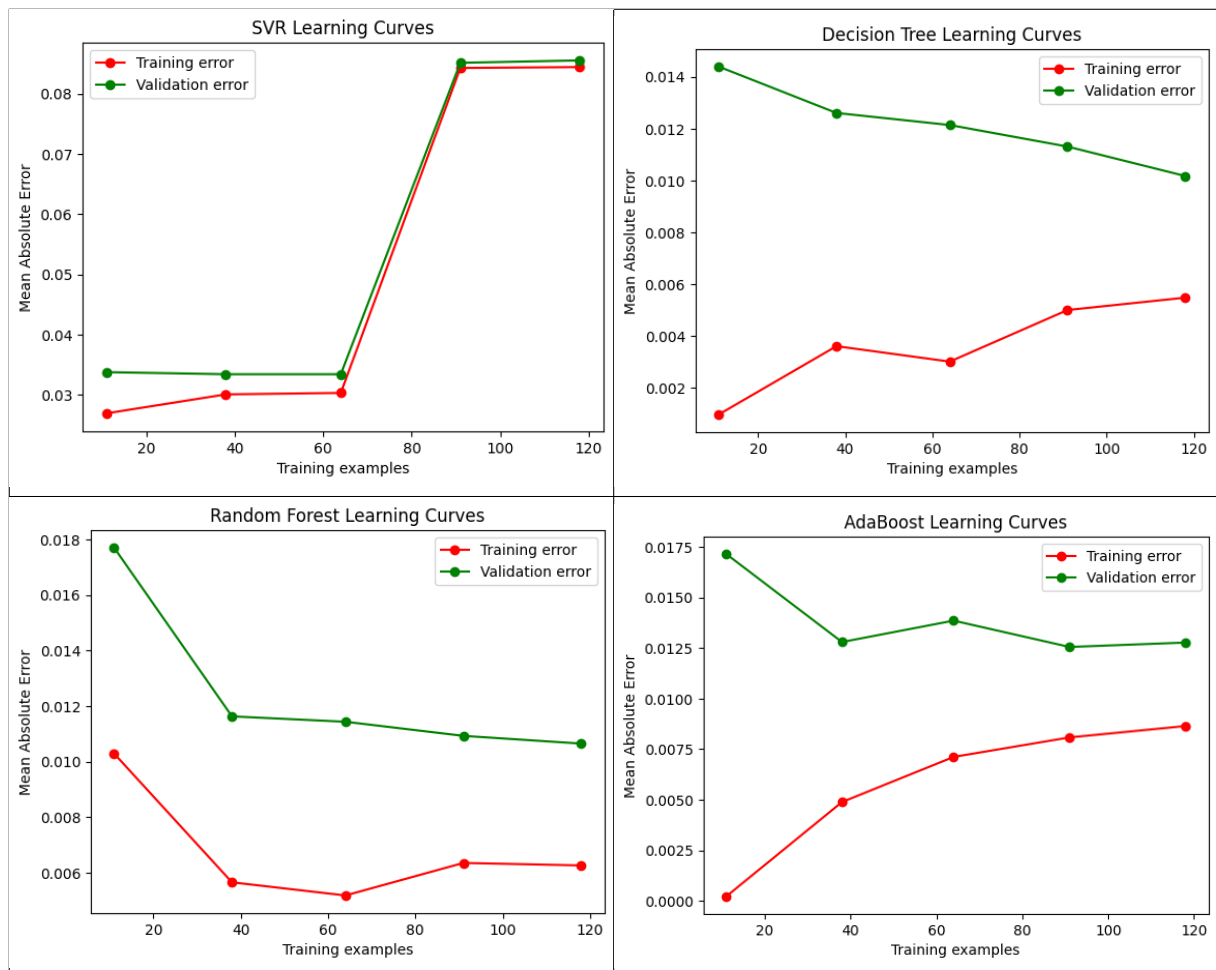


Tabella 29: Learning Curves con il nuovo dataset split 50/50

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, entrambe le curve crescono all'aumentare delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor. La learning curve mostra che

la curva di training aumenta all'aumentare del numero di istanze mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curves di train e di validation set diminuiscono all'aumentare del numero di istanze. Quella di training tra 60 in poi aumenta ma intorno a 100 si stabilizza. Il modello sembra cominciare a generalizzare.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Decision Tree Regressor ma migliori rispetto al SVR. La learning curve mostrano che il modello non è riuscito a generalizzare bene in quanto la curva di validation aumenta e diminuisce all'aumentare delle istanze, mentre quella di training aumenta.

Conclusioni Random Forest sembra essere il migliore sia dal punto di vista delle metriche che delle curve. Segue dietro il Decision Tree. AdaBoost e SVR non sono riusciti a generalizzare bene.

Split 60/40

Regressor	MAE	RMSE	MSLE
SVR	0.1082652	0.0237656	0.0144029
Decision Tree	0.0439202	0.0183129	0.0096602
Random Forest	0.0405412	0.0188566	0.0096752
AdaBoost	0.0519845	0.0186018	0.0099135

Tabella 30: Risultati ottenuti con il nuovo dataset split 60/40

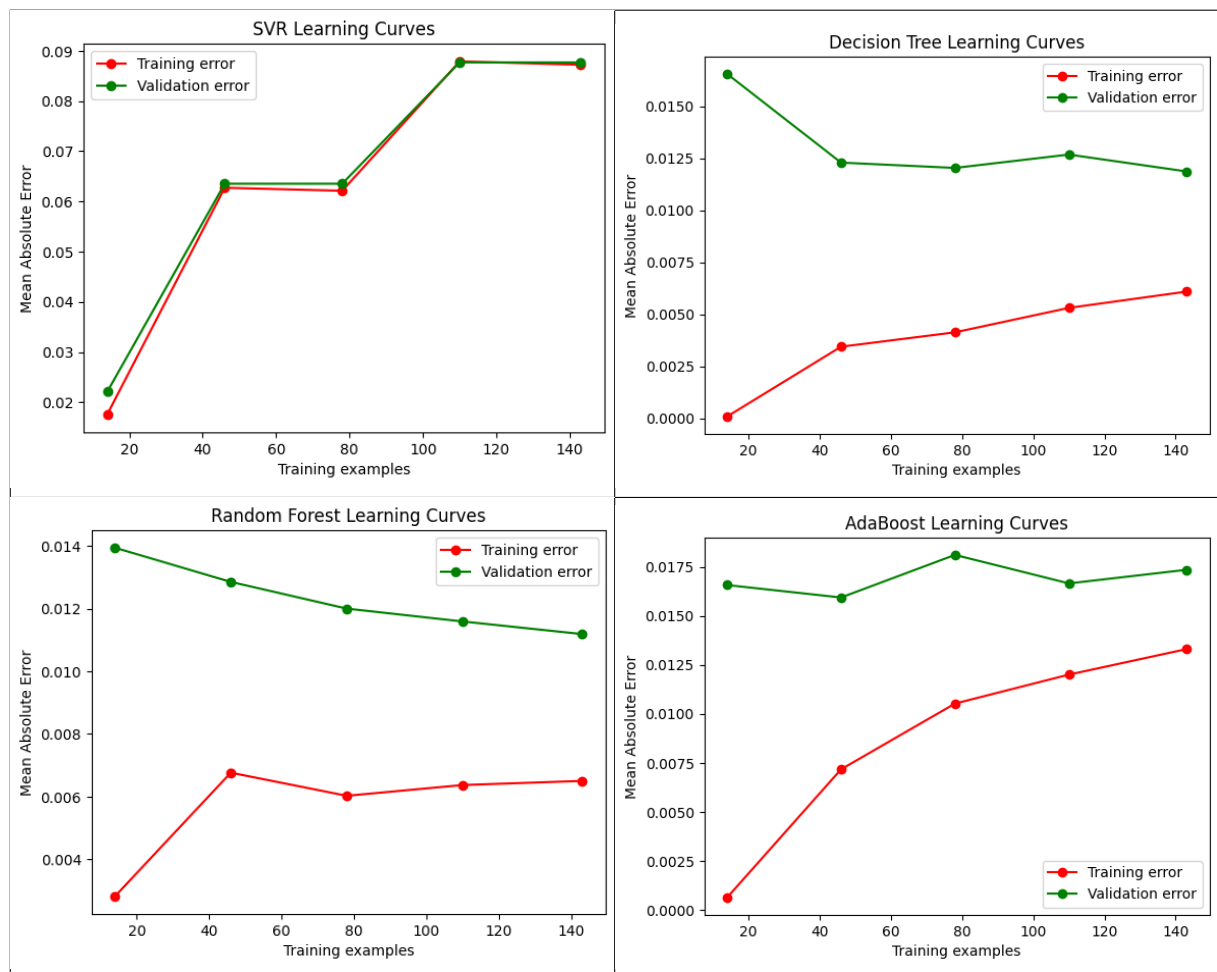


Tabella 31: Learning Curves con il nuovo dataset split 60/40

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, entrambe le curve crescono all'aumentare delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor. La learning curve mostra che la curva di training aumenta all'aumentare del numero di istanze mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curves di training diminuisce all'aumentare del numero di istanze. Quella di training intorno a 80 comincia ad aumentare leggermente e a stabilizzarsi.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Decision Tree Regressor ma migliori rispetto al SVR. La learning curve mostrano che il modello non è riuscito a generalizzare bene in quanto la curva di validation aumenta e diminuisce all'aumentare delle istanze, mentre quella di training aumenta.

Conclusioni Random Forest sembra essere il migliore sia dal punto di vista delle metriche che delle curve. Segue dietro il Decision Tree. AdaBoost e SVR non sono riusciti a generalizzare bene. In generale i modelli peggiorano rispetto allo split 50/50.

Split 70/30

Regressor	MAE	RMSE	MSLE
SVR	0.1103392	0.0268736	0.0154947
Decision Tree	0.0419923	0.0139534	0.0067713
Random Forest	0.0410102	0.0179366	0.0081916
AdaBoost	0.0486434	0.0160941	0.0074143

Tabella 32: Risultati ottenuti con il nuovo dataset split 70/30

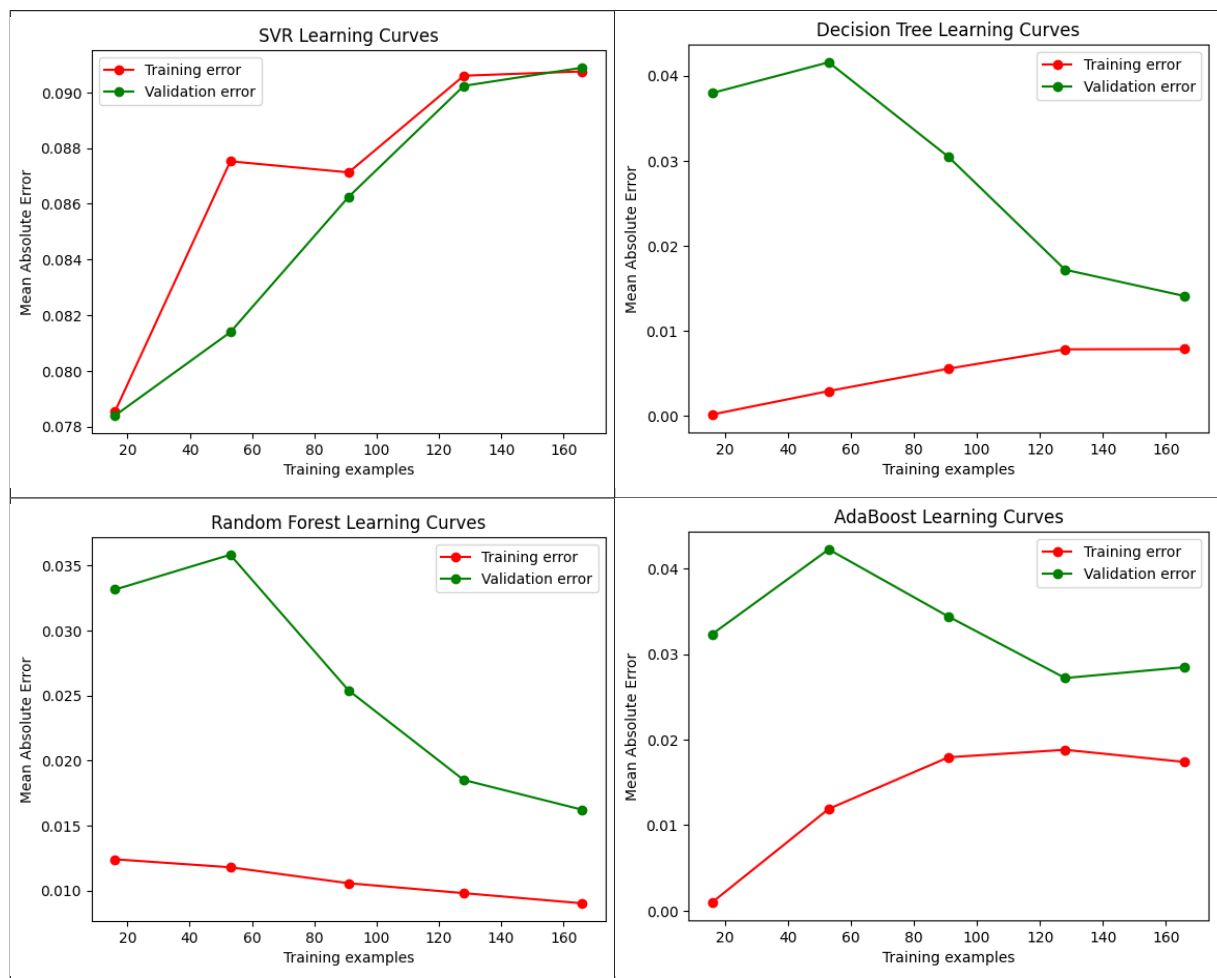


Tabella 33: Learning Curves con il nuovo dataset split 70/30

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, l'errore aumenta all'aumentare delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche di RMSE e MSLE, il Decision Tree Regressor ha ottenuto i migliori risultati. La learning curve mostra una curva di training che aumenta all'aumentare del numero di istanze, mentre la curva di validation diminuisce. Le

due curve si avvicinano, ma non si sovrappongono. Sembra esserci un leggero miglioramento rispetto al passato.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il secondo miglior modello, tranne per MAE dove è il migliore. La learning curves di train e di validation set diminuiscono entrambe all'aumentare del numero di istanze, ma non si sovrappongono. Anche qui rispetto al passato si nota un leggero miglioramento.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Random Forest Regressor e al Decision Tree Regressor ma migliori rispetto al SVR. La learning curve mostrano che il modello non è riuscito a generalizzare bene ma si è comportato meglio rispetto al passato.

