Analyse Predictive pour le Scoring Jeunesse

RAPPORT

Réalisée par : Cheriette Sofiane
Encadrants : Julien Aligon & Moncef Garouani
Institut de Recherche en Informatique de Toulouse (IRIT)
Équipe Systèmes d'Informations Généralisés (SIG)
Mai - Août 2025

TABLE DES MATIÈRES

8. Modélisation et prédiction de la classification d'âge

8.1. Choix des variables explicatives
8.2. Présentation des modèles de machine learning utilisés
8.3. Évaluation des modèles : métriques et exemples concrets
8.4. Résultats améliorés des modèles
8.5. Limites des modèles et déséquilibre des classes

---

9. Bilan personnel

9.1. Compétences techniques développées
9.2. Apports méthodologiques et interdisciplinaires

10. Conclusion

11. Bibliographie

12. Annexes

3. Remerciements.

Je tiens tout d'abord à exprimer ma profonde gratitude à toute l'équipe de l'IRIT, qui m'a accueillie et accompagnée tout au long de ces trois mois de stage enrichissants.

Je remercie tout particulièrement mon encadrant, Monsieur Julien Aligon, ainsi que Monsieur Moncef Garouani, qui a co-encadré mon stage et dont l'expertise, les conseils et l'accompagnement ont été déterminants dans ma réussite personnelle de ce projet.
Leur bienveillance et leur disponibilité ont grandement facilité mon intégration et le bon déroulement de mes missions.

Mes remerciements s'adressent également aux différentes équipes de recherche de l'IRIT pour leur accueil chaleureux et leurs échanges, qui ont contribué à rendre mon expérience au sein du laboratoire encore plus agréable.

Mes pensées vont aussi à Mme Laurence Calendri et Salma El Mallassi du laboratoire IDETCOM, pour leurs échanges et leur collaboration dans le cadre de ce travail.

Enfin, je souhaite remercier M. Alain Berro pour son accompagnement lors de la recherche de stage.

Ce stage m'a permis de développer des compétences techniques et humaines solides, et je suis reconnaissant d'avoir pu contribuer à un projet ambitieux mêlant droit, données, analyse prédictive et éthique, de plus sur un sujet qui me passionne.

Cheriette Sofiane

4. Introduction.

4. 1. Objectifs du rapport.

Le présent rapport a pour objectif de présenter de façon détaillée le travail réalisé durant mon stage au sein de l'IRIT, consistant à développer un ou plusieurs modèles prédictifs de classification d'âge pour les contenus vidéo (films et séries) proposés sur différentes plateformes de SVOD, à partir de données de la plateforme IMDb (il s'agit du point de départ)[1].

Il vise non seulement à exposer le sujet du stage, à détailler les techniques employée et à analyser les résultats obtenus, mais également à étudier les logiques de classification, si elles existent, mises en place par les plateformes, dans le but de mieux comprendre leurs pratiques et les éventuels écarts entre elles.

4. 2. Contexte général du stage.

Ce travail s'inscrit dans une approche pluridisciplinaire, menée en collaboration avec des équipes spécialisées en droit des médias, afin d'éclairer les enjeux réglementaires, commerciaux et légaux liés à la classification des œuvres audiovisuelles.

4. 3. Problématique.

La classification d'âge des contenus audiovisuels relève uniquement de la responsabilité des plateformes de SVOD, ce qui peut entraîne des méthodes variables, parfois controversées d'une plateforme à une autre et surprenantes pour certains spectateurs.

Certaines décisions inadaptées peuvent avoir des conséquences importantes : par exemple, des parents ont réclamé réparation après qu'un film mal classé a traumatisé leur enfant[2], et l'ARCOM a récemment mis en garde une chaîne pour la diffusion de la série "It's a Sin" sans la classification appropriée[3].

Ces cas illustrent le risque lié à l'absence d'un référentiel unifié et amènent à s'interroger sur la manière de garantir une classification conforme aux exigences légales et éthiques, tout en prenant en compte les spécificités de chaque plateforme.

