Netfloox System Recommandation

Présenté par Jonathan Caillaux, Arnaud Rambourg et Maximilien Proust

1) La base de données

Nom de la table	Nombre de lignes
name_basics	14 118 391
title_akas	49 705 848
title_basics	11 399 048
title_crew	10 826 426
title_episode	6 965 725
title_principals	90 465 504
title_ratings	1 526 047

1) La base de données

name_basics	
nconst	varchar(12)
primaryName	varchar(32)
birthYear	int2
deathYear	int2
primaryProfession	varchar(128)
knownForTitles	varchar(256)

title_akas	
titleId	varchar(12)
title	varchar(128)
ordering	int4
region	varchar(3)
language	varchar(3)
types	varchar(32)
attributes	varchar(32)
isOriginalTitle	bool

title_basics	
tconst	varchar(12)
titleType	varchar(8)
primaryTitle	varchar(128)
originalTitle	varchar(128)
isAdult	bool
startYear	int2
endYear	int2
runtimeMinutes	int2
genres	varchar(12)

1) La base de données

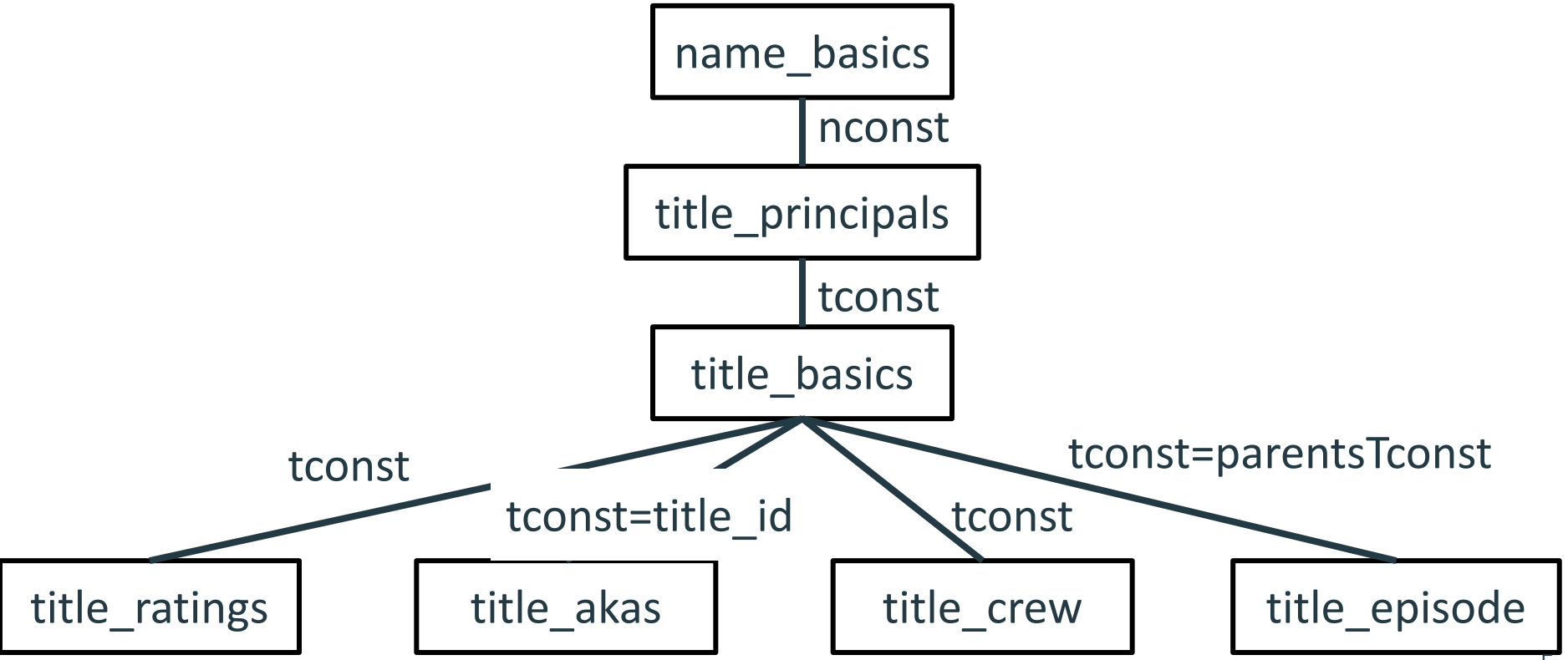
title_crew	
tconst	varchar(12)
directors	varchar(12)
writers	varchar(12)

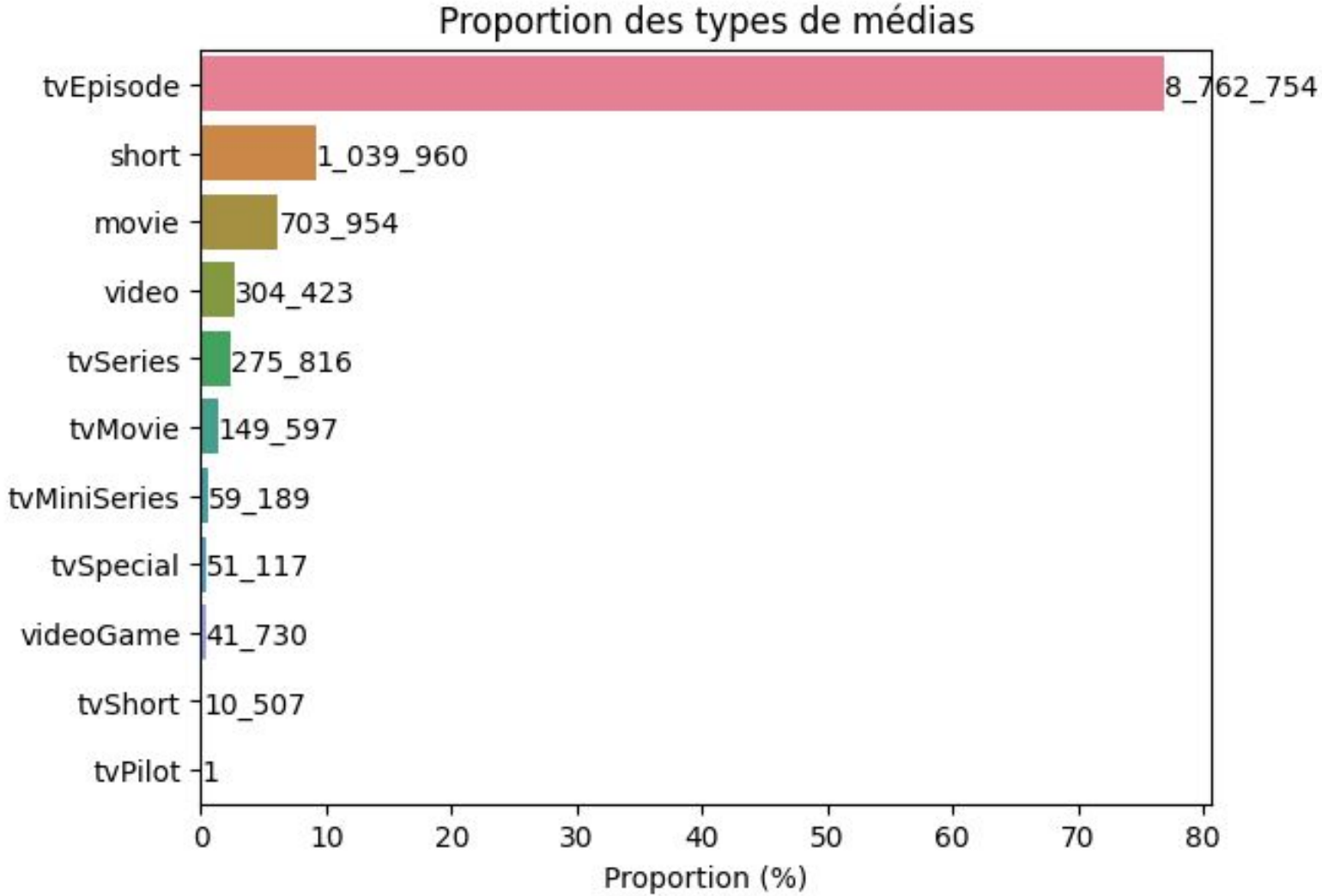
title_episode	
tconst	varchar(12)
parentTconst	varchar(12)
seasonNumber	int2
episodeNumber	int2