Conclusioni Vi è un generale miglioramento rispetto allo split 60/40. Random Forest e Decision Tree sono i migliori modelli. AdaBoost e SVR non sono riusciti a generalizzare bene.
Split 80/20

Regressor	MAE	RMSE	MSLE
SVR	0.1211949	0.0364312	0.0196837
Decision Tree	0.0511053	0.0211890	0.0095072
Random Forest	0.0538403	0.0262660	0.0118334
AdaBoost	0.0617220	0.0230938	0.0104140

Tabella 34: Risultati ottenuti con il nuovo dataset split 80/20

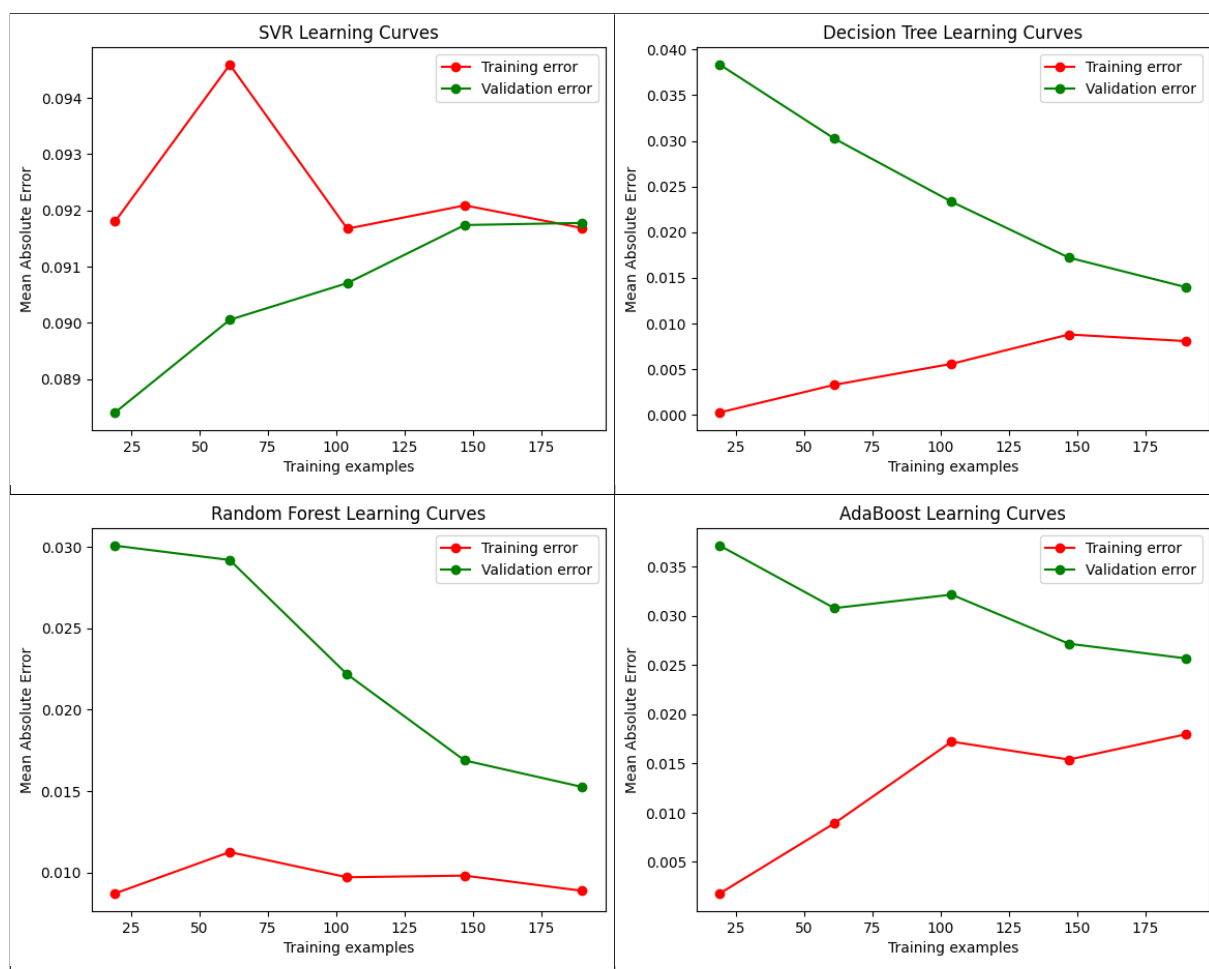


Tabella 35: Learning Curves con il nuovo dataset split 80/20

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve di validation aumenta all'aumentare delle istanze, mentre quella di training è più instabile. Il modello non è riuscito a generalizzare bene.

Decision Tree Regressor. Dal punto di vista delle metriche il Decision Tree Regressor ha ottenuto i migliori risultati. La learning curve mostra una curva di training che aumenta all'aumentare del numero di istanze cominciando una lenta discesa verso le 150 istanze, mentre la curva di validation diminuisce. Le due curve si avvicinano, Sembra che il modello cominci a generalizzare.

Random Forest Regressor. Random Forest Regressor è secondo migliore in MAE, terzo in RMSE e MSLE. La learning curves di train e di validation set diminuiscono entrambe all'aumentare del numero di istanze, ma non si sovrappongono. Comportamento analogo al passato.

AdaBoost Regressor. AdaBoost è il secondo migliore in RMSE e MSLE, terzo in MAE. La learning curve mostrano che il modello non è riuscito a generalizzare bene.

Conclusioni Decision Tree è il miglior modello. Tuttavia rispetto allo split precedente vi è un generale peggioramento delle performance.

Split 90/10

Regressor	MAE	RMSE	MSLE
SVR	0.1193075	0.0308946	0.0181817
Decision Tree	0.1011067	0.0860050	0.0377845
Random Forest	0.0857748	0.0539394	0.0275298
AdaBoost	0.1282122	0.0921047	0.0410386

Tabella 36: Risultati ottenuti con il nuovo dataset split 90/10

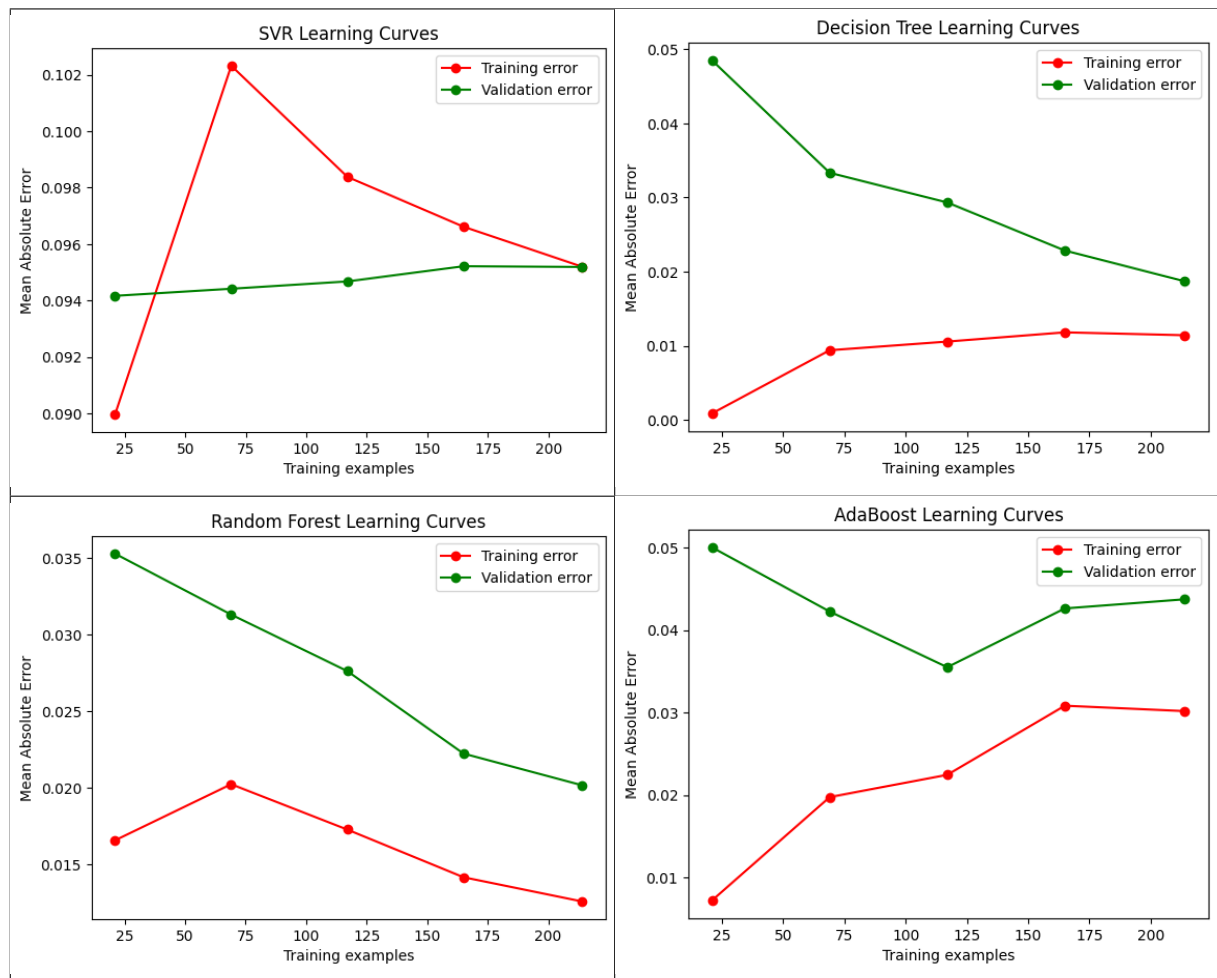


Tabella 37: Learning Curves con il nuovo dataset split 90/10

SVR. Dal punto di vista delle metriche il regressore SVR ha un pessimo risultato in MAE ma ha i migliori risultati in RMSE e MSLE. La learning curve di validation aumenta all'aumentare delle istanze, mentre quella di training è più instabile. Il modello non è riuscito a generalizzare bene.

Decision Tree Regressor. Dal punto di vista delle metriche, Decision è il secondo miglior modello. La learning curve mostra una curva di training che aumenta all'aumentare del numero di istanze, mentre la curva di validation diminuisce. Le due curve si avvicinano, ma non si sovrappongono. Comportamento analogo al passato.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curves di train e di validation set diminuiscono entrambe all'aumentare del numero di istanze, ma non si sovrappongono. Comportamento analogo al passato.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto i risultati peggiori. La learning curve mostrano che il modello non è riuscito a generalizzare bene in quanto entrambe le curve crescono all'aumentare delle istanze.

Conclusioni Il miglior modello si conferma Random Forest. Il secondo migliore sembra essere SVR. In generale vi è un drastico peggioramento delle performance rispetto allo split precedente.

Conclusioni generali In generale Random Forest sembra essere il miglior modello. Decision Tree è il secondo migliore. AdaBoost e SVR non sono riusciti a generalizzare bene. In particolare lo split dove le performance sono migliori osservando sia learning curve che metriche è il 70/30 a cui segue il 60/40.

9.1.2 Dataset Azure

Parallelamente si è deciso di procedere eliminando tutti i risultati non prodotti sulla macchina Azure in modo da avere un regressore specifico per tali esperimenti.

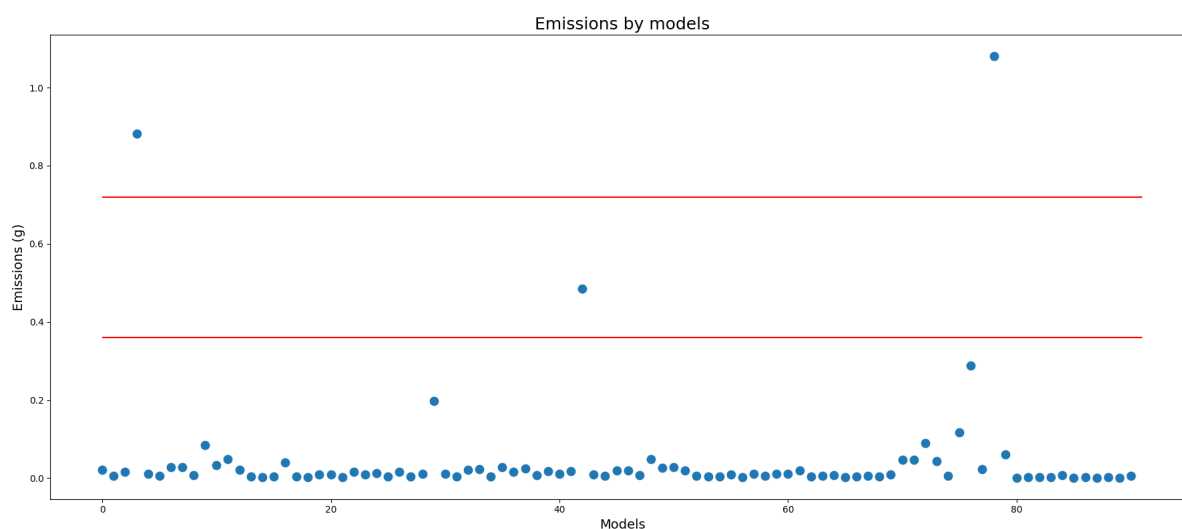


Figura 8: Distribuzione delle emissioni nel dataset Azure

Anche qui si è deciso di eseguire addestramenti con diversi split tra training e test set.