5. Présentation de l'organisme d'accueil.

5.1.L'InstitutdeRechercheenInformatiquedeToulouse (IRIT) .

L’Institut de Recherche en Informatique de Toulouse (IRIT) est une unité mixte de recherche associée au CNRS et rattachée à plusieurs établissements universitaires toulousains, notamment l’Université Toulouse Paul Sabatier, l’INP Toulouse, l’Université Toulouse Capitole et l’Université Toulouse Jean Jaurès. L’IRIT est un centre majeur en France dans le domaine de l’informatique, regroupant environ 700 chercheurs, enseignants-chercheurs, ingénieurs et doctorants.

L’institut couvre un large spectre de disciplines informatiques, allant des algorithmes et théories du calcul aux systèmes distribués, à l’intelligence artificielle, ainsi qu’au traitement du signal et des images. Son activité comprend aussi des recherches appliquées dans des domaines variés, dont la gestion de contenus multimédias.

L’IRIT possède plusieurs plateformes technologiques, notamment Osirim (Occidata), qui offre un environnement homogène pour la recherche d’information dans de larges corpus de contenus multimédias. Cet outil est notamment en lien avec le projet de prédiction de classification des contenus audiovisuels auquel j’ai contribué durant mon stage.

5. 2. L'Institut de Recherche en Informatique de Toulouse (IRIT)

L’équipe SIG (Systèmes d’Informations Généralistes) fait partie des équipes de recherche de l’IRIT et concentre ses travaux dans le domaine de la donnée. Ses recherches couvrent l’ensemble de la chaîne de traitement des données (pipeline), depuis leur collecte brute jusqu’à leur transformation en informations structurées et analysables.

Les principaux axes de l’équipe SIG incluent :

- La conception et le développement d’algorithmes décisionnels, permettant d'augmenter l'efficience de l'exploitabilité les données.
- L’étude des documents et données semi-structurées, avec un intérêt particulier pour les usages variés et les modalités de structuration des données.
- Le traitement de données, qu’elles soient d’origine scientifique, industrielle ou issues de bases de données d’entreprise.

5. 3. La collaboration avec IDETCOM et son rôle dans le projet.

Le laboratoire IDETCOM (Institut du Droit de l'Espace, des Territoires, de la Culture et de la Communication) a joué un rôle fondamental dans le cadre de ce projet. C'est en effet l'IDETCOM qui a identifié la problématique initiale ainsi que le contentieux autour de la classification des contenus sur les plateformes de SVOD. Leur questionnement principal était de déterminer s'il existait une logique cohérente dans la manière dont ces plateformes classifient leurs œuvres, dans une optique d'éventuelles démarches judiciaires à venir.

Au sein de ce laboratoire, nous avons collaboré étroitement avec Mme Laurence Calandri, spécialisée en droit du numérique, ainsi qu'avec sa stagiaire, Salma El Mallassi. Des réunions hebdomadaires étaient organisées, durant lesquelles elles nous présentaient leurs travaux, notamment sur les problématiques liées aux systèmes de plaintes, qu'ils soient gérés par l'ARCOM, le Conseil supérieur de l'audiovisuel (CSA) ou par les plateformes de SVOD elles-mêmes.

Ces échanges réguliers ont permis d'enrichir le travail technique grâce à une compréhension approfondie des enjeux juridiques et réglementaires, favorisant une meilleure construction du jeu de données selon leurs besoins, ainsi qu'une approche pluridisciplinaire essentielle pour appréhender les difficultés liées à la classification par âge des contenus audiovisuels.

6. Présentation du projet de stage.

6. 1. Objectifs globaux du projet.

Le projet vise à explorer la logique de classification d'âge appliquée par les principales plateformes de SVOD. À partir de données publiques, plusieurs modèles d'analyse automatisée ont été développés pour comparer les pratiques de classement des films et séries selon les services étudiés.