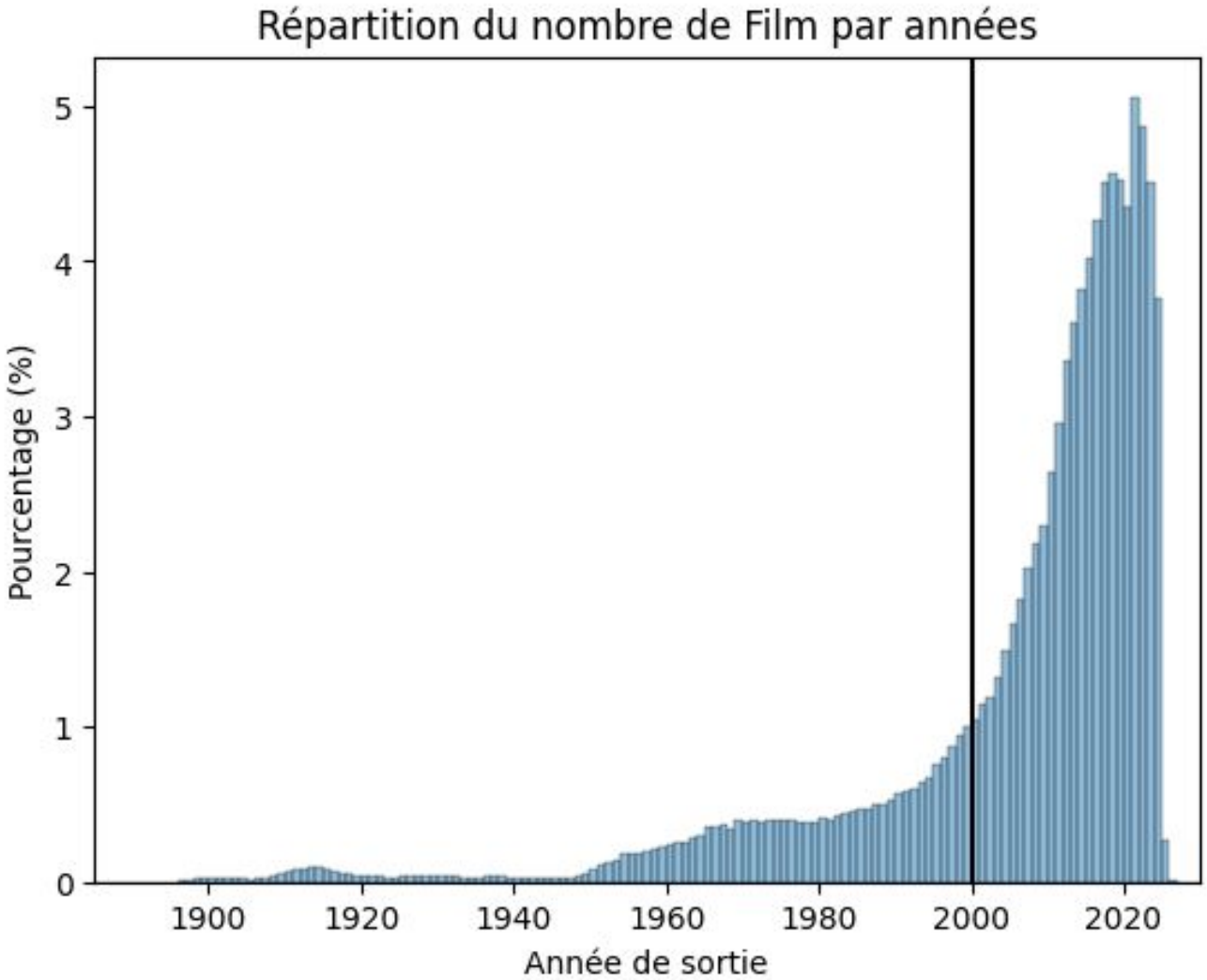
title_ratings	
tconst	varchar(12)
averageRating	float4
numVotes	int4

title_principals	
tconst	vatrchar(12)
ordering	int4
nconst	varchar(12)
category	varchar(20)
job	varchar(256)
characters	varchar(32)