Split 50/50

Regressor	MAE	RMSE	MSLE
SVR	0.1043015	0.0213416	0.0133855
Decision Tree	0.0583675	0.0329582	0.0154626
Random Forest	0.0459593	0.0179844	0.0092711
AdaBoost	0.0615526	0.0331120	0.0155884

Tabella 38: Risultati ottenuti con il nuovo dataset Azure con split 50/50

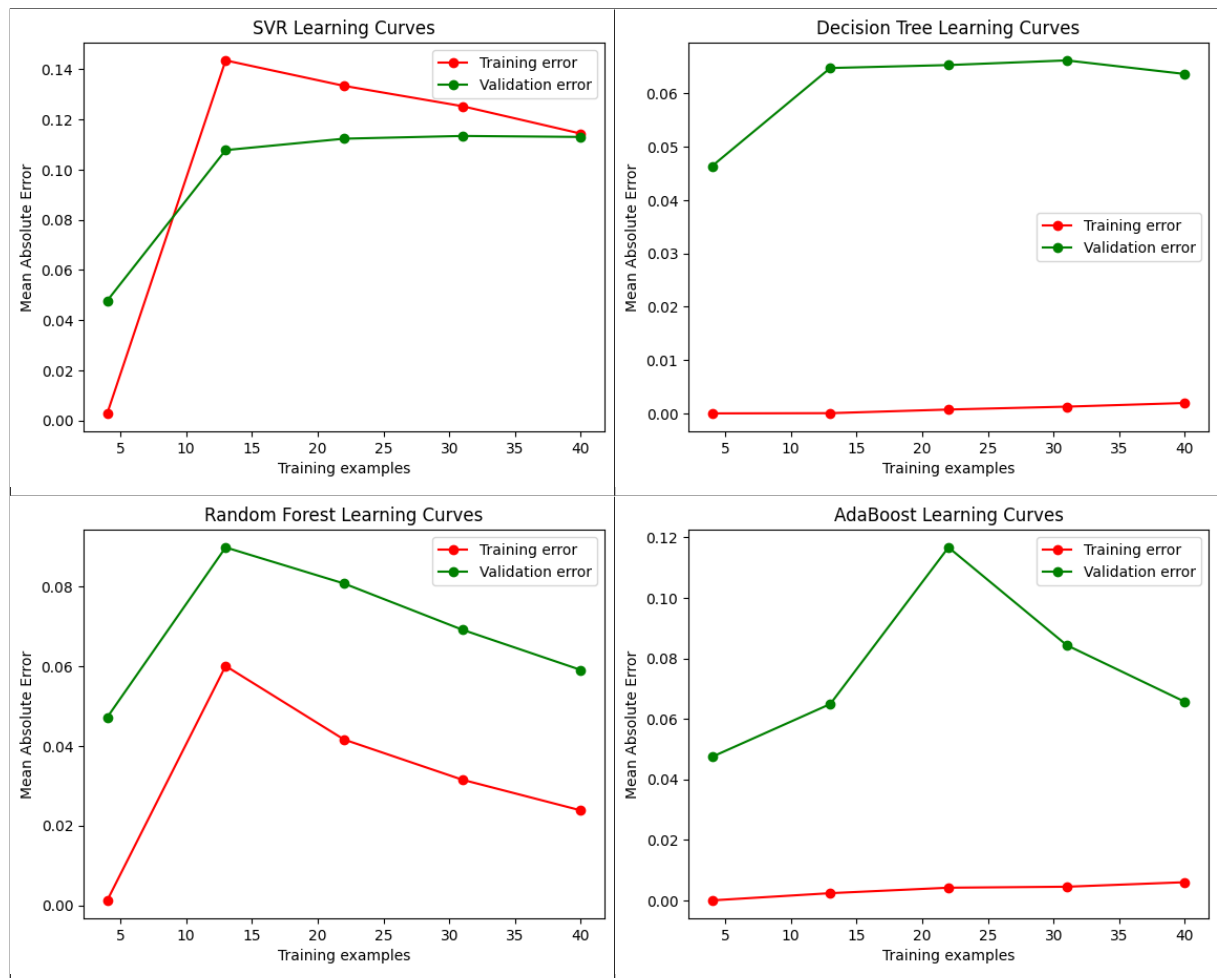


Tabella 39: Learning Curves con il nuovo dataset Azure con split 50/50

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE e i secondi migliori in RMSE e MSLE. La training curve inizialmente cresce velocemente per poi diminuire, mentre quella di validation cresce all'aumentare delle istanze. Le due arrivano a sovrapporsi ma non si incrociano. Il modello non sembra aver generalizzato bene.

Decision Tree Regressor. Dal punto di vista delle metriche il DecisionTree non si comporta bene. L'errore sul training è molto basso mentre sul validation è molto alto. Il modello non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curves di train e di validation hanno lo stesso andamento e dalle 10 istanze circa in poi diminuiscono all'aumentare del numero delle istanze. Il modello generalizza meglio degli altri.

AdaBoost Regressor. AdaBoost dal punto di vista delle metriche è il secondo peggior modello. La learning curve mostra che il modello non è riuscito a generalizzare bene in quanto molto irregolari.

Conclusioni RandomForest sembra essere il miglior modello. Dal punto di vista delle metriche segue Decision Tree anche se le curve non mostrano una buona generalizzazione.

Split 60/40

Regressor	MAE	RMSE	MSLE
SVR	0.1097324	0.0248018	0.0150739
Decision Tree	0.0556196	0.0342894	0.0152874
Random Forest	0.0498345	0.0213098	0.0107539
AdaBoost	0.0754194	0.0411506	0.0193612

Tabella 40: Risultati ottenuti con il nuovo dataset Azure con split 60/40

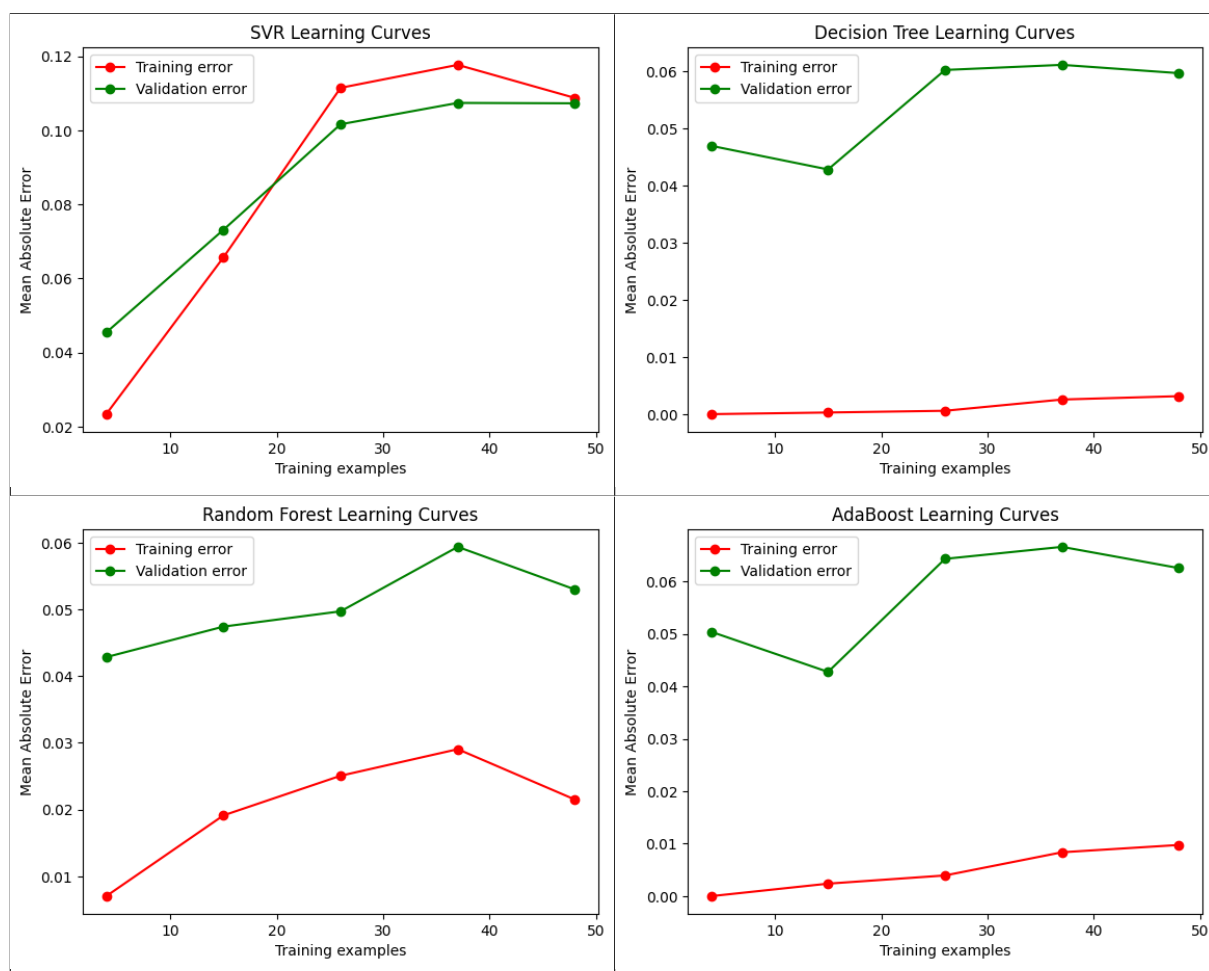


Tabella 41: Learning Curves con il nuovo dataset Azure con split 60/40

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE e i secondi migliori in RMSE e MSLE. La training curve inizialmente cresce velocemente per poi diminuire, mentre quella di validation cresce all'aumentare delle istanze. Le due arrivano a sovrapporsi ma non si incrociano. Il modello non sembra aver generalizzato bene.

Decision Tree Regressor. Dal punto di vista delle metriche il DecisionTree non si comporta bene. L'errore sul training è molto basso mentre sul validation è molto alto. Il modello non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curves di train e di validation hanno lo stesso andamento e dalle 35 istanze circa in poi diminuiscono all'aumentare del numero delle istanze. Il modello generalizza meglio degli altri.

AdaBoost Regressor. AdaBoost dal punto di vista delle metriche è il secondo peggior modello. La learning curve mostra che il modello non è riuscito a generalizzare bene in quanto molto irregolari.

Conclusioni RandomForest sembra essere il miglior modello. Dal punto di vista delle metriche segue Decision Tree anche se le curve non mostrano una buona generalizzazione. In generale si nota un peggioramento rispetto allo split precedente.

Split 70/30

Regressor	MAE	RMSE	MSLE
SVR	0.1165169	0.0301642	0.0175917
Decision Tree	0.0691471	0.0450743	0.0199880
Random Forest	0.0641987	0.0306766	0.0153200
AdaBoost	0.0910959	0.0542499	0.0254505

Tabella 42: Risultati ottenuti con il nuovo dataset Azure

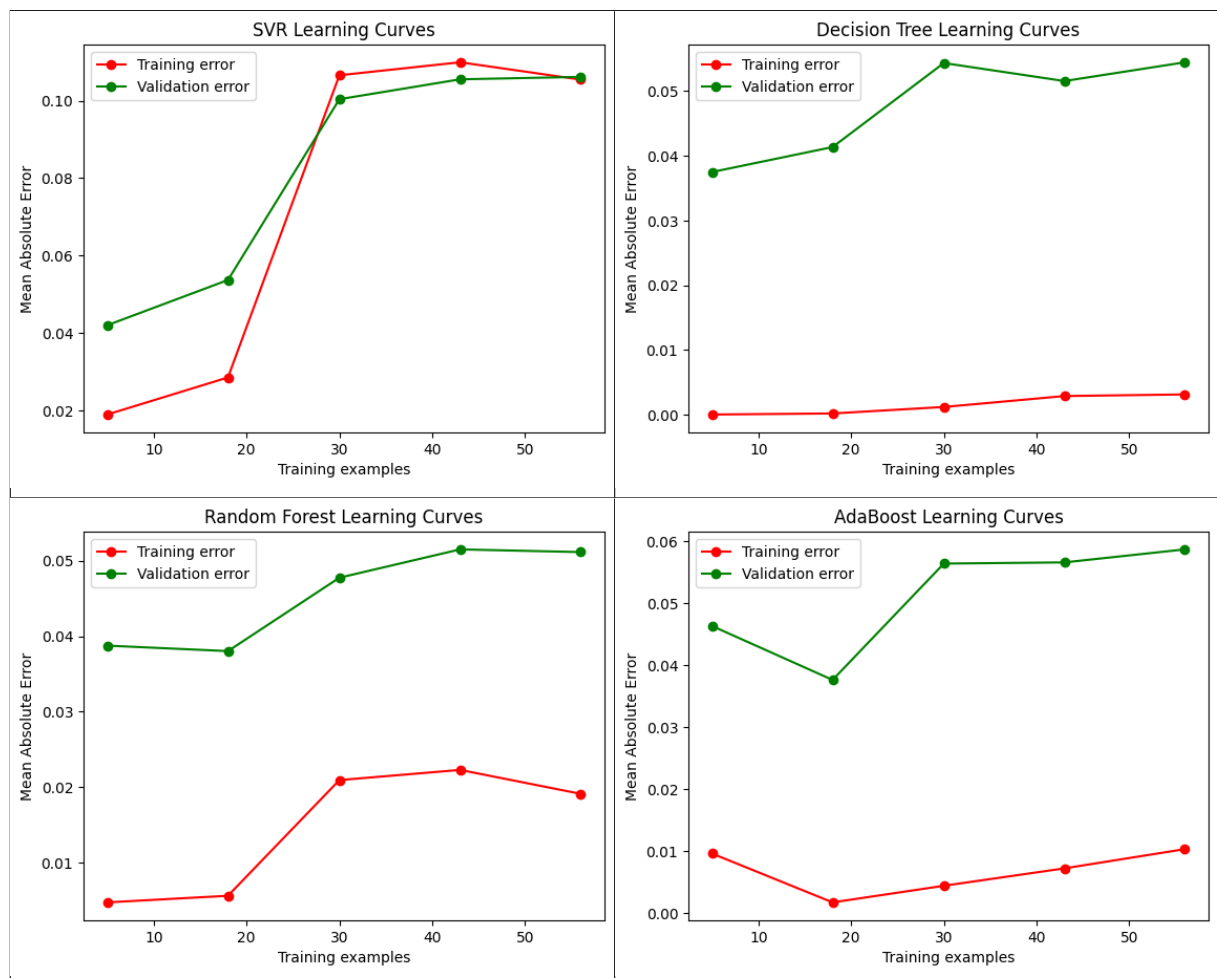


Tabella 43: Learning Curves con il nuovo dataset Azure

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti l'errore di training supera anche quello di validation.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor. La learning curve mostra che entrambe le curve di training e di validation aumentano all'aumentare del numero di istanze. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curves di train e di validation set crescono all'aumentare del numero delle istanze, quella di training infine comincia una lenta discesa. Il modello comunque non ha appreso bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Decision Tree Regressor ma migliori rispetto al SVR. La learning curve mostrano che il modello non è riuscito a generalizzare bene

Conclusioni RandomForest sembra essere il miglior modello. Decision Tree è il secondo migliore. AdaBoost e SVR non sono riusciti a generalizzare bene. In generale si nota un peggioramento rispetto allo split precedente.