6. 2. Enjeux juridiques et techniques liés à la classification des œuvres.

En France, le CNC définit officiellement les tranches d'âge pour les films ("Tous publics", "-12", "-16", "-18"), tandis que l'ARCOM impose des signalétiques pour la télévision ("-10", "-12", "-16", "-18").

Pourtant, sur les plateformes de SVOD, chaque acteur propose sa propre grille, parfois très éloignée des normes nationales (exemple : la catégorie "0+" sur Disney+ n'existe pas dans la réglementation française). Cette situation rend toute comparaison délicate et soulève de véritables enjeux, aussi bien en droit qu'en technologie (la comparaison inter-plateforme devient plus complexe car les éléments de comparaisons diffèrent)[4].

6. 3. Problématique de la prédiction du scoring jeunesse.

Prédire et comparer les classifications d'âge devient complexe lorsque les référentiels sont différents. Les modèles développés ont d'abord pour but de vérifier si une logique, même implicite, existe dans la façon dont chaque plateforme classe ses contenus. Ensuite, le projet vise à utiliser des méthodes d'explication des résultats pour tenter d'expliquer cette logique en confrontant les critères retenus (les variables des jeux de données).

6. 4. Contexte des plateformes de SVOD.

Certaines plateformes telles que Canal+ sont domiciliées en France et soumises au cadre légal national pour la classification des contenus, tandis que d'autres (comme Netflix, Disney+, Prime Video) opèrent depuis l'étranger tout en diffusant largement sur le territoire français.
Les plateformes étrangères sont donc non soumises aux mêmes obligations que celles installées en France, ce qui peut créer des distorsions de concurrence et laisser place à des pratiques parfois moins encadrées.

7. Méthodologie et construction des jeux de données.

7. 1. Sélection de la base IMDb et justification.

IMDb a été retenue comme base principale pour plusieurs raisons. D'abord, grâce à son système de contributeurs communautaires, elle dispose de la plus grande quantité d'informations publiques sur les œuvres audiovisuelles (films, séries et episodes). Cette richesse couvre un large éventail de métadonnées indispensables à l'analyse et à la classification.

La plateforme est largement utilisée dans le domaine informatique, notamment pour développer et évaluer des systèmes de recommandation. Cela atteste de la robustesse et de la fiabilité de son infrastructure technique ainsi que de la qualité de ses données.

Sa popularité auprès de la communauté des développeurs représente également un grand avantage car elle garantit l'existence de nombreux outils et ressources associés (API, fichiers de données, bibliothèques d'accès), facilitant ainsi l'exploitation et l'enrichissement des informations.

Suivant cette argument, le jeu de données utilisé n'a pas été construit intégralement depuis zéro , il provient d'un dépôt existant sur la plateforme Kaggle[1]. Toutefois, les données initiales qui le composent sont bien issues d'IMDb, ce qui conserve tous les avantages liés à cette source.

D'autres alternatives ont été étudiées, comme CineCheck, qui se concentre principalement sur la classification d'âge et propose un nombre beaucoup plus limité de contenus, JustWatch, qui sert avant tout à identifier les diffuseurs d'une œuvre. Ces dernières ne présentent pas la même profondeur d'informations qu'IMDb. Toutefois, JustWatch et d'autres plateformes similaires seront mobilisées ultérieurement lors de la phase d'enrichissement des données.

Le jeu de données utilisé correspond à la base complète d'IMDb, comprenant plus d'un million d'œuvres, au 19 mai 2025.
Pour chaque œuvre les variables suivantes sont présentées:

- id : identifiant unique de l'œuvre sur IMDb, indispensable pour éviter les confusions avec des titres homonymes et pour retrouver précisément un contenu.
- title : titre officiel de l'œuvre.
- type : nature de l'œuvre (film, série ou mini-série ces dernières étant considérées comme des séries dans ce projet).
- genres : liste des genres associés à l'œuvre, information particulièrement pertinente pour inférer la classification d'âge.
- averageRating : note moyenne attribuée par les utilisateurs d'IMDb.
- numVotes : nombre de votes ayant servi à établir la moyenne, indicateur clé de la popularité d'un contenu.
- releaseYear : année de sortie, utile pour contextualiser l'œuvre et éventuellement analyser des tendances temporelles.