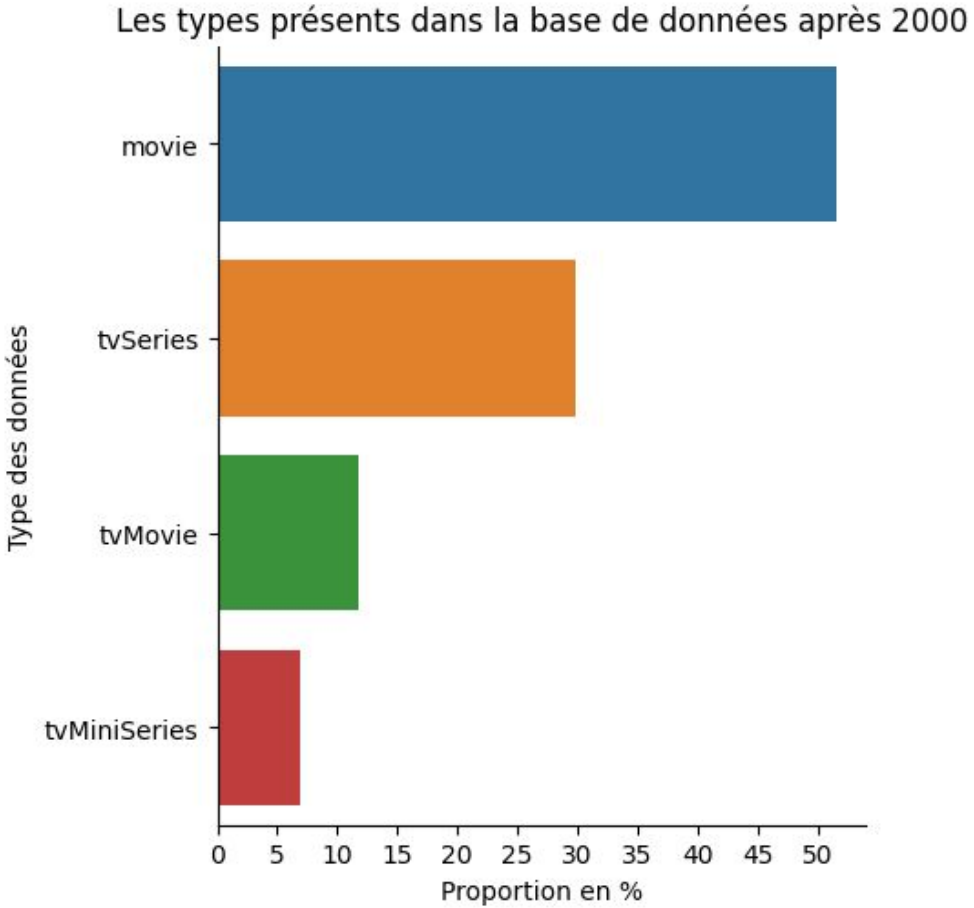
2) Le diagramme entité-relation

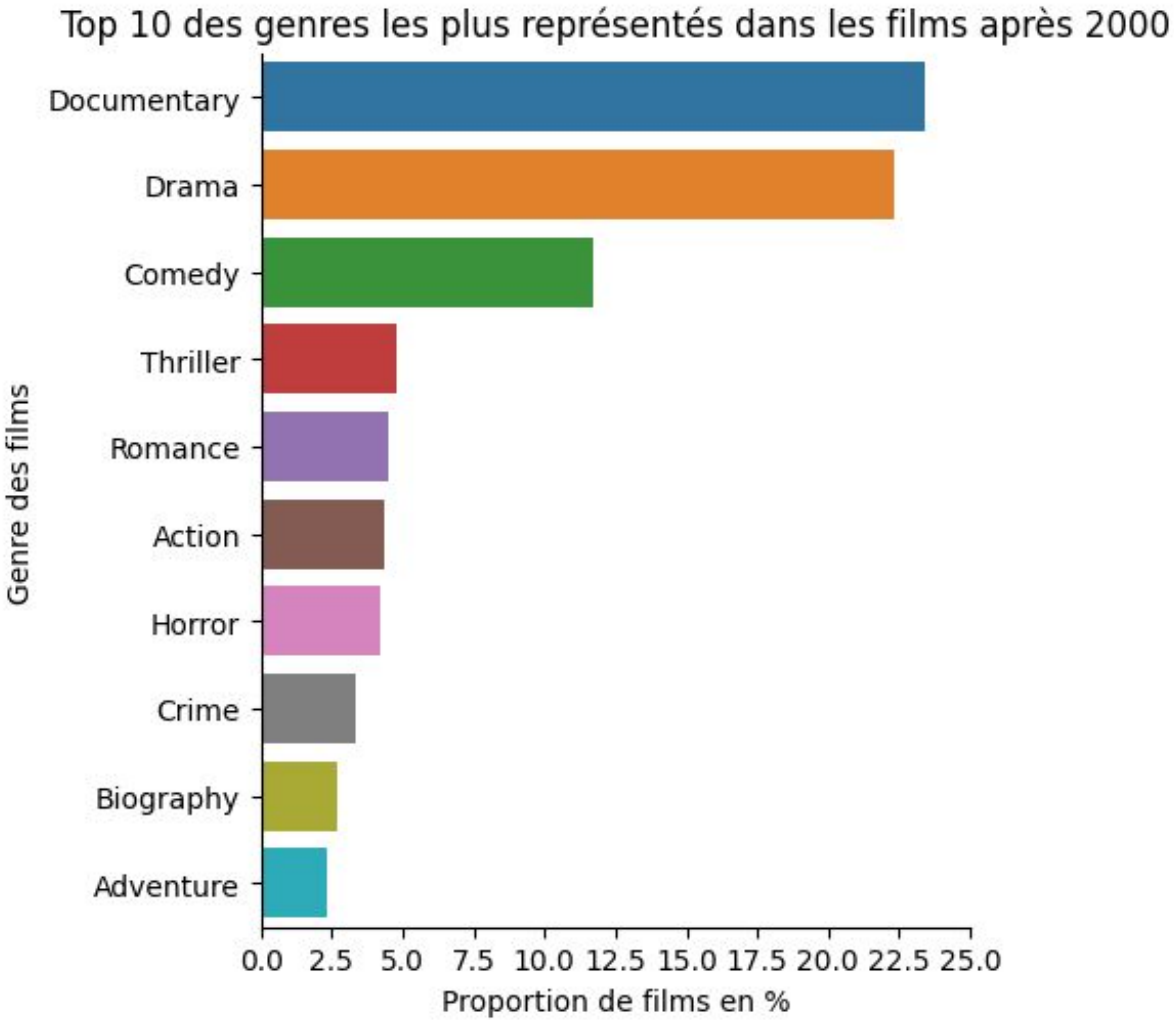



Proportion des types de médias
tvEpisode
short
movie
video
tvSeries
tvMovie
tvMiniSeries
tvSpecial
videoGame
tvShort
tvPilot
8_762_754
1_039_960
703_954
304_423
275_816
149_597
59_189
51_117
41_730
10_507
1
0
10
20
30
40
50
60
70
80
Proportion (%)

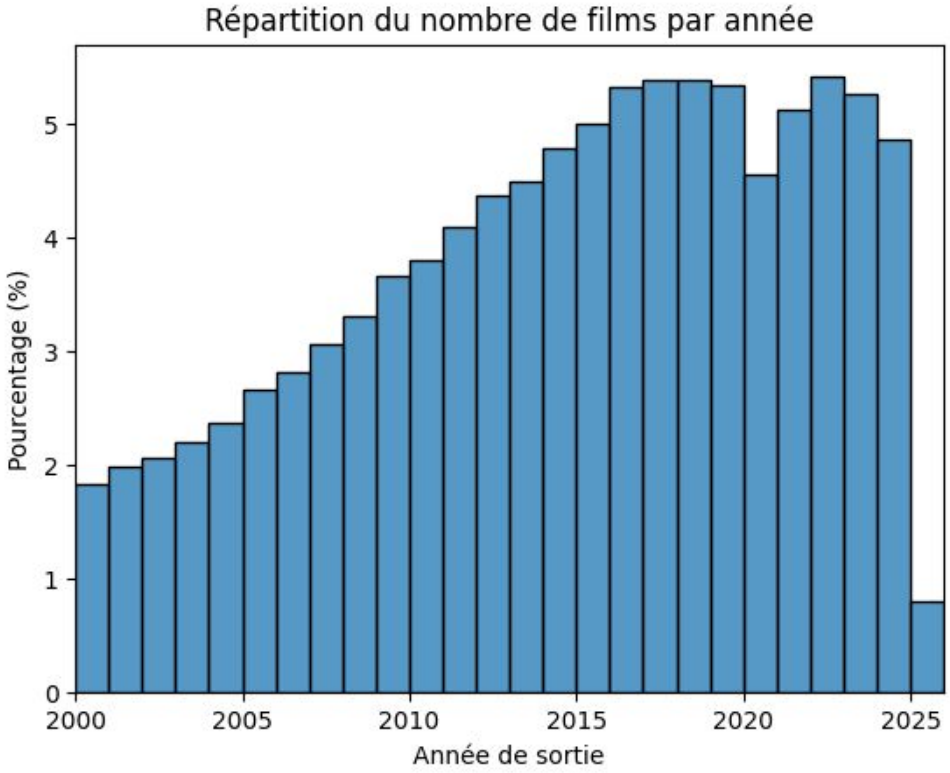

Répartition du nombre de Film par années
Pourcentage (%)
0
1
2
3
4
5
1900
1920
1940
1960
1980
2000
2020
Année de sortie