Split 80/20

Regressor	MAE	RMSE	MSLE
SVR	0.1308274	0.0407490	0.0226068
Decision Tree	0.0470062	0.0118862	0.0057503
Random Forest	0.0557055	0.0218310	0.0098565
AdaBoost	0.0822472	0.0254612	0.0138452

Tabella 44: Risultati ottenuti con il nuovo dataset Azure con split 80/20

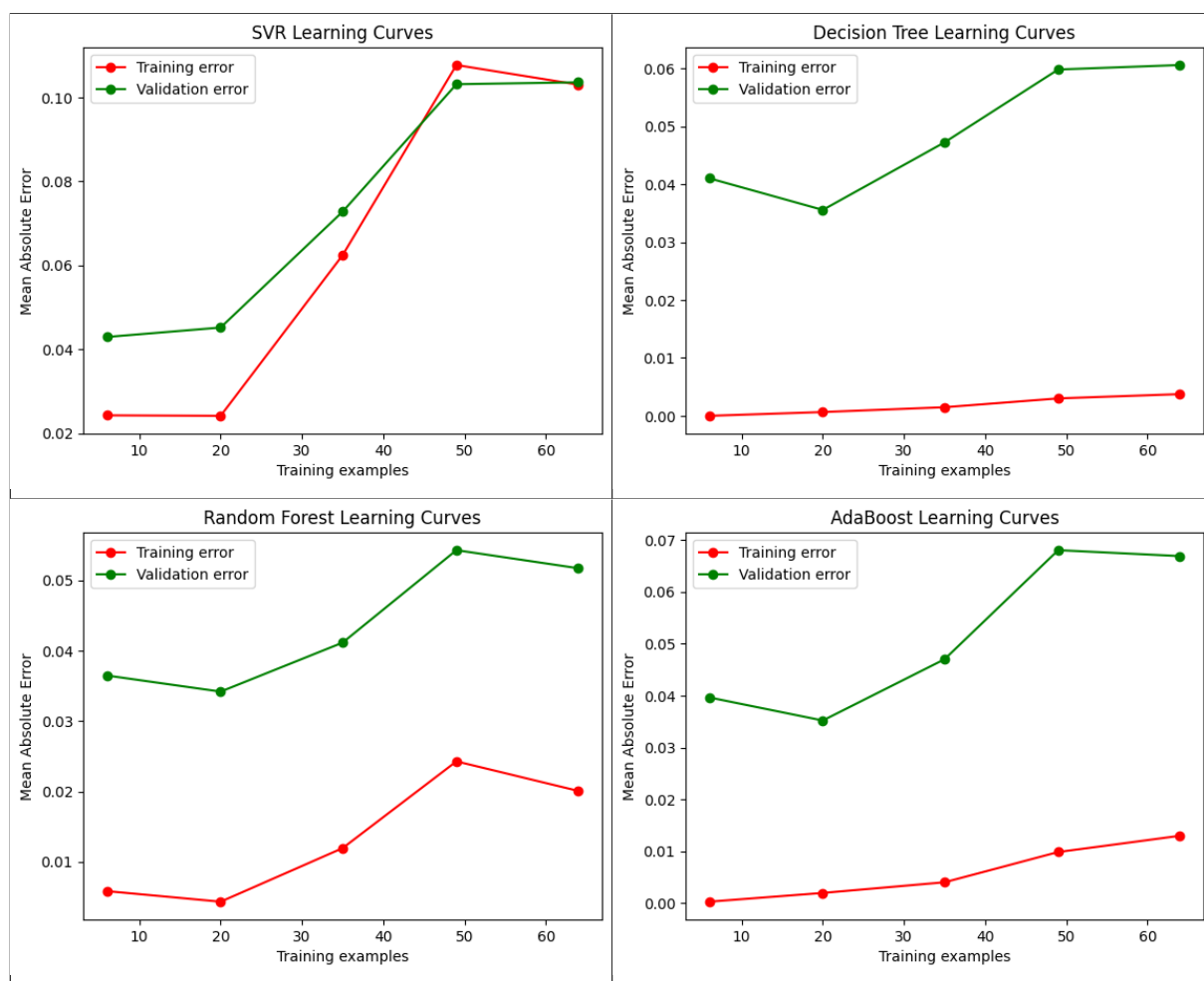


Tabella 45: Learning Curves con il nuovo dataset Azure con split 80/20

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti l'errore di training supera anche quello di validation.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i migliori risultati. La learning curve mostra che entrambe le curve di training e di validation aumentano all'aumentare del numero di istanze. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il secondo miglior modello. La learning curves di train e di validation set seguono lo stesso andamento, crescendo fino a 50 istanze e cominciando poi a diminuire. Il Comportamento è analogo al passato.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Random Forest Regressor ma migliori rispetto al SVR. La learning curve mostrano che il modello non è riuscito a generalizzare ben in quanto entrambe le curve crescono all'aumentare delle istanze.

Conclusioni DecisionTree è il miglior modello dal punto di vista delle metriche, il secondo migliore è Random Forest. Dal punto di vista delle curve quelle della Random Forest sono le migliori. Si nota un netto miglioramento rispetto allo split precedente.

Split 90/10

Regressor	MAE	RMSE	MSLE
SVR	0.0896299	0.0081883	0.0073278
Decision Tree	0.0220335	0.0025806	0.0020781
Random Forest	0.0231489	0.0023014	0.0018764
AdaBoost	0.0384317	0.0063016	0.0047659

Tabella 46: Risultati ottenuti con il nuovo dataset Azure con split 90/10

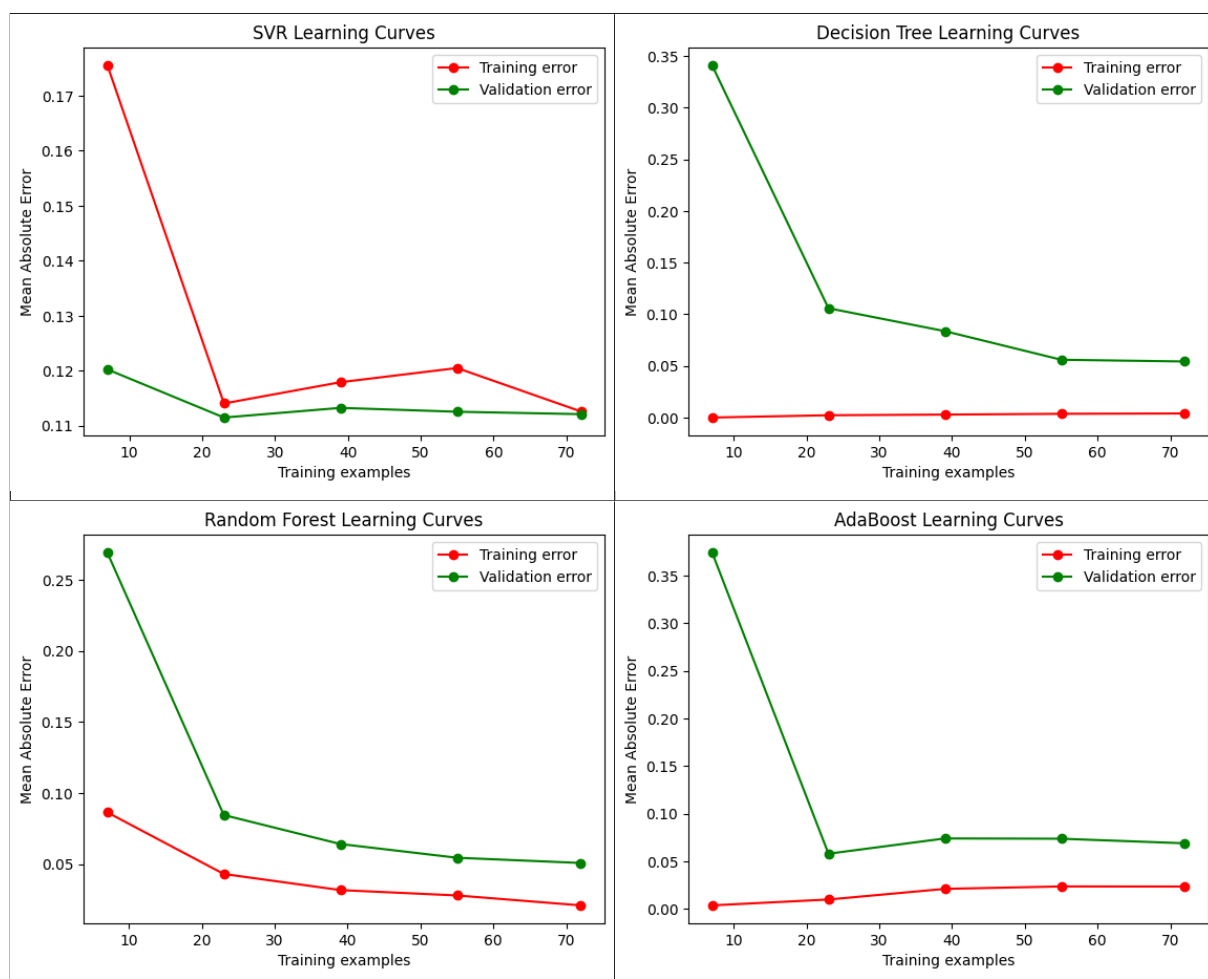


Tabella 47: Learning Curves con il nuovo dataset Azure con split 90/10

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati. Entrambe le learning curve diminuiscono all'aumentare delle istanze. Il modello sembra aver generalizzato da questo punto di vista, ma l'errore è ancora alto.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor, tranne in MAE dove è il migliore. La learning curve di validation diminuisce all'aumentare delle istanze, mentre quella di training aumenta. Il modello non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello, MAE escluso. Entrambe le learning curve diminuiscono all'aumentare delle istanze. Il modello sembra aver generalizzato bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Decision Tree Regressor e Random Forest Regressor ma migliori rispetto al SVR. La learning curve mostrano che il modello non è riuscito a generalizzare bene in quanto quella di validation diminuisce all'aumentare delle istanze, mentre quella di training aumenta.

Conclusioni Questo sembra essere il miglior split per il dataset Azure. Random Forest sembra essere il miglior modello. Decision Tree è il secondo migliore. AdaBoost e SVR non sono riusciti a generalizzare bene.

Conclusioni generali In generale Random Forest sembra essere il miglior modello. Decision Tree 'e il secondo migliore. AdaBoost e SVR non sono riusciti a generalizzare bene. In particolare lo split dove le performance sono migliori osservando sia learning curve che metriche 'e il 90/10 a cui segue il 60/40.

9.2 Dataset ridotto

Per provare a migliorare ancor di più le performance dei modelli di regressione, si è deciso di eliminare tutti gli outlier dal dataset(in questo caso le emissioni più alte). In particolare si sono eliminate 11 emissioni dal dataset completo e 8 dal dataset AZURE.

9.2.1 Dataset completo ridotto

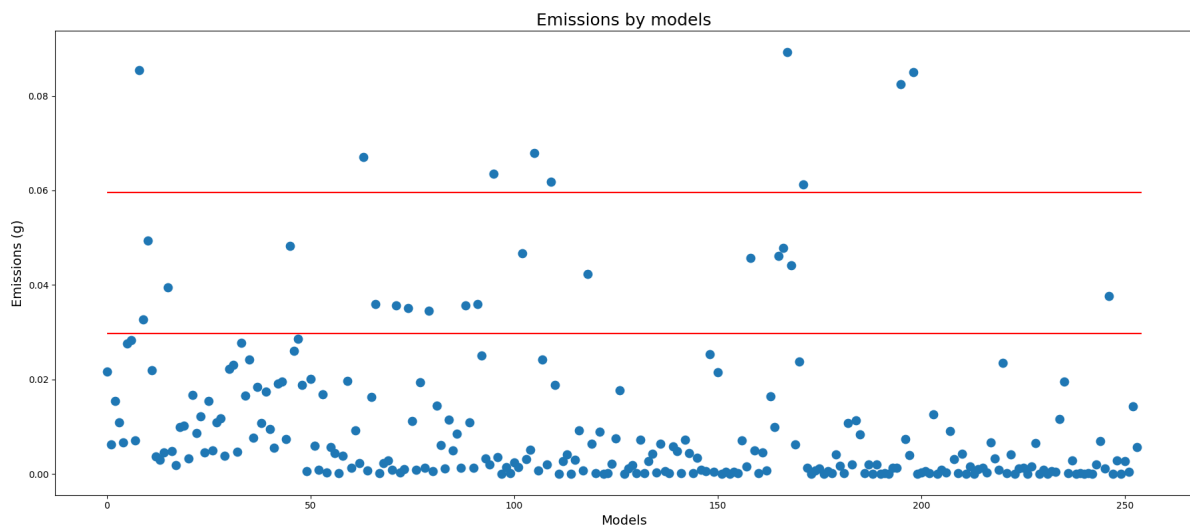


Figura 9: Distribuzione delle emissioni nel dataset completo ridotto

E' possibile notare un miglior bilanciamento delle emissioni. Per quanto riguarda i modelli, sono stati eseguiti addestramenti con diversi split tra training e test set

Split 50/50

Regressor	MAE	RMSE	MSLE
SVR	0.0357360	0.0014172	0.0013505
Decision Tree	0.0071074	0.0001741	0.0001635
Random Forest	0.0067896	0.0001391	0.0001312
AdaBoost	0.0080218	0.0001498	0.0001417

Tabella 48: Risultati regressore con dataset completo ridotto 50/50

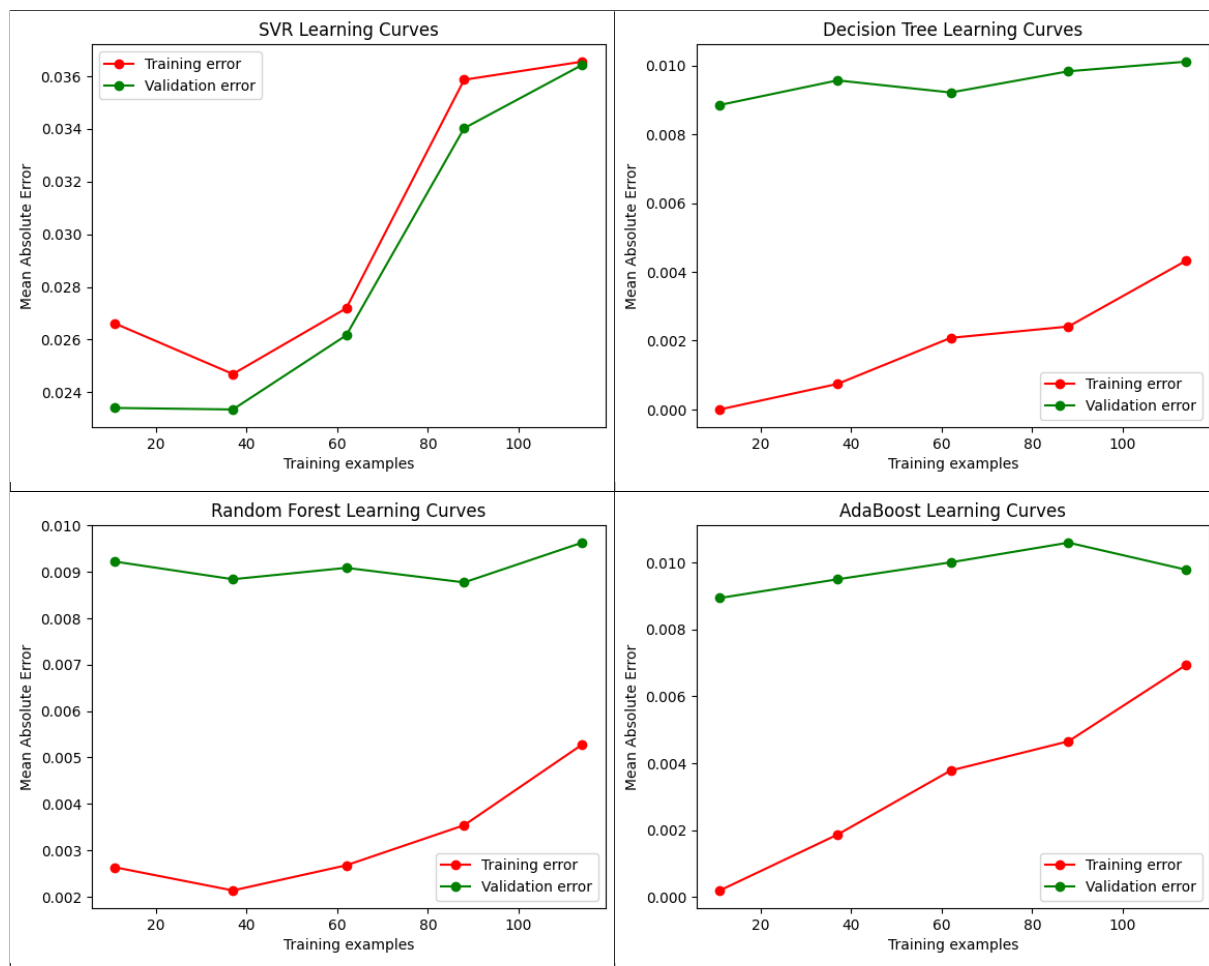


Tabella 49: Learning Curves con dataset completo ridotto 50/50

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor. La learning curve mostra che entrambe le curve di training e di validation aumentano all'aumentare del numero di istanze. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curve mostra che entrambe le curve di training e di validation aumentano all'aumentare del numero di istanze ma la distanza tra gli errori è minore rispetto al Decision Tree Regressor. Il modello dunque non ha generalizzato bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Decision Tree Regressor in MAE ma risultati molto simili al Random Forest Regressor. La learning curve mostrano che il modello non è riuscito a generalizzare bene.