7. 2. Critères de sélection des films et séries.

Il a été nécessaire d'établir dès le départ une logique de sélection précise afin de constituer, à partir de la base complète d'IMDb, un ou plusieurs sous-jeux de données pertinents. L'objectif était de disposer d'un corpus suffisamment représentatif et exploitable.

L'approche initiale consistait à sélectionner les 1000 séries et les 1000 films les mieux notés sur IMDb, en se basant uniquement sur leur note moyenne. Cette stratégie semblait pertinente sur le papier, puisqu'elle permettait de travailler sur un échantillon censé correspondre aux œuvres jugées comme les meilleures par les utilisateurs. Cependant, une fois cette première extraction réalisée, deux obstacles majeurs se sont imposés.

Le premier problème concernait la taille même de l'échantillon. Avec seulement 2000 œuvres, il est vite apparu que le volume des données était trop faible dans une perspective de modélisation statistique et d'apprentissage automatique. Comme cela sera montré plus loin dans ce rapport, ce n'était pas la dernière fois que des ajustements furent nécessaires afin d'augmenter la taille du dataset et ainsi tenter d'améliorer la robustesse des analyses

Le second problème, plus important encore, tenait à la nature des œuvres extraites. IMDb étant une plateforme internationale ouverte aux contributions du monde entier, de nombreux titres parmi les mieux notés provenaient de cinématographies très actives comme l'industrie Bollywoodienne ou le marché chinois.
Si ces productions bénéficiaient parfois d'excellentes notes, elles étaient souvent peu ou pas documentées en termes d'informations disponibles sur la plateforme IMDb, ce qui limitait fortement leur intérêt pour une analyse fine. Pire encore, la plupart d'entre elles n'étaient disponibles sur aucune des plateformes de SVOD accessibles depuis la France, réduisant considérablement la proportion réellement exploitable du jeu de départ, qui tombait alors autour de 60-70% seulement.

Face à ces contraintes, il a fallu repenser entièrement la stratégie de sélection. La solution adoptée a été d'augmenter la taille de l'échantillon à 1500 films et 1500 séries, portant ainsi le total des œuvres retenues à 3000. Ce choix permettait mécaniquement d'obtenir plus de données exploitables. Mais la modification la plus déterminante fut d'abandonner le critère de la meilleure note moyenne au profit d'un indicateur de popularité, à savoir le nombre total de votes enregistrés pour une œuvre sur IMDb.
Ce changement augmentait la probabilité de travailler sur des œuvres largement vues, mieux renseignées et donc plus faciles à analyser puisque plus une œuvre est populaire, plus elle bénéficie d'une fiche IMDb riche en informations, grâce à une contribution communautaire plus active.

Cela représente un atout indéniable pour l'étape ultérieure de l'apprentissage automatique, car des données plus complètes permettent non seulement d'affiner la modélisation mais aussi d'améliorer l'interprétation des résultats.

Parallèlement, un filtre supplémentaire a été ajouté au processus de sélection des œuvres afin de ne conserver que les films et séries disposant d'au moins un diffuseur officiel en France. C'est précisément là que la plateforme JustWatch joue un rôle clé, elle recense pour chaque pays les diffuseurs officiels des œuvres ainsi que le lien vers leur page d'informations sur le site de la plateforme SVOD. Cette ressource est essentielle pour la réussite du projet, car la section "streaming" d'IMDb, qui indique les diffuseurs pour les créations originales, s'avère très peu fiable pour les autres contenus et que les liens des oeuvres sur les plateformes de SVOD sont nécessaires afin d'en extraire la classification d'âge.