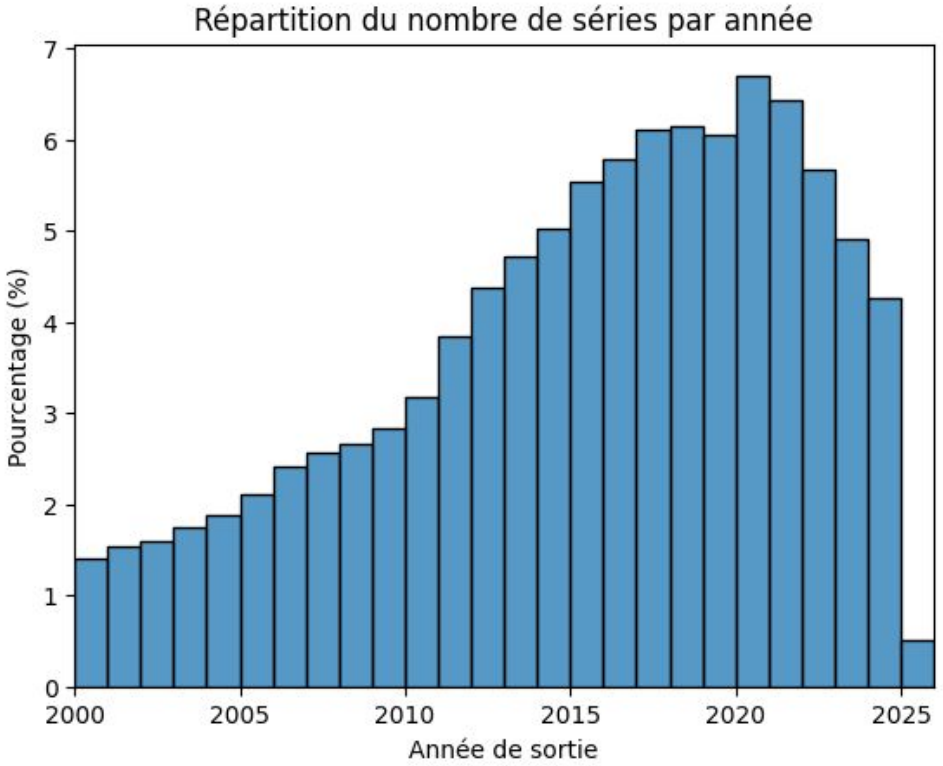
3) L’analyse exploratoire

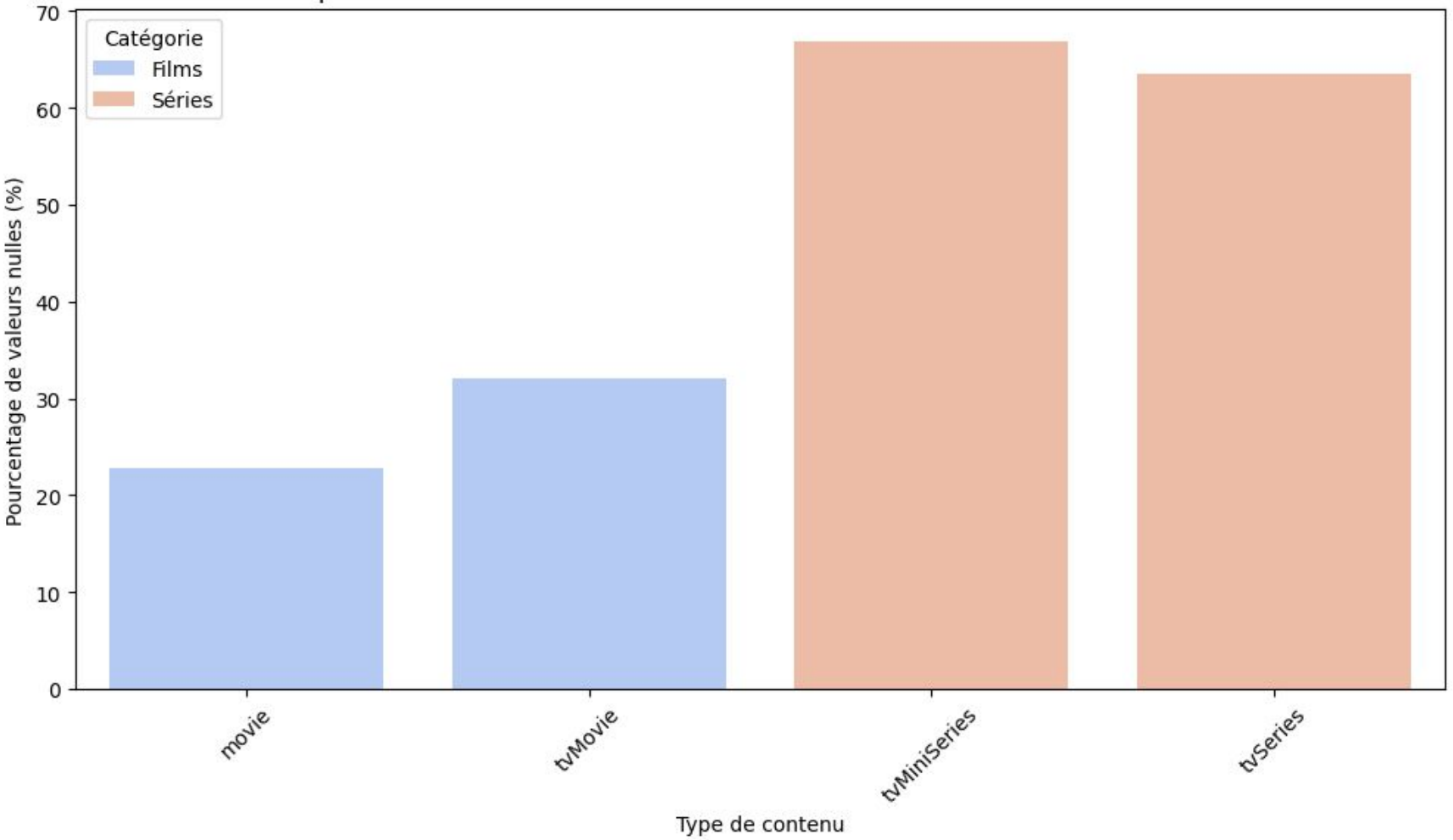


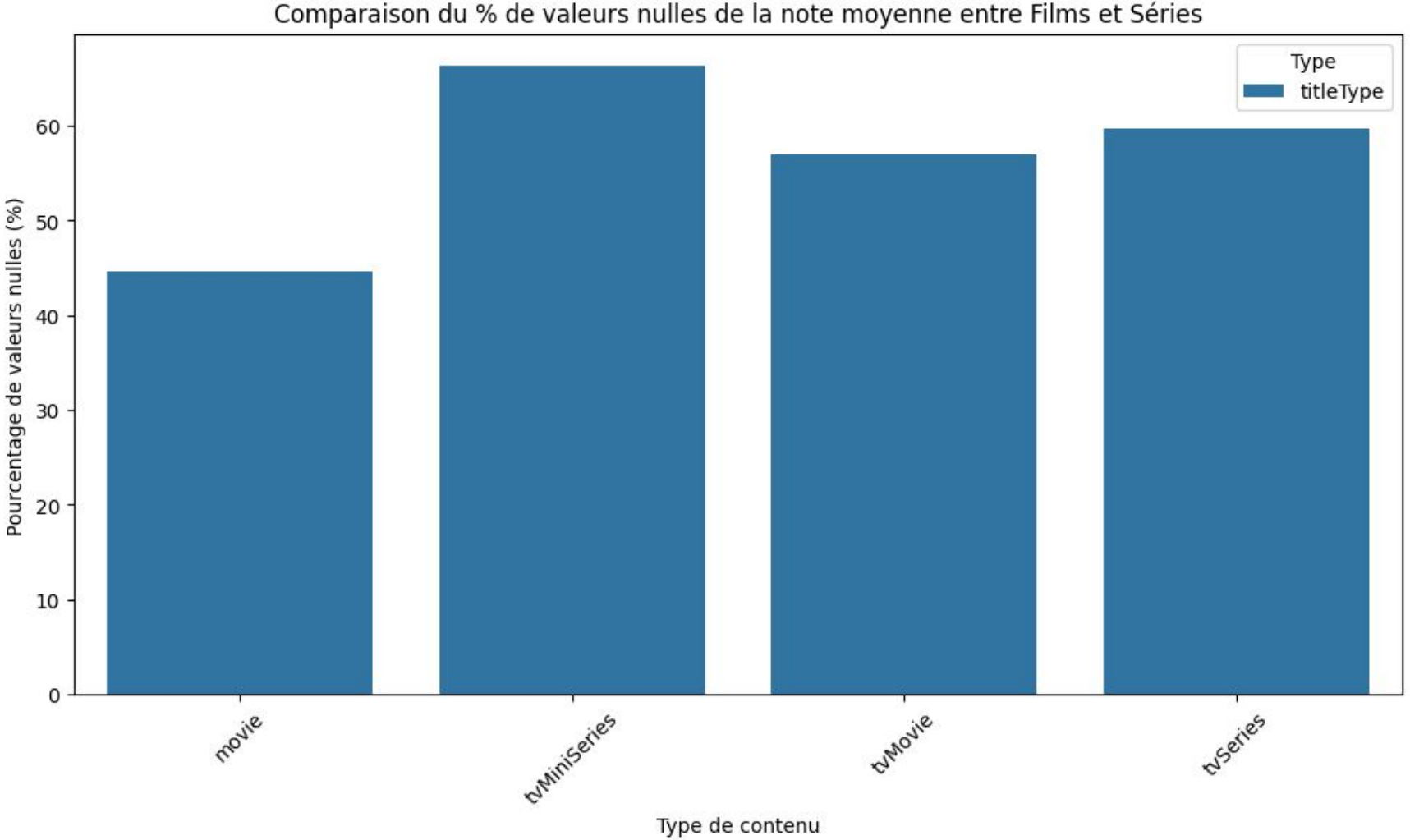

Top 10 des genres les plus représentés dans les films après 2000
Documentary
Drama
Comedy
Thriller
Romance
Action
Horror
Crime
Biography
Adventure
Genre des films
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0 22.5 25.0
Proportion de films en %


Top 10 des genres les plus représentés dans les séries après 2000
Comedy
Drama
Documentary
Reality-TV
Talk-Show
Animation
Romance
Family
Action
Crime
Genre des séries
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Proportion de séries en %


Répartition du nombre de films par année
Pourcentage (%)
0
1
2
3
4
5
2000
2005
2010
2015
2020
2025
Année de sortie


Répartition du nombre de séries par année
Pourcentage (%)
0
1
2
3
4
5
6
7
2000
2005
2010
2015
2020
2025
Année de sortie


Comparaison du % de valeurs nulles de la durée en minutes entre Films et Séries
Catégorie
Films
Séries
Pourcentage de valeurs nulles (%)
0
10
20
30
40
50
60
70
movie
tvMovie
tvMiniSeries
tvSeries
Type de contenu


Comparaison du % de valeurs nulles de la note moyenne entre Films et Séries
Type
titleType
Pourcentage de valeurs nulles (%)
0
10
20
30
40
50
60
movie
tvMiniSeries
tvMovie
tvSeries
Type de contenu