Conclusioni Con lo split 50/50 il RandomForest sembrerebbe essere il migliore e l'AbaBoost sembra migliorare rispetto al passato, anche se ancora non riesce a generalizzare bene.

Split 60/40

Regressor	MAE	RMSE	MSLE
SVR	0.0351053	0.0013804	0.0013161
Decision Tree	0.0072612	0.0001487	0.0001409
Random Forest	0.0064122	0.0000973	0.0000930
AdaBoost	0.0088515	0.0001644	0.0001578

Tabella 50: Risultati regressore con dataset completo ridotto 60/40

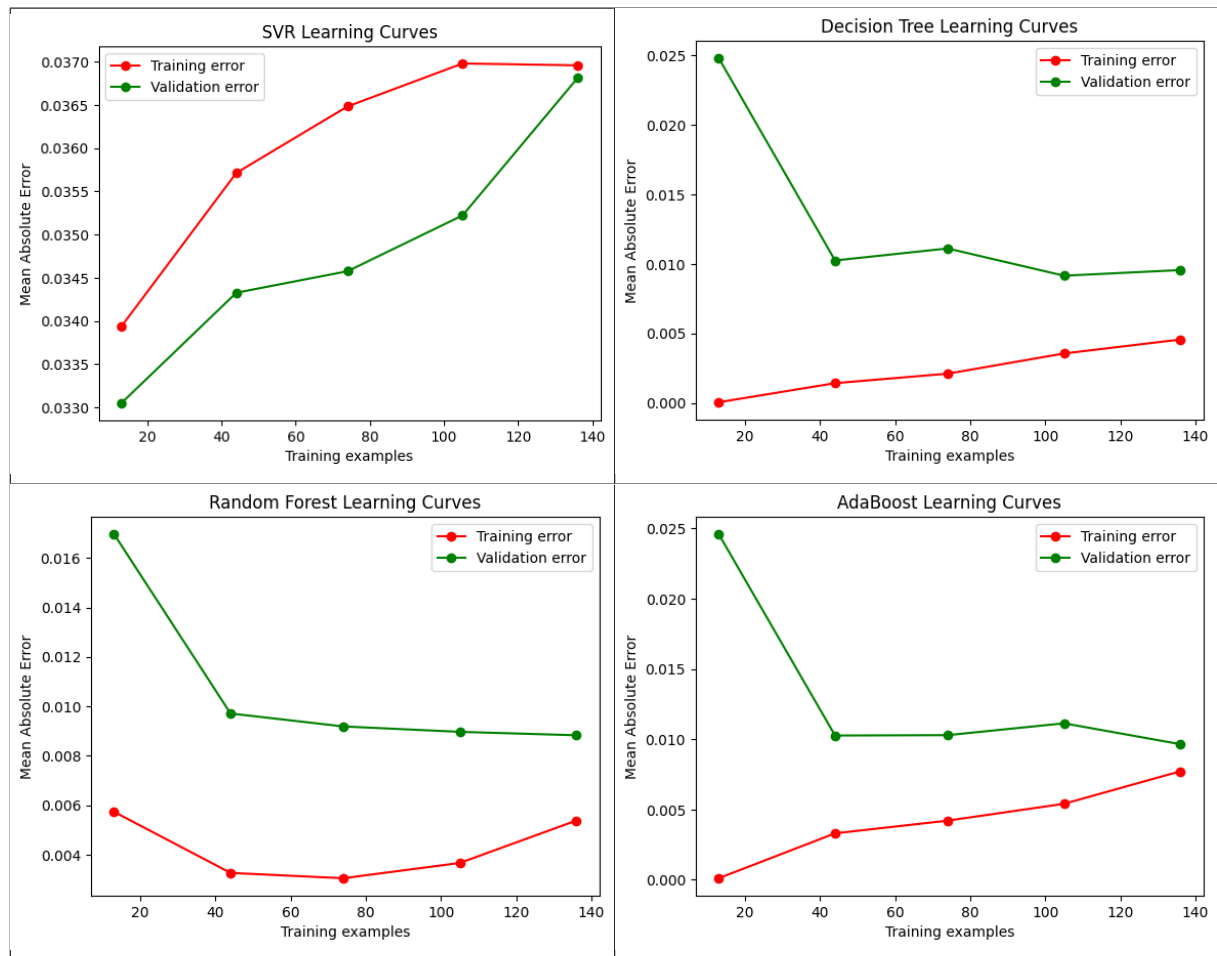


Tabella 51: Learning Curves con dataset completo ridotto 60/40

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor. La learning curve mostra una curva di training che aumenta all'aumentare del numero di istanze, mentre la curva di validation diminuisce. Le due curve si avvicinano, ma non si sovrappongono. Sembra esserci un leggero miglioramento rispetto al passato.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. Le curve sono molto simili a quelle del Decision Tree Regressor, ma sembra esserci un leggero miglioramento in quanto più vicine tra loro.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati simili al Decision Tree Regressor nelle metriche. La curva di validation diminuisce all'aumentare delle istanze, quella di training aumenta. Sono molto vicine tra loro tuttavia l'errore è più alto rispetto agli altri modelli (tranne SVR).

Conclusioni Con lo split 60/40 il RandomForest sembrerebbe essere il migliore. AdaBoost si comporta peggio rispetto allo split 50/50. DecisionTree anche migliora rispetto al 50/50.

Split 70/30

Regressor	MAE	RMSE	MSLE
SVR	0.0351675	0.0013755	0.0013111
Decision Tree	0.0066729	0.0001254	0.0001196
Random Forest	0.0067262	0.0001148	0.0001100
AdaBoost	0.0080581	0.0001397	0.0001339

Tabella 52: Risultati regressore con dataset completo ridotto 70/30

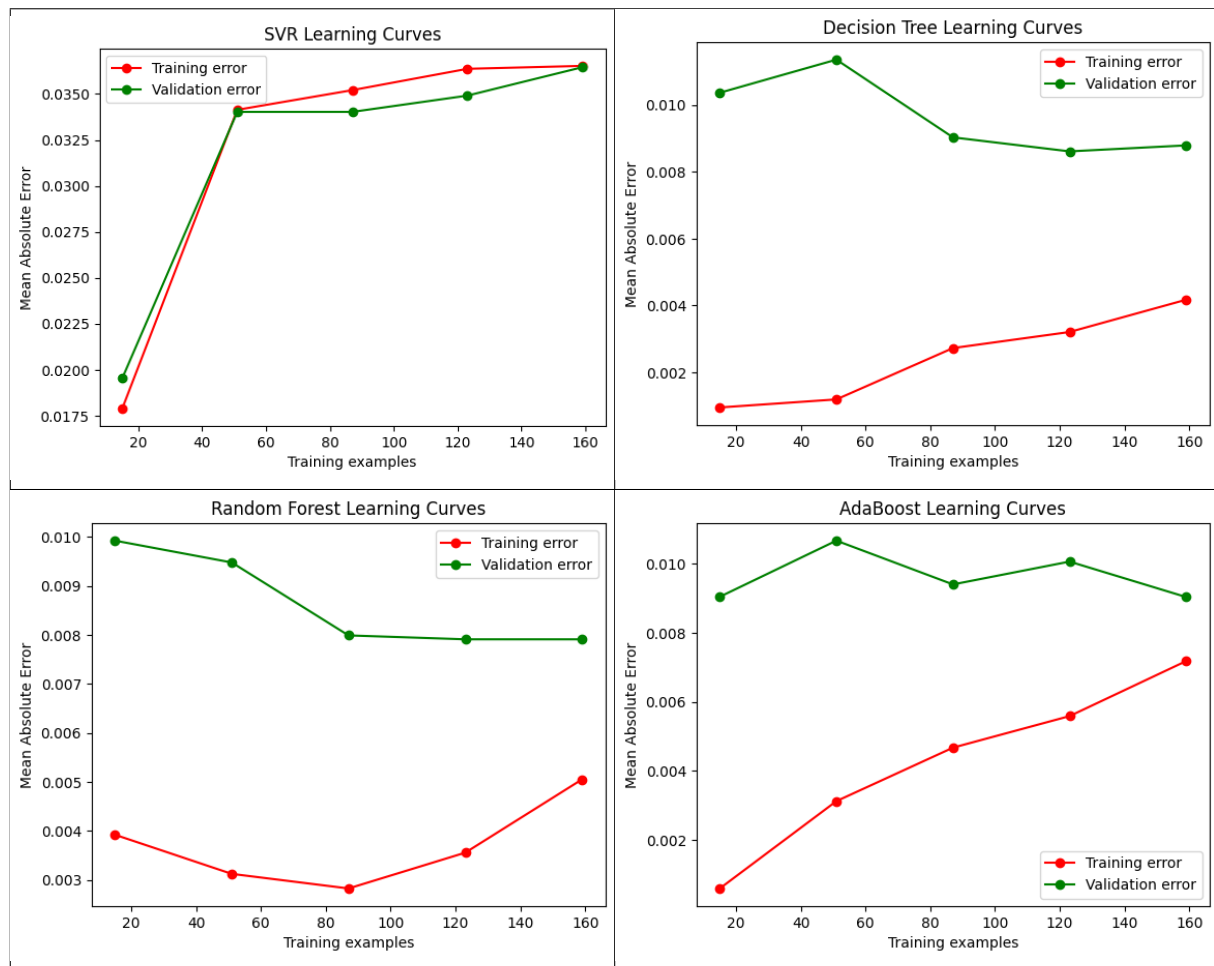


Tabella 53: Learning Curves con dataset completo ridotto 70/30

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor, MAE esclusa dove si comporta meglio. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation si stabilizza intorno alle 120 istanze. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curve mostra che la curva di training all'inizio diminuisce all'aumentare del numero delle istanze ma da 85 in poi aumenta. Quella di validation, invece, sulle 80 istanze si stabilizza. Il modello dunque non ha generalizzato bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Decision Tree Regressor e al Random Forest Regressor. La learning curve mostrano che il modello non è riuscito a generalizzare bene, in quanto quella di validation aumenta e diminuisce all'aumentare delle istanze, mentre quella di training aumenta.

Conclusioni Con lo split 70/30 il RandomForest sembrerebbe essere il migliore, ma si comporta peggio rispetto al 60/40. DecisionTree e AdaBoost si comportano peggio rispetto al 60/40.

Split 80/20

Regressor	MAE	RMSE	MSLE
SVR	0.0344696	0.0013440	0.0012808
Decision Tree	0.0074149	0.0001473	0.0001397
Random Forest	0.0069751	0.0001199	0.0001140
AdaBoost	0.0094008	0.0001633	0.0001562

Tabella 54: Risultati regressore con dataset completo ridotto 80/20

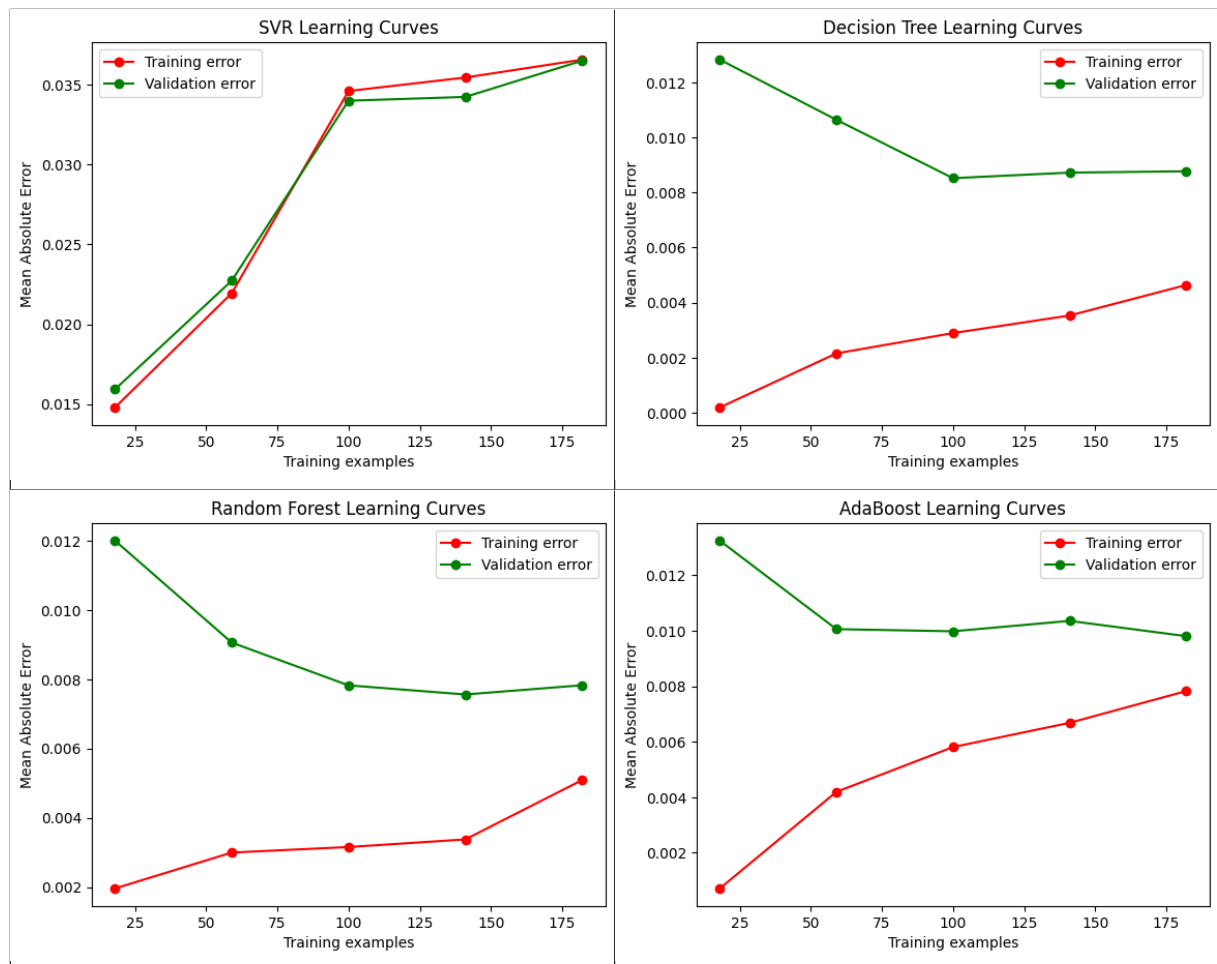


Tabella 55: Learning Curves con dataset completo ridotto 80/20

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation si stabilizza intorno alle 100 istanze. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curve mostra una curva di training che aumenta all'aumentare del numero delle istanze, mentre quella di validation si stabilizza intorno alle 100 istanze. Il modello dunque non ha generalizzato bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Decision Tree Regressor e al Random Forest Regressor. La learning curve mostrano che il modello non è riuscito a generalizzare bene in quanto quella di validation diminuisce all'aumentare delle istanze, mentre quella di training aumenta.