Au cours de mes recherches, j'ai constaté que JustWatch ne propose pas d'API officielle. Cependant, une API non officielle, développée et maintenue par la communauté informatique, permet de rechercher les diffuseurs d'une œuvre dans un pays donné[5]. Un des grands défis ici est que les titres des œuvres peuvent être homonymes, et que les titres officiels peuvent varier d'un pays à l'autre, ce qui rend la recherche par titre seule très imprécise.

Deux options s'offraient à moi, soit d'utiliser la section streaming d'IMDb, ce qui était peu satisfaisant pour les raisons évoquées précédemment, soit d'étudier et adapter l'API communautaire afin de réaliser les recherches à partir de l'identifiant IMDb unique, plus fiable que le titre. J'ai opté pour cette seconde solution.
Néanmoins, un autre problème est apparu, cette API n'est pas basée sur une base de données classique, mais sur un algorithme de web scraping qui fonctionne comme une barre de recherche, ici celle de la plateforme JustWatch.

J'ai donc dû effectuer les recherches par titre puisque l'API ne proposait pas d'autres possibilités, puis vérifier pour chaque œuvre proposée si l'identifiant IMDb retourné correspondait à celui de mon jeu de données. Lorsque les identifiants correspondaient, je pouvais alors intégrer les liens vers les plateformes SVOD pour chaque œuvre dans mon jeu de données.

D'autres données issues de cette API ont également été ajoutées, qui seront détaillées plus tard dans la section 7.4 consacrée au nettoyage, traitement et enrichissement des données.

7. 3. Techniques de web scraping et outils utilisés.

Cette sous-partie vise à présenter les outils et méthodes mis en œuvre pour la collecte des données. Elle constitue une introduction à la section suivante, en offrant les éléments nécessaires pour comprendre les raisons qui ont motivé le choix de certains outils.

Le web scraping désigne l'ensemble des techniques qui permettent d'extraire automatiquement des données à partir de pages internet. Cette méthode a été essentielle durant mon projet, car elle m'a permis d'enrichir mon jeu de données, notamment en récupérant les informations sur la classification d'âge directement auprès de chaque plateforme.

Concrètement, des algorithmes de web scraping adaptés ont été conçus pour extraire les classifications d'âge des plateformes Prime Video, Canal+, Disney+ et Netflix. Pour chacune, il a fallu s'adapter à la structure spécifique du site et aux éléments affichés à l'écran, même si la logique générale restait identique.

Initialement, la collecte des classifications sur Paramount+ et Apple TV+ nécessitait une démarche légèrement différente. Nous verrons plus tard que, pour ces deux plateformes, il a fallu recourir à un web scraping traditionnel. En effet, l'information concernant la classification d'âge transite par des requêtes techniques du navigateur (XHR).
Les outils développés ont donc été conçus pour interroger directement les serveurs de ces plateformes et récupérer les données à la source, sans passer par l'extraction du contenu visible sur la page web, car cette méthode s'avère légèrement plus simple à coder.

Au final, même si les processus employés pour Paramount+, AppleTV+ et les autres plateformes sont assez proches, ils reposent moins sur le web scraping traditionnel et davantage sur la récupération automatisée de données via des requêtes serveur (ce n'est pas du web scraping "classique").

Plusieurs autres outils informatiques ont été mobilisés pour collecter, enrichir et traiter les données. Le premier outil que j'ai découvert au cours de mes recherches sur les jeux de données et les solutions d'agrégation est IMDbPY[6].
Il s'agit d'une bibliothèque Python open source permettant d'interroger la base IMDb et de récupérer des métadonnées détaillées sur les films et séries. Bien que cette API ne soit pas officielle, elle présente l'avantage d'être totalement gratuite et de disposer d'une documentation solide. L'accès aux API officielles proposées par IMDb est payant, parfois à des tarifs très élevés, et n'est pas réellement nécessaire, car la grande majorité des informations qu'elles fournissent peuvent être obtenues par d'autres moyens