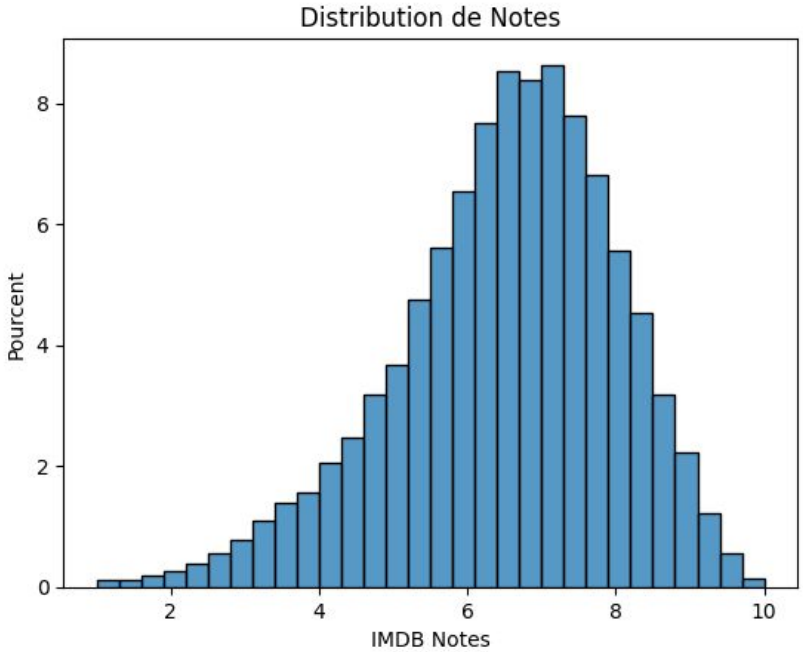

Répartition des notes films et séries confondus
Proportion (%)
0
2
4
6
8
10
12
14
2
4
6
8
10
Note IMDb

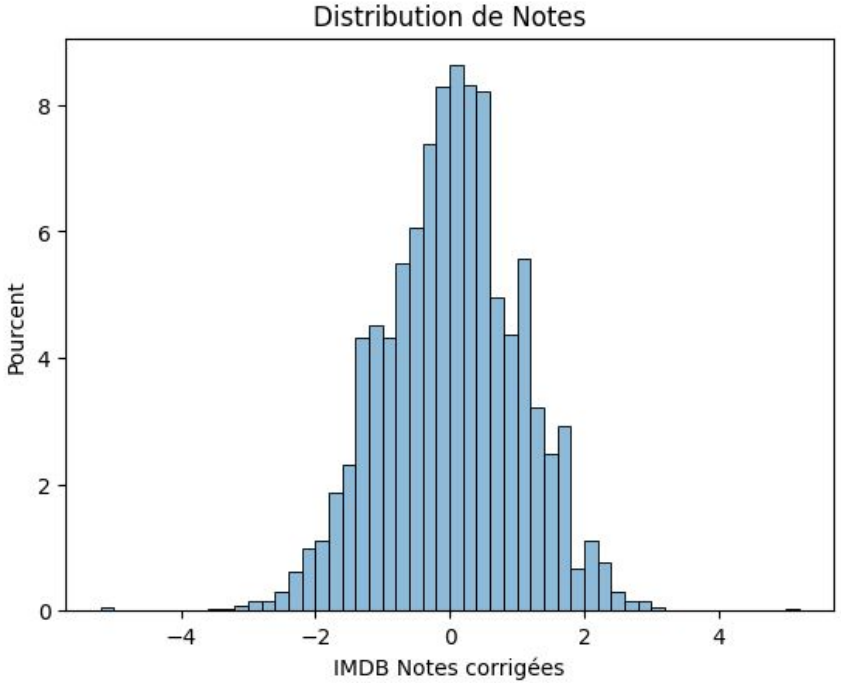
Analyse des votes

Les films	**Les séries**
Moyenne : 3769	Moyenne : 1658
Ecart-type : 36373	Ecart-type : 20677
Q1 : 18	Q1 : 14
Médiane : 60	Médiane : 35
Q3 : 331	Q3: 156

4) L’algorithme de prédiction

Transformation de la cible





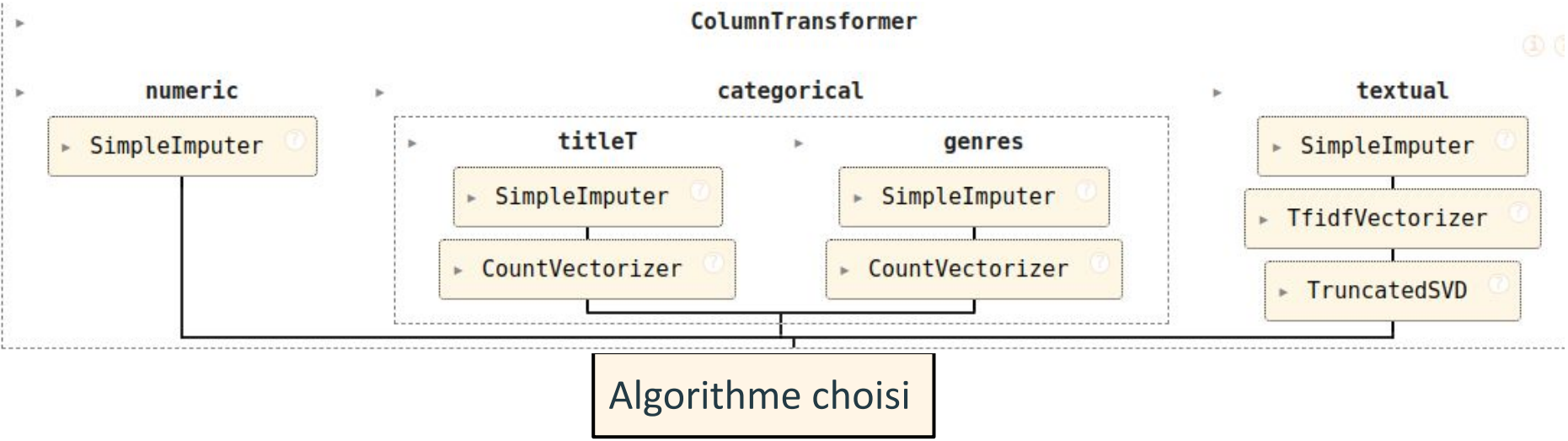
Cible : Notes

4) L’algorithme de prédiction

Les features utilisées :

Numériques :	**Catégorielles :**	**Textuelles :**
startYear	genre_1	actor_1, 2 et 3
	genre_2	director_1, 2 et 3
	genre_3	

La pipeline de regression



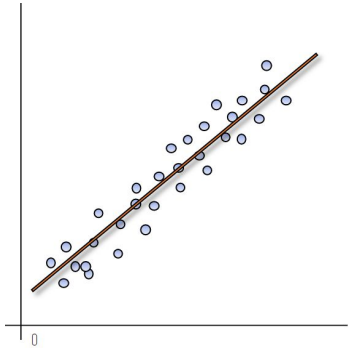
Train/Test split : 70-30

Optimization: Gridsearch + Validation Croisée (k=5)