Conclusioni Con lo split 80/20 il RandomForest sembrerebbe essere il migliore. In generale tutti i modelli si comportano peggio rispetto allo split 70/30.

Split 90/10

Regressor	MAE	RMSE	MSLE
SVR	0.0296920	0.0011083	0.0010553
Decision Tree	0.0096568	0.0002100	0.0002000
Random Forest	0.0088683	0.0001676	0.0001582
AdaBoost	0.0115947	0.0002257	0.0002146

Tabella 56: Risultati regressore con dataset completo ridotto 90/10

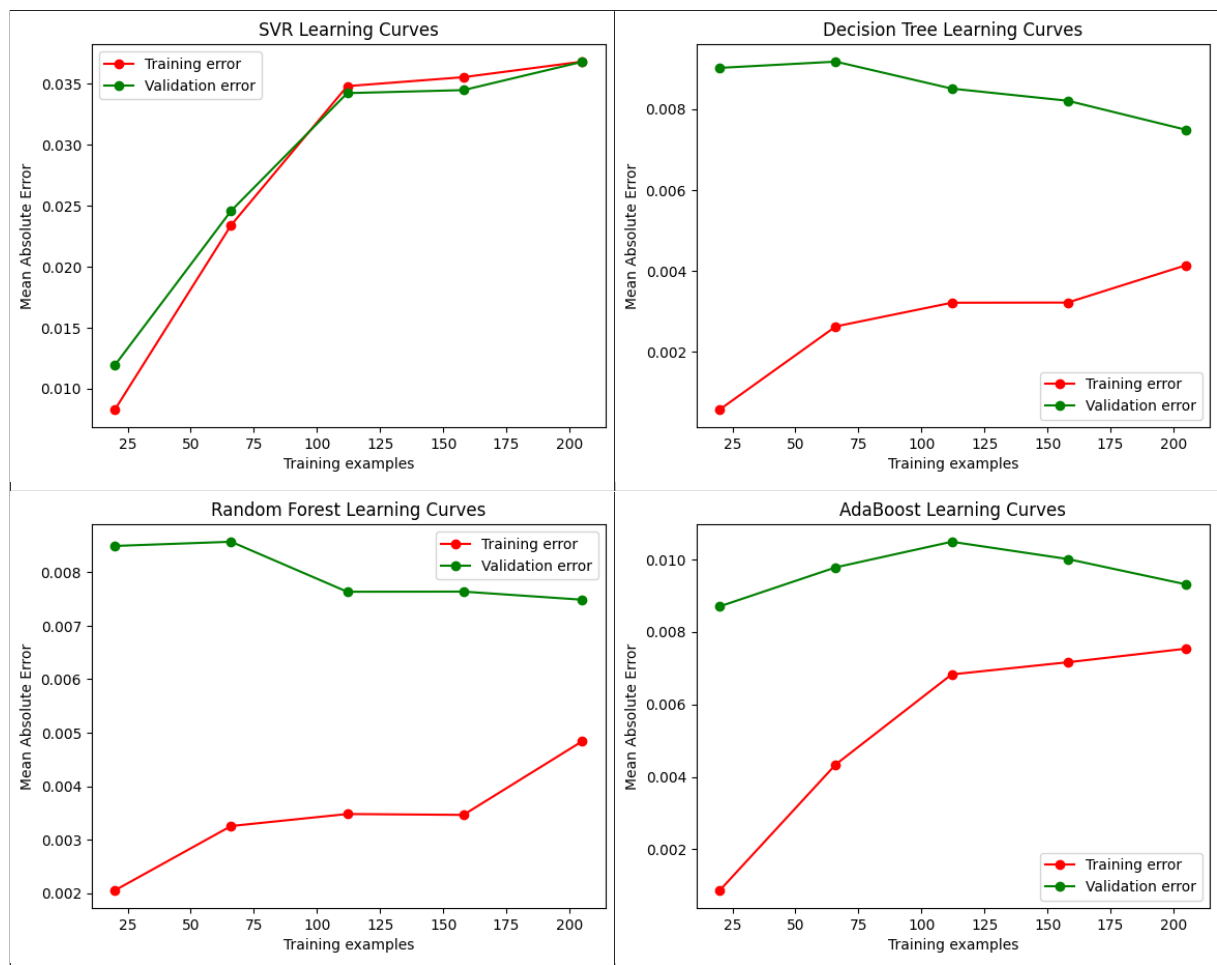


Tabella 57: Learning Curves con dataset completo ridotto 90/10

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati secondi al Random Forest Regressor. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation si stabilizza sulle 150 istanze. Il modello dunque non ha generalizzato bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Decision Tree Regressor in MAE ma risultati molto simili al Random Forest Regressor. La learning curve mostrano che il modello non è riuscito a generalizzare bene in quanto quella di training aumenta all'aumentare delle istanze, mentre quella di validation prima aumenta e poi comincia una lenta discesa.

Conclusioni Con lo split 90/10 il RandomForest sembrerebbe essere il migliore e l'AbaBoost sembra migliorare rispetto al passato, anche se ancora non riesce a generalizzare bene. In generale i modelli si comportano peggio rispetto agli split precedenti. SVR migliora rispetto agli split precedenti, ma rimane comunque il peggior modello.

Conclusioni generali Come si può notare in generale la riduzione del dataset ha portato a miglioramenti. Per quanto riguarda i diversi split si può notare che alcuni modelli performano meglio con alcuni split, altri modelli con altri. RandomForestRegressor risulta il migliore con lo split 60/40

9.2.2 Dataset AZURE

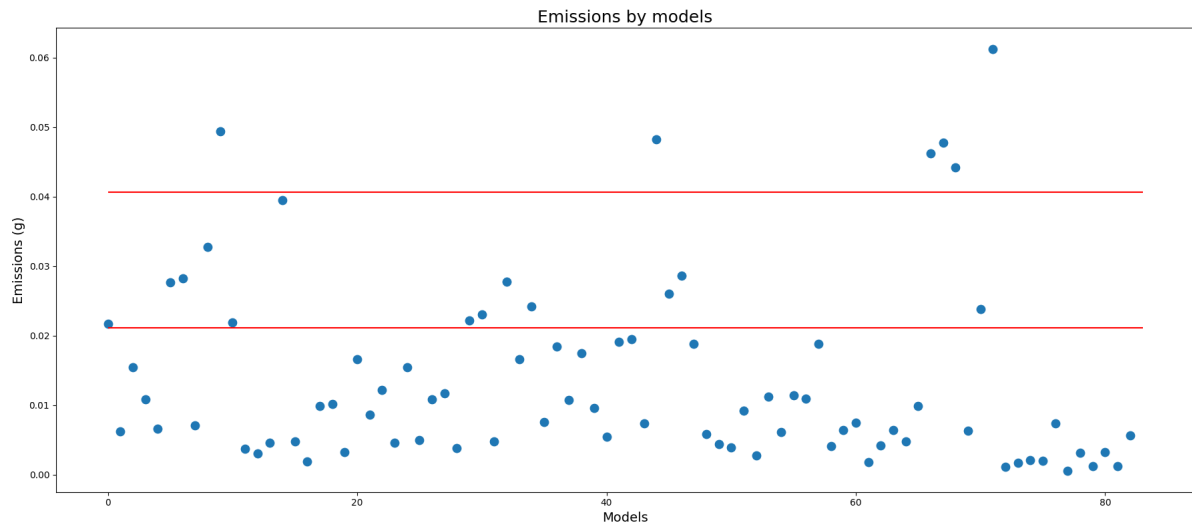


Figura 10: Distribuzione delle emissioni nel dataset Azure ridotto

E' possibile notare un miglior bilanciamento delle emissioni. Per quanto riguarda i modelli, sono stati eseguiti addestramenti con diversi split tra training e test set

Split 50/50

Regressor	MAE	RMSE	MSLE
SVR	0.0201690	0.0004759	0.0004582
Decision Tree	0.0088806	0.0001921	0.0001823
Random Forest	0.0083870	0.0001706	0.0001619
AdaBoost	0.0087987	0.0001824	0.0001728

Tabella 58: Risultati ottenuti con il dataset Azure ridotto 50/50

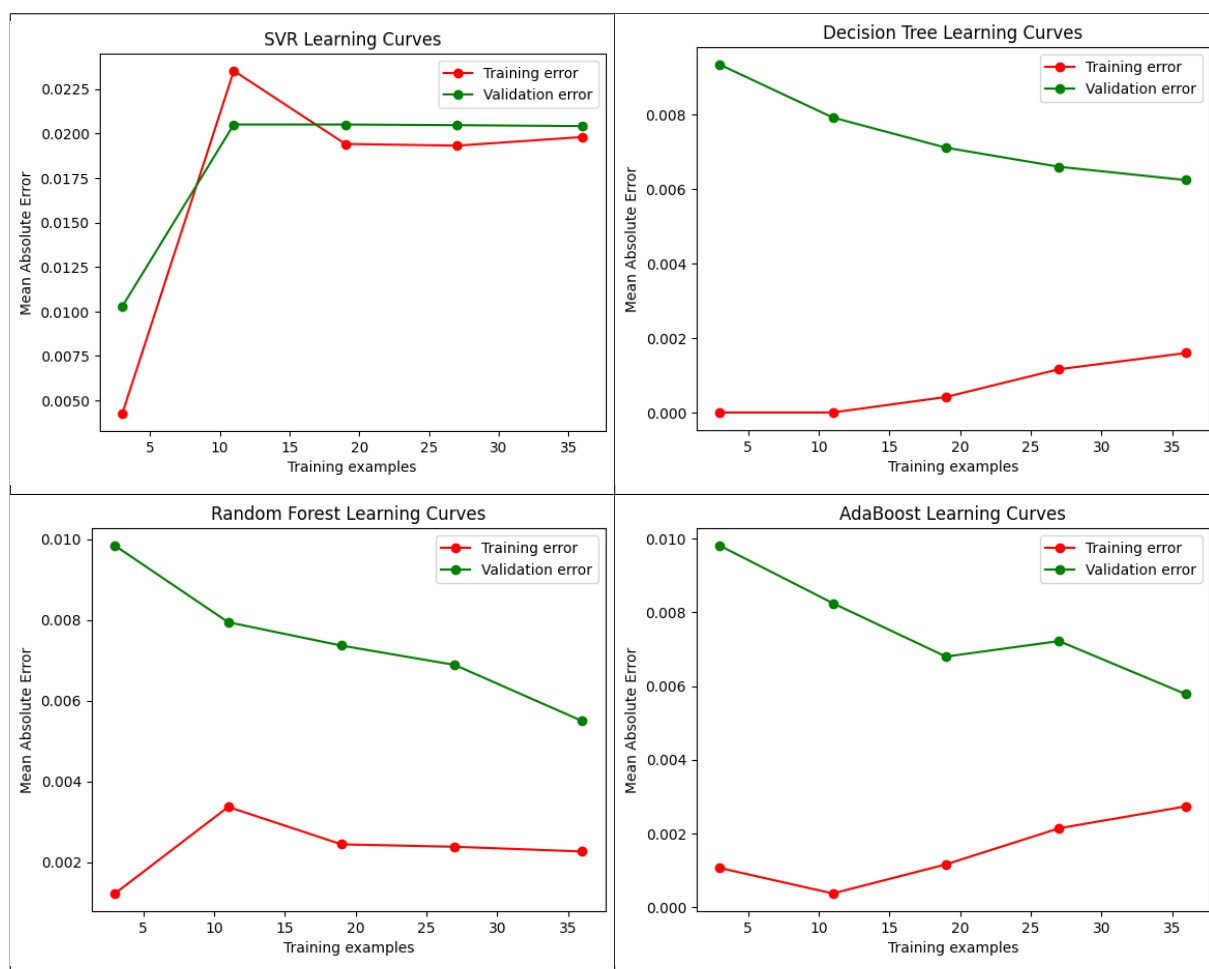


Tabella 59: Learning Curves con il dataset Azure ridotto 50/50

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze (quella di training intorno alle 10 istanze supera quella di validation).

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati peggiori al Random Forest Regressor e ad AdaBoost. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curve mostra che entrambe le curve di training e di validation diminuiscono all'aumentare del numero di istanze. Il modello sembrerebbe iniziare a generalizzare bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati secondi al Random Forest Regressor. La curva di training aumenta all'aumentare delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Conclusioni Con lo split 50/50 il RandomForest sembrerebbe essere il migliore. Decision-Tree si comporta meglio rispetto al passato. AdaBoost peggiora rispetto al passato.