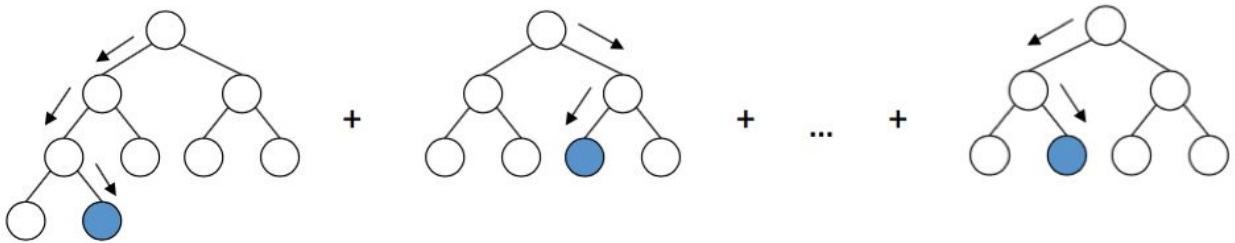
Métriques : R2, RMSE

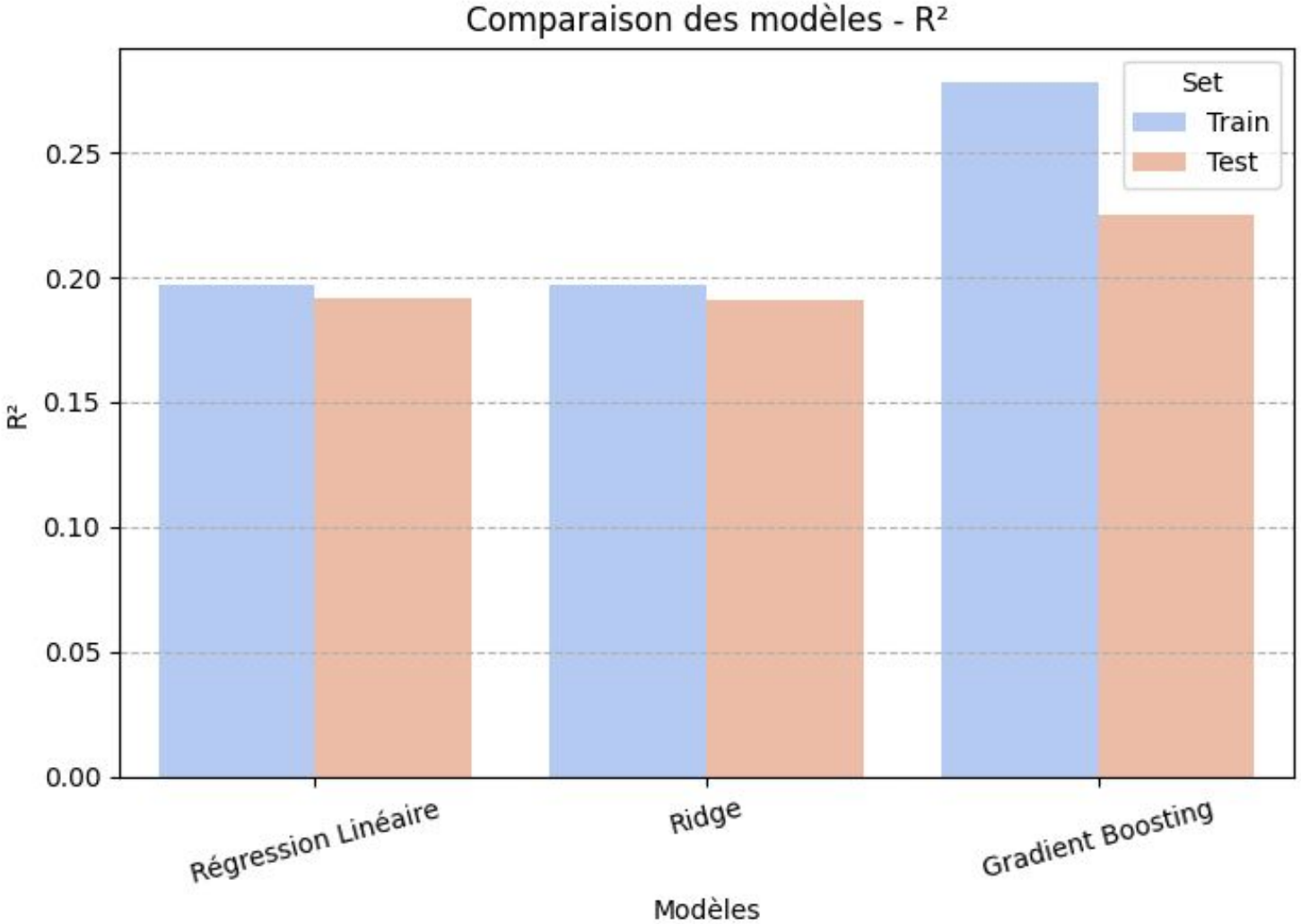
Choix des Modèles

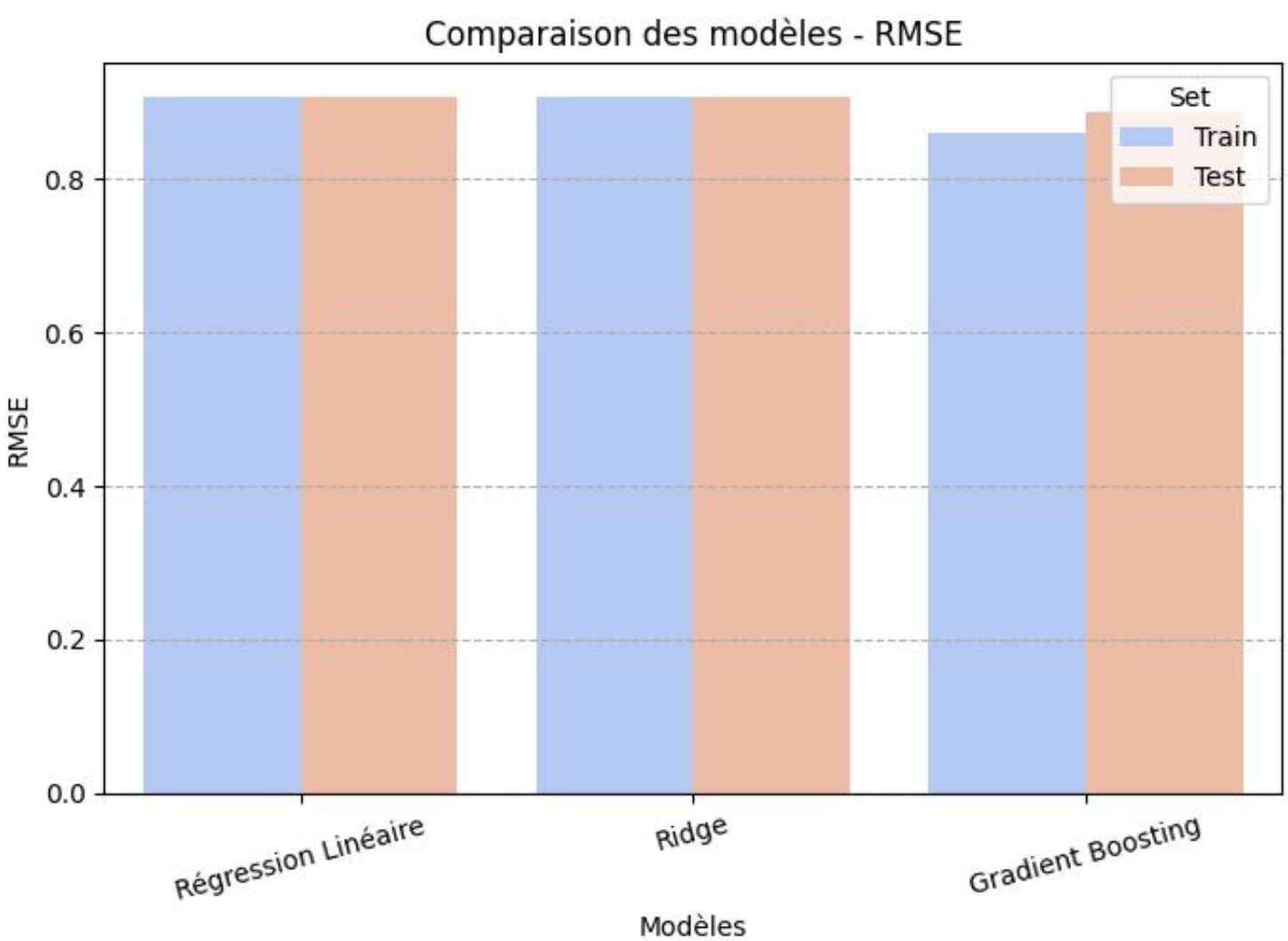
Régression Linéaire/ Ridge



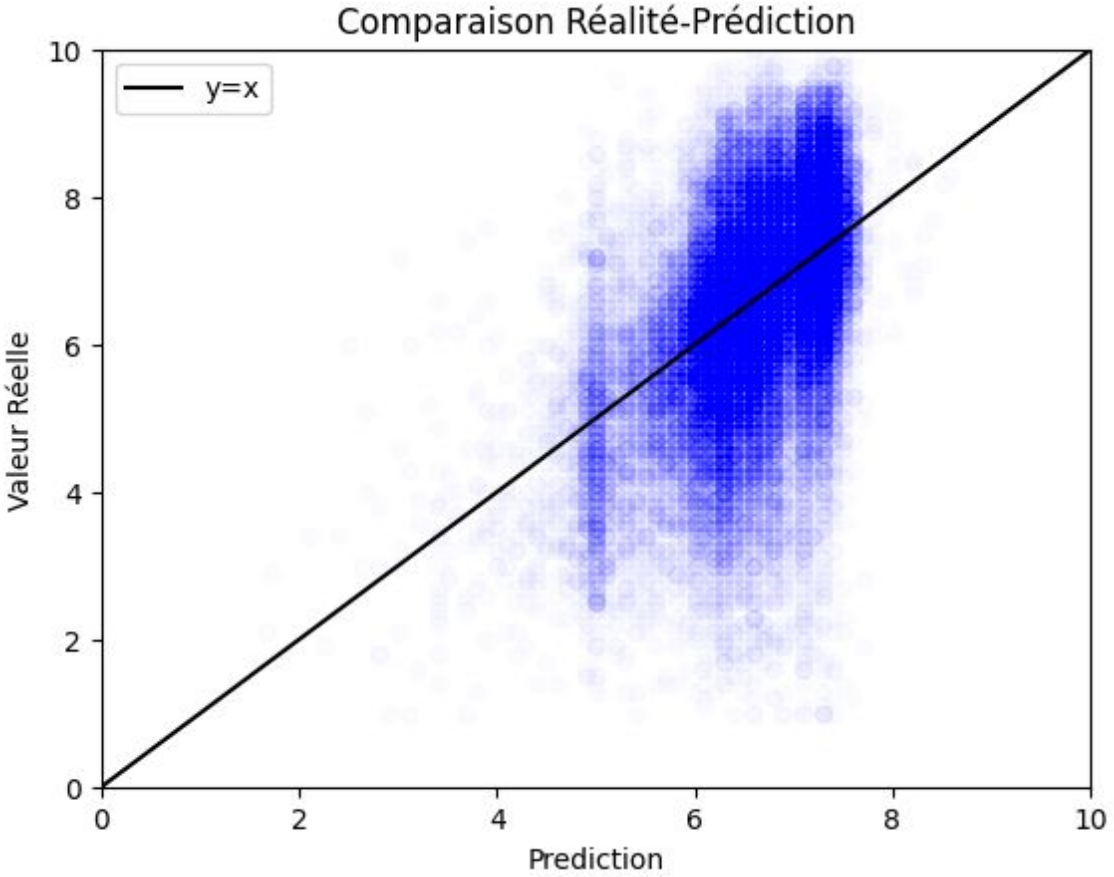
Gradient Boosting




Comparaison des modèles - R²
Set
Train
Test
0.25
0.20
0.15
0.10
0.05
0.00
R²
Régression Linéaire
Ridge
Gradient Boosting
Modèles


Comparaison des modèles - RMSE
Set
Train
Test
RMSE
0.0
0.2
0.4
0.6
0.8
Régression Linéaire
Ridge
Gradient Boosting
Modèles

Comparaison réalité/prédiction du “meilleur” modèle

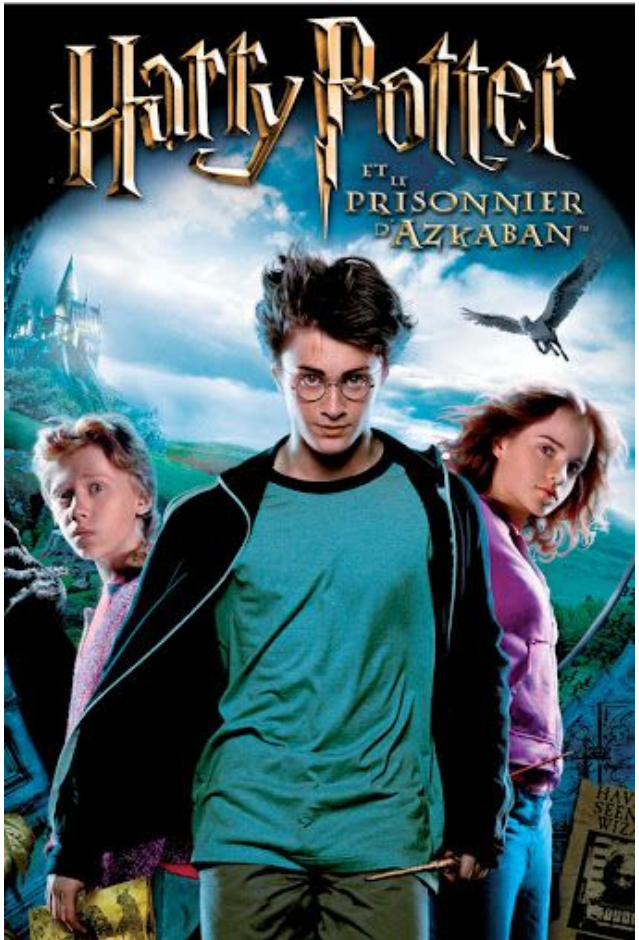


5) L’algorithme de recommandation

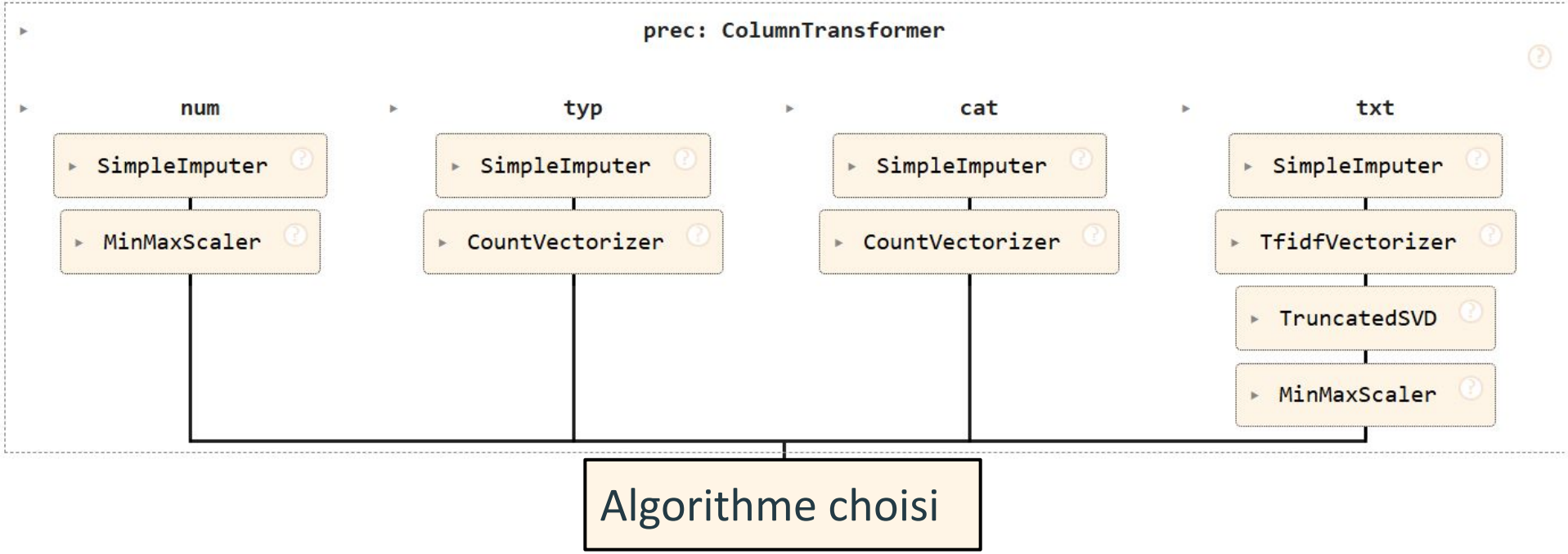
Les features utilisées :

Numériques :	**Catégorielles :**	**Textuelles :**
startYear	genre_1	title
averageRating	genre_2	actor_1, 2 et 3
logVotes	genre_3	director_1, 2 et 3

Les films testés



La pipeline :


prec: ColumnTransformer
num
SimpleImputer
MinMaxScaler
typ
SimpleImputer
CountVectorizer
cat
SimpleImputer
CountVectorizer
txt
SimpleImputer
TfidfVectorizer
TruncatedSVD
MinMaxScaler
Algorithme choisi

Algorithme de cosine similarity / KNN avec Harry Potter et le prisonnier d'Azkaban

startYear	titleType	genre_1	genre_2	genre_3	primaryTitle	averageRating	numVotes
2004	movie	Adventure	Family	Fantasy	Harry Potter and the Prisoner of Azkaban	7.9	719001.0

startYear	**titleType**	**genre_1**	**genre_2**	**genre_3**	**primaryTitle**	**averageRating**	**numVotes**
2018	movie	Comedy	Drama	None3	Jatt vs. Ielts	4.4	121.0
2002	movie	Adventure	Family	Fantasy	Harry Potter and the Chamber of Secrets	7.4	719337.0
2021	movie	Drama	Thriller	War	Chess Story	6.8	5069.0
2010	movie	Adventure	Family	Fantasy	Harry Potter and the Deathly Hallows: Part 1	7.7	622383.0
2011	movie	Adventure	Family	Fantasy	Harry Potter and the Deathly Hallows: Part 2	8.1	982979.0

Quelques incohérences !!