Split 60/40

Regressor	MAE	RMSE	MSLE
SVR	0.0203710	0.0004819	0.0004645
Decision Tree	0.0079792	0.0001632	0.0001549
Random Forest	0.0076229	0.0001404	0.0001337
AdaBoost	0.0078253	0.0001293	0.0001236

Tabella 60: Risultati ottenuti con il dataset Azure ridotto 60/40

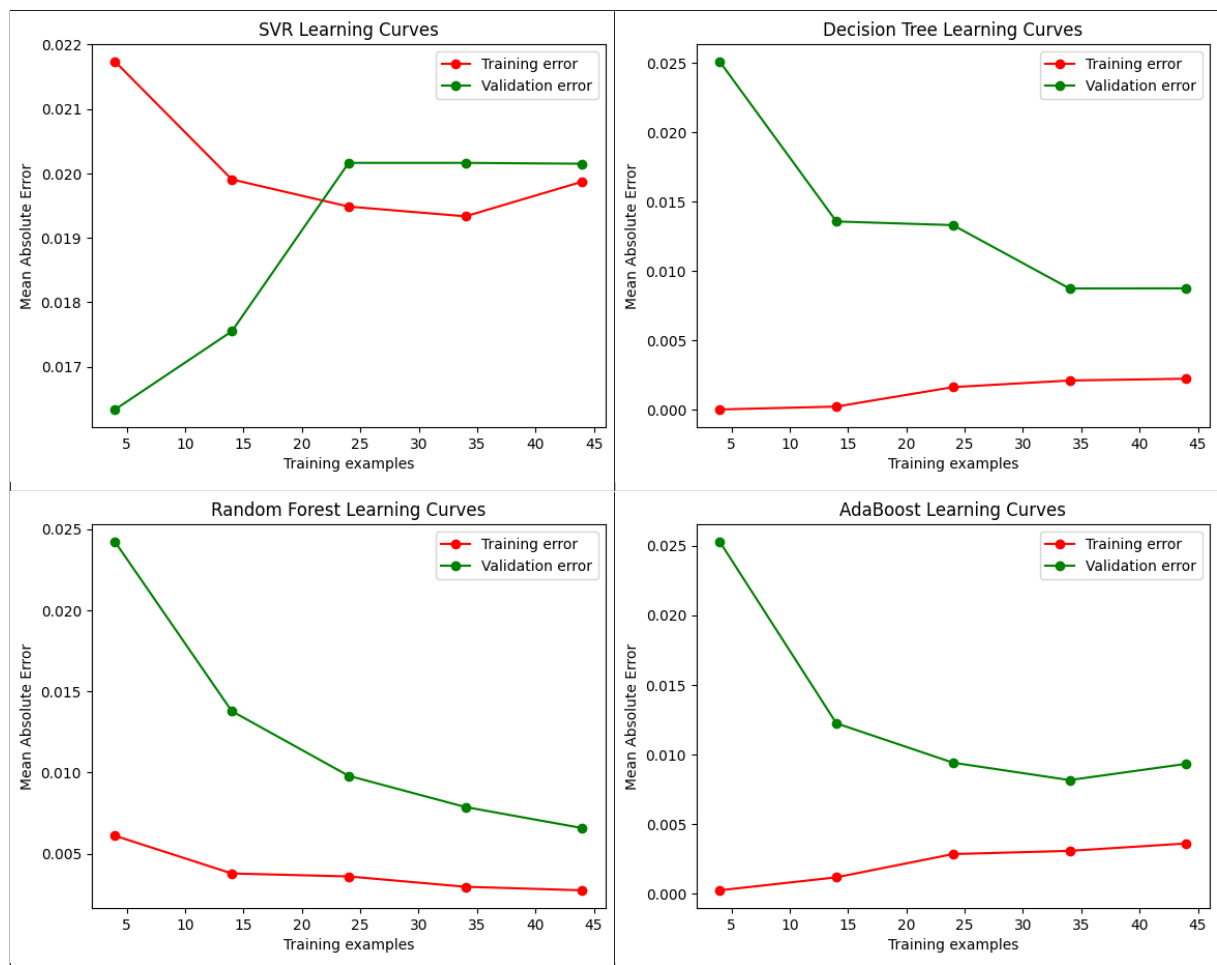


Tabella 61: Learning Curves con il dataset Azure ridotto 60/40

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati peggiori al Random Forest Regressor e ad AdaBoost. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curve mostra che entrambe le curve di training e di validation diminuiscono all'aumentare del numero di istanze. Il modello sembrerebbe iniziare a generalizzare bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati secondi al Random Forest Regressor. La curva di training aumenta all'aumentare delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Conclusioni Con lo split 60/40 il RandomForest sembrerebbe essere il migliore. Decision-Tree si comporta meglio rispetto allo split precedente. AdaBoost migliora rispetto allo split precedente.

Split 70/30

Regressor	MAE	RMSE	MSLE
SVR	0.0205812	0.0004825	0.0004649
Decision Tree	0.0086633	0.0001952	0.0001848
Random Forest	0.0079448	0.0001688	0.0001604
AdaBoost	0.0087171	0.0001541	0.0001471

Tabella 62: Risultati ottenuti con il dataset Azure ridotto 70/30

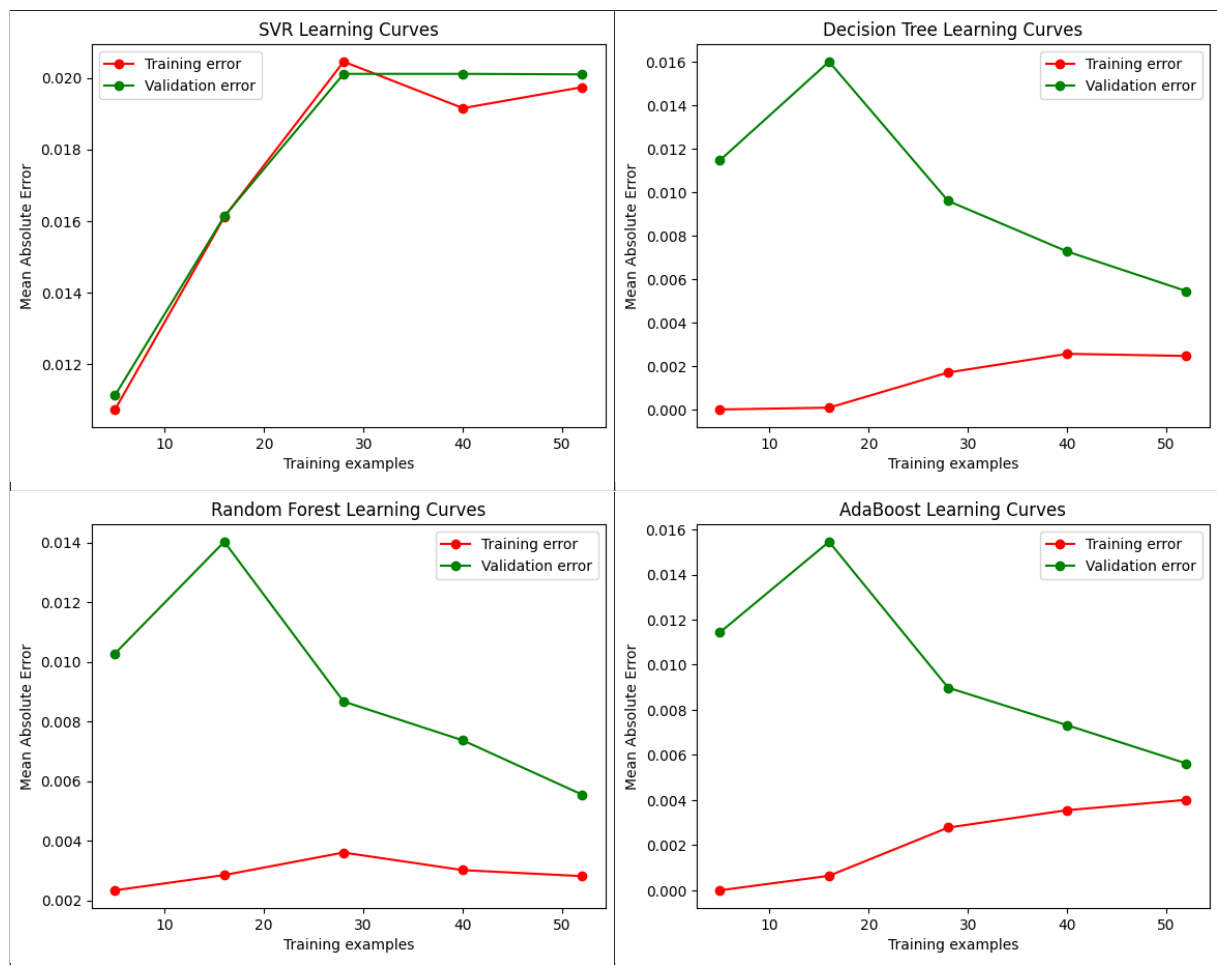


Tabella 63: Learning Curves con il dataset Azure ridotto 70/30

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati peggiori al Random Forest Regressor e ad AdaBoost. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello PER MAE. La learning curve mostra che entrambe le curve di training e di validation diminuiscono all'aumentare del numero di istanze. Il modello sembrerebbe iniziare a generalizzare bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto i migliori risultati in RMSE e MSLE. La curva di training aumenta all'aumentare delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Conclusioni Con lo split 70/30 il RandomForest sembrerebbe essere il migliore. Decision-Tree si comporta peggio rispetto allo split precedente. AdaBoost peggiora rispetto allo split precedente.

Split 80/20

Regressor	MAE	RMSE	MSLE
SVR	0.0207895	0.0004925	0.0004749
Decision Tree	0.0089766	0.0001894	0.0001786
Random Forest	0.0067808	0.0001078	0.0001032
AdaBoost	0.0076901	0.0001143	0.0001092

Tabella 64: Risultati ottenuti con il dataset Azure ridotto 80/20

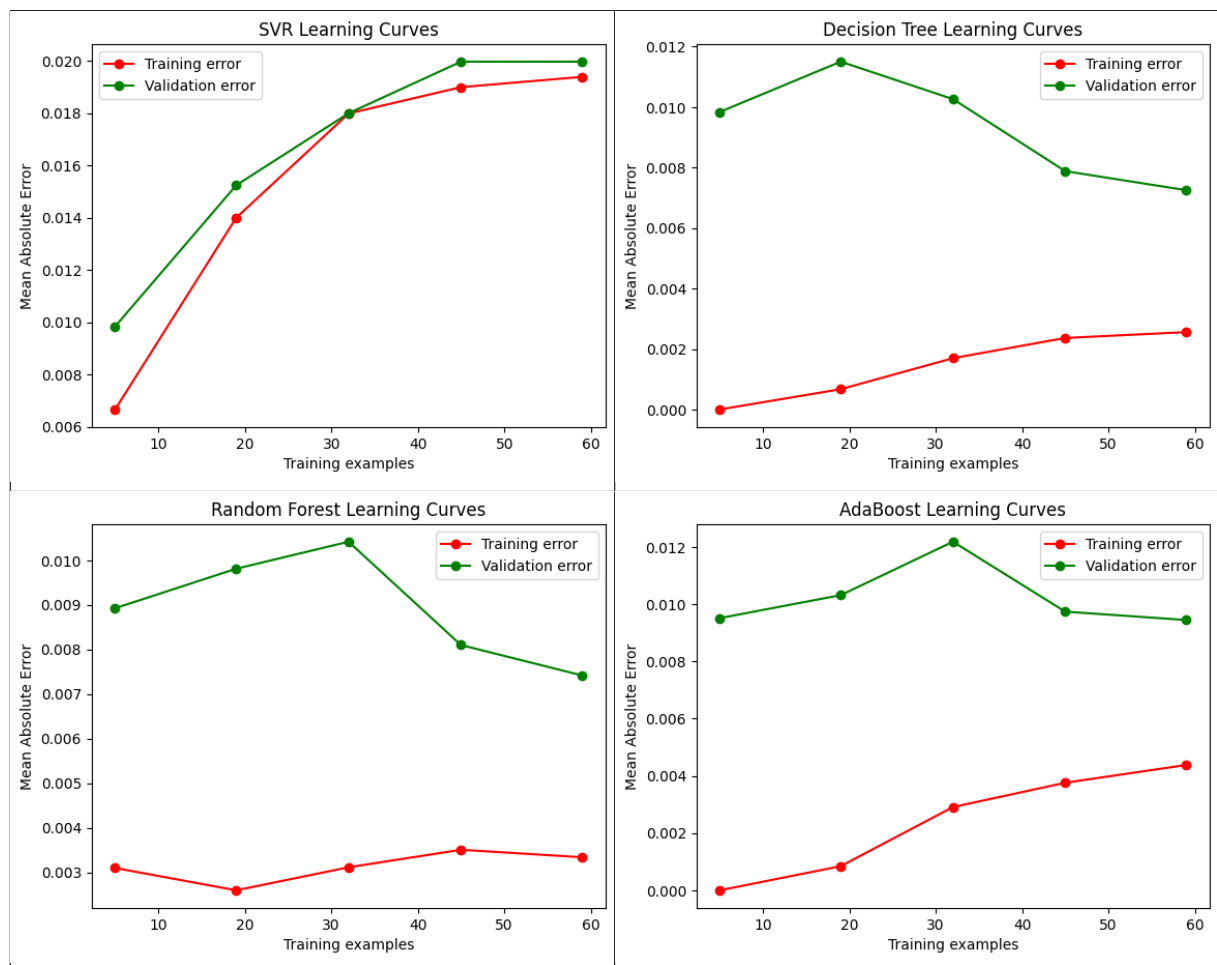


Tabella 65: Learning Curves con il dataset Azure ridotto 80/20

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze.

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto i risultati peggiori al Random Forest Regressor e ad AdaBoost. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curve mostra che la curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation prima aumenta e poi diminuisce. Il modello dunque non ha generalizzato bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati secondi al Random Forest Regressor. La curva di training aumenta all'aumentare delle istanze, mentre quella di validation aumenta e poi diminuisce. Il modello dunque non ha generalizzato bene.

Conclusioni Con lo split 80/20 il RandomForest sembrerebbe essere il migliore. Decision-Tree si comporta peggio rispetto allo split precedente. AdaBoost migliora rispetto allo split precedente.