Algorithme de cosine similarity / KNN
avec “Harry Potter et le prisonnier d’Azkaban”

startYear	titleType	genre_1	genre_2	genre_3	primaryTitle	averageRating	numVotes
2004	movie	Adventure	Family	Fantasy	Harry Potter and the Prisoner of Azkaban	7.9	719001.0

startYear	titleType	genre_1	genre_2	genre_3	primaryTitle	averageRating	numVotes
2005	movie	Adventure	Family	Fantasy	Harry Potter and the Goblet of Fire	7.7	707778.0
2002	movie	Adventure	Family	Fantasy	Harry Potter and the Chamber of Secrets	7.4	719337.0
2001	movie	Adventure	Family	Fantasy	Harry Potter and the Sorcerer's Stone	7.7	892057.0
2010	movie	Adventure	Family	Fantasy	Harry Potter and the Deathly Hallows: Part 1	7.7	622383.0
2011	movie	Adventure	Family	Fantasy	Harry Potter and the Deathly Hallows: Part 2	8.1	982979.0

Cela semble plus cohérent !!

Algorithme de cosine similarity / KNN avec “Le robot sauvage”

startYear	titleType	genre_1	genre_2	genre_3	primaryTitle	averageRating	numVotes
2024	movie	Animation	Sci-Fi	None3	The Wild Robot	8.2	118794.0

startYear	titleType	genre_1	genre_2	genre_3	primaryTitle	averageRating	numVotes
2023	movie	Animation	Sci-Fi	None3	The Missing	7.2	318.0
2015	movie	Animation	Sci-Fi	None3	Arpeggio of Blue Steel: Ars Nova - Cadenza	7.2	105.0
2009	movie	Romance	None2	None3	Evaraina Eppudaina	4.8	64.0
2020	movie	Animation	Sci-Fi	None3	The Intruder	8.1	17.0
2021	movie	Animation	Sci-Fi	None3	Bigfoot vs Krampus	2.9	81.0

Quelques incohérences !!

Algorithme de cosine similarity / KNN avec “Le robot sauvage”

startYear	titleType	genre_1	genre_2	genre_3	primaryTitle	averageRating	numVotes
2024	movie	Animation	Sci-Fi	None3	The Wild Robot	8.2	118794.0

startYear	titleType	genre_1	genre_2	genre_3	primaryTitle	averageRating	numVotes
2023	movie	Animation	Sci-Fi	None3	The Missing	7.2	318.0
2015	movie	Animation	Sci-Fi	None3	Arpeggio of Blue Steel: Ars Nova - Cadenza	7.2	105.0
2017	movie	Animation	Sci-Fi	None3	ChäoS;Child: Silent Sky	6.5	68.0
2020	movie	Animation	Sci-Fi	None3	The Intruder	8.1	17.0
2021	movie	Animation	Sci-Fi	None3	Bigfoot vs Krampus	2.9	81.0

C’est un peu mieux…

6) Streamlit

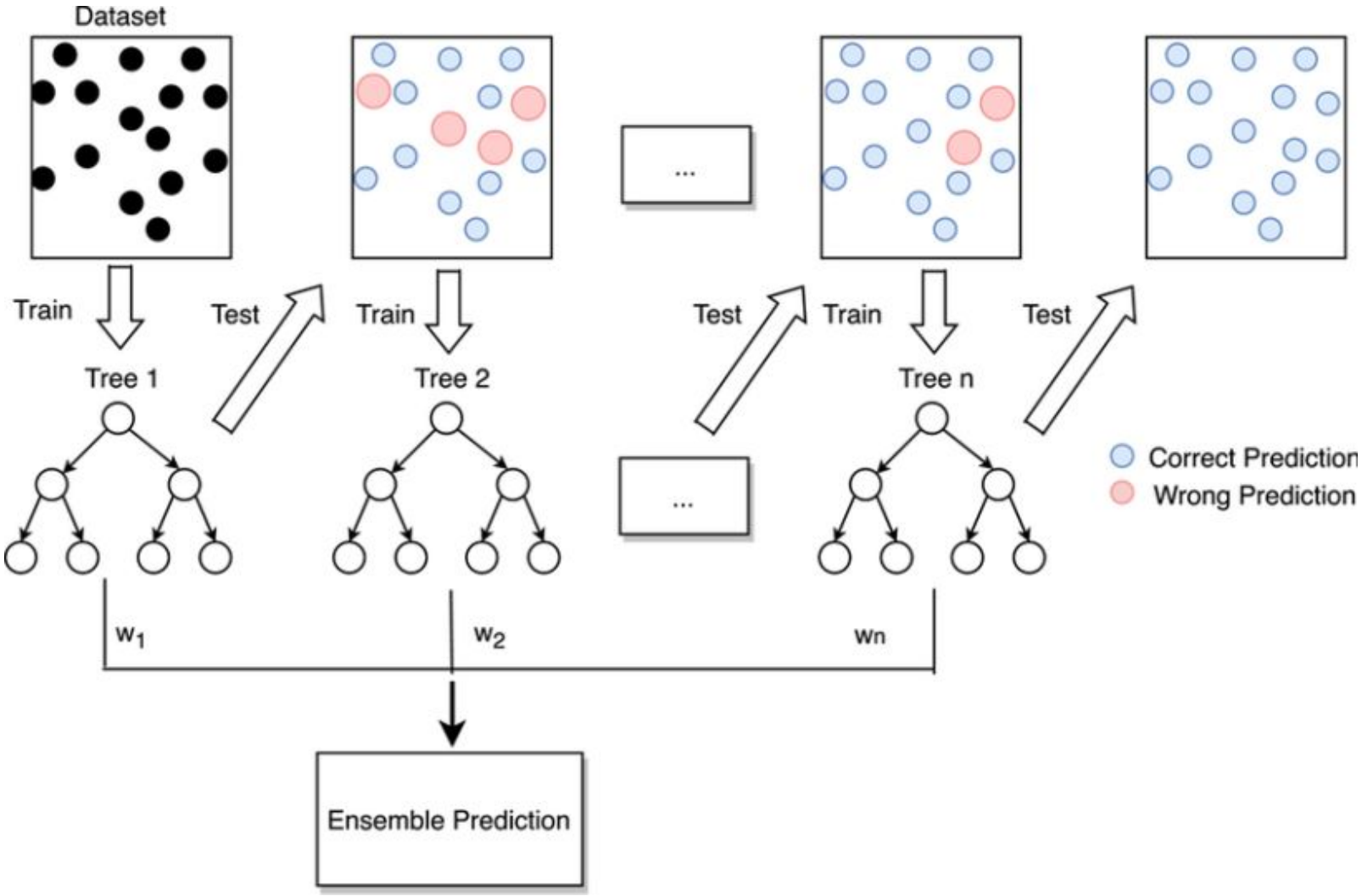
7) Les axes d’améliorations

➔ **Base de données :**
- Intégration en temps réel des mises à jours IMDb

➔ **Système de recommandation :**
- Utiliser les mots-clés d’IMDb afin d’améliorer la recommandation de films
- Intégrer d’autres données comme le temps moyen de visionnage

➔ **Prédiction de la Popularité**
- Prendre en compte l’historique des notes pour chaque acteurs/directeurs
- Prise en compte de l’impact du nombre de vote sur les notes

Merci pour votre attention, avez-vous des questions ?


Dataset
Train
Test
Train
Test
Train
Test
Tree 1
Tree 2
Tree n
...
...
Correct Prediction
Wrong Prediction
w1
w2
wn
Ensemble Prediction