Split 90/10

Regressor	MAE	RMSE	MSLE
SVR	0.0229610	0.0005560	0.0005361
Decision Tree	0.0049908	0.0000352	0.0000345
Random Forest	0.0045819	0.0000323	0.0000317
AdaBoost	0.0049257	0.0000379	0.0000371

Tabella 66: Risultati ottenuti con il dataset Azure ridotto 90/10

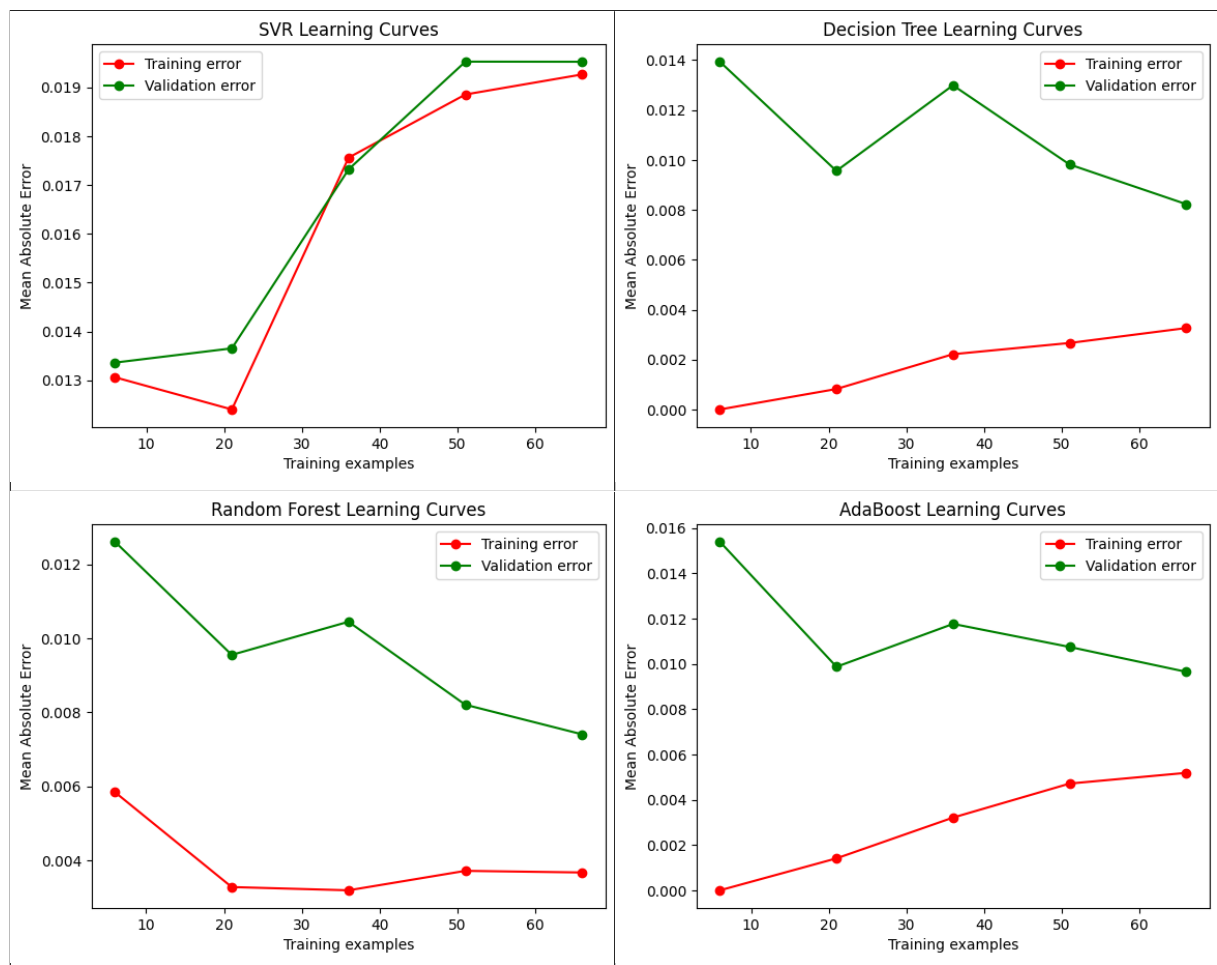


Tabella 67: Learning Curves con il dataset Azure ridotto 90/10

SVR. Dal punto di vista delle metriche il regressore SVR ha ottenuto i peggiori risultati in termini di MAE, RMSE e MSLE. La learning curve mostra che il modello non è in grado di generalizzare, infatti gli errori aumentano all'aumentare del numero delle istanze (quella di training intorno alle 10 istanze supera quella di validation).

Decision Tree Regressor. Dal punto di vista delle metriche, il Decision Tree Regressor ha ottenuto risultati secondi al Random Forest Regressor. La learning curve mostra che la

curva di training aumenta all'aumentare del numero delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Random Forest Regressor. Dal punto di vista delle metriche il Random Forest Regressor risulta essere il miglior modello. La learning curve mostra che entrambe le curve di training e di validation diminuiscono all'aumentare del numero di istanze. Il modello sembrerebbe iniziare a generalizzare bene.

AdaBoost Regressor. Dal punto di vista delle metriche il AdaBoost Regressor ha ottenuto risultati peggiori rispetto al Random Forest Regressor e al Decision Tree Regressor. La curva di training aumenta all'aumentare delle istanze, mentre quella di validation diminuisce. Il modello dunque non ha generalizzato bene.

Conclusioni Con lo split 90/10 il RandomForest sembrerebbe essere il migliore. In genere tutti i modelli (SVR escluso) migliorano rispetto allo split precedente

Conclusioni generali RandomForest sembra essere ancora il migliore. Inoltre il dataset completo si comporta meglio rispetto a quello Azure senza e con riduzione

9.2.3 Confronto tra i quattro dataset

Si può notare che in generale il dataset completo performa meglio del dataset di Azure. E' possibile anche notare come i dataset ridotti performano meglio di quelli completi. Il dataset che performa meglio in assoluto è dunque il dataset completo ridotto. Riassumendo:

- Dataset completo: RandomForest con split 70/30
- Dataset Azure: RandomForest con split 90/10
- Dataset completo ridotto: RandomForest con split 60/40
- Dataset Azure ridotto: RandomForest con split 90/10

10 Errori delle classi

Lo step successivo è stato quello di analizzare gli errori di predizione delle classi per ogni dataset. Si è deciso di procedere per ogni dataset con il modello del RandomForestRegressor in quanto risulta essere il migliore in ogni caso. Per quanto riguarda lo splitting dei dati, si è deciso di utilizzare uno split 70-30, ovvero il 70% dei dati per il training e il 30% per il testing, questo per avere delle misurazioni "confrontabili". Il dataset è stato diviso in 3 classi, ovvero **low**, **medium** e **high**, in base ai valori di emissioni di CO2 (rispettivamente codificate con 0,1,2). Per ogni elemento del test set, sono stati calcolati gli errori e successivamente si è proceduto a calcolare la media degli errori per ogni classe. Gli errori sono stati calcolati come segue:

- **Errore assoluto:**

$$|y - \hat{y}|$$

- **Errore percentuale:**

$$\frac{|y - \hat{y}|}{y} \cdot 100$$

dove y è il valore reale e \hat{y} è il valore predetto.

10.1 Dataset Completo

Il dataset completo conta 265 righe, di cui 185 per il training e 80 per il testing. Il calcolo delle classi vede come **low_bound** delle emissioni 0.36027045739563296 mentre l' **high_bound** è 0.7205403996360675. Nella totalità abbiamo 262 righe nella classe **low**, 1 nella classe **medium** e 2 nella classe **high**.

Classe	Numero elementi	Errore assoluto medio	Errore percentuale medio
low	78	0.021805	813
medium	0	-	-
high	2	0.790026	80

Tabella 68: Errori delle classi per il dataset completo

10.2 Dataset Completo Ridotto

Il dataset completo "ridotto" conta 254 righe, di cui 177 per il training e 77 per il testing. Il calcolo delle classi vede come **low_bound** delle emissioni 0.029759662206275274 mentre l' **high_bound** è 0.059518809257352256. Nella totalità abbiamo 228 righe nella classe **low**, 17 nella classe **medium** e 9 nella classe **high**.

Classe	Numero elementi	Errore assoluto medio	Errore percentuale medio
low	69	0.004995	1723.95
medium	7	0.021113	51.97
high	1	0.025475	41.23

Tabella 69: Errori delle classi per il dataset completo ridotto

10.3 Dataset Azure

Il dataset Azure conta 91 righe, di cui 63 per il training e 28 per il testing. Il calcolo delle classi vede come **low_bound** delle emissioni 0.3610033427811713 mentre l' **high_bound** è 0.7203571782896829. Nella totalità abbiamo 88 righe nella classe **low**, 1 nella classe **medium** e 2 nella classe **high**.

Classe	Numero elementi	Errore assoluto medio	Errore percentuale medio
low	78	0.044774	188.63
medium	0	-	-
high	2	0.588654	66.76

Tabella 70: Errori delle classi per il dataset Azure

10.4 Dataset Azure Ridotto

Il dataset Azure "ridotto" conta 83 righe, di cui 58 per il training e 25 per il testing. Il calcolo delle classi vede come **low_bound** delle emissioni 0.021137526211647432 mentre l' **high_bound** è 0.04062554515063527. Nella totalità abbiamo 64 righe nella classe **low**, 13 nella classe **medium** e 6 nella classe **high**.

Classe	Numero elementi	Errore assoluto medio	Errore percentuale medio
low	20	0.005607	103.15
medium	5	0.017295	61.7
high	0	-	-

Tabella 71: Errori delle classi per il dataset Azure ridotto

Riferimenti bibliografici

- [1] Qingyao Ai, Vahid Azizi, Xu Chen, and Yongfeng Zhang. Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms*, 11(9), 2018.
- [2] Fabio Aiolli. Efficient top-n recommendation for very large scale binary rated datasets. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 273–280, New York, NY, USA, 2013. Association for Computing Machinery.
- [3] Ting Bai, Ji-Rong Wen, Jun Zhang, and Wayne Xin Zhao. A neural collaborative filtering model with interaction-based neighborhood. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, page 1979–1982, New York, NY, USA, 2017. Association for Computing Machinery.
- [4] Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *The World Wide Web Conference*, WWW '19, page 151–161, New York, NY, USA, 2019. Association for Computing Machinery.
- [5] Chong Chen, Min Zhang, Yongfeng Zhang, Yiqun Liu, and Shaoping Ma. Efficient neural matrix factorization without sampling for recommendation. *ACM Trans. Inf. Syst.*, 38(2), jan 2020.
- [6] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 639–648, New York, NY, USA, 2020. Association for Computing Machinery.
- [7] Xiangnan He, Xiaoyu Du, Xiang Wang, Feng Tian, Jinhui Tang, and Tat-Seng Chua. Outer product-based neural collaborative filtering. *arXiv preprint arXiv:1808.03912*, 2018.
- [8] Xiangnan He, Zhankui He, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. Nais: Neural attentive item similarity model for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 30(12):2354–2366, December 2018.
- [9] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 173–182, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.
- [10] Santosh Kabbur, Xia Ning, and George Karypis. Fism: factored item similarity models for top-n recommender systems. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 659–667, New York, NY, USA, 2013. Association for Computing Machinery.
- [11] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 689–698, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.
- [12] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 689–698, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

- [13] Zihan Lin, Changxin Tian, Yupeng Hou, and Wayne Xin Zhao. Improving graph collaborative filtering with neighborhood-enriched contrastive learning. In *Proceedings of the ACM Web Conference 2022, WWW '22*. ACM, April 2022.
- [14] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. *Learning disentangled representations for recommendation*. Curran Associates Inc., Red Hook, NY, USA, 2019.
- [15] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. Simplex: A simple and strong baseline for collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21*, page 1243–1252, New York, NY, USA, 2021. Association for Computing Machinery.
- [16] Xia Ning and George Karypis. Slim: Sparse linear methods for top-n recommender systems. In *2011 IEEE 11th International Conference on Data Mining*, pages 497–506, 2011.
- [17] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09*, page 452–461, Arlington, Virginia, USA, 2009. AUAI Press.
- [18] Ilya Shenbin, Anton Alekseev, Elena Tutubalina, Valentin Malykh, and Sergey I. Nikolenko. Recvae: A new variational autoencoder for top-n recommendations with implicit feedback. In *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20*, page 528–536, New York, NY, USA, 2020. Association for Computing Machinery.
- [19] Giuseppe Spillo, Allegra De Filippo, Cataldo Musto, Michela Milano, and Giovanni Semeraro. Towards sustainability-aware recommender systems: analyzing the trade-off between algorithms performance and carbon footprint. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 856–862, 2023.
- [20] Harald Steck. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference, WWW '19*, page 3251–3257, New York, NY, USA, 2019. Association for Computing Machinery.
- [21] Harald Steck, Maria Dimakopoulou, Nickolai Riabov, and Tony Jebara. Admm slim: Sparse recommendations for many users. In *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20*, page 555–563, New York, NY, USA, 2020. Association for Computing Machinery.
- [22] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web, WWW '15*, page 1067–1077, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.
- [23] Rianne van den Berg, Thomas N. Kipf, and Max Welling. Graph convolutional matrix completion, 2017.
- [24] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *Proceedings of the 27th ACM International Conference on Information*

- and Knowledge Management*, CIKM '18, page 417–426, New York, NY, USA, 2018. Association for Computing Machinery.
- [25] Hongwei Wang, Fuzheng Zhang, Mengdi Zhang, Jure Leskovec, Miao Zhao, Wenjie Li, and Zhongyuan Wang. Knowledge-aware graph neural networks with label smoothness regularization for recommender systems, 2019.
- [26] Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. Multi-task feature learning for knowledge graph enhanced recommendation, 2019.
- [27] Hongwei Wang, Miao Zhao, Xing Xie, Wenjie Li, and Minyi Guo. Knowledge graph convolutional networks for recommender systems. In *The World Wide Web Conference*, WWW '19, page 3307–3313, New York, NY, USA, 2019. Association for Computing Machinery.
- [28] Wenjie Wang, Yiyang Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 832–841, New York, NY, USA, 2023. Association for Computing Machinery.
- [29] Wenjie Wang, Yiyang Xu, Fuli Feng, Xinyu Lin, Xiangnan He, and Tat-Seng Chua. Diffusion recommender model. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 832–841, New York, NY, USA, 2023. Association for Computing Machinery.
- [30] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 165–174, New York, NY, USA, 2019. Association for Computing Machinery.
- [31] Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat-Seng Chua. Learning intents behind interactions with knowledge graph for recommendation. In *Proceedings of the Web Conference 2021*, WWW '21, page 878–887, New York, NY, USA, 2021. Association for Computing Machinery.
- [32] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. Disentangled graph collaborative filtering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1001–1010, New York, NY, USA, 2020. Association for Computing Machinery.
- [33] Ga Wu, Maksims Volkovs, Chee Loong Soon, Scott Sanner, and Himanshu Rai. Noise contrastive estimation for one-class collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 135–144, New York, NY, USA, 2019. Association for Computing Machinery.
- [34] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. Self-supervised graph learning for recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21. ACM, July 2021.
- [35] Yao Wu, Christopher DuBois, Alice X. Zheng, and Martin Ester. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, page 153–162, New York, NY, USA, 2016. Association for Computing Machinery.

- [36] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep matrix factorization models for recommender systems. In *IJCAI*, volume 17, pages 3203–3209. Melbourne, Australia, 2017.
- [37] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 353–362, New York, NY, USA, 2016. Association for Computing Machinery.
- [38] Lei Zheng, Chun-Ta Lu, Fei Jiang, Jiawei Zhang, and Philip S Yu. Spectral collaborative filtering. In *Proceedings of the 12th ACM conference on recommender systems*, pages 311–319, 2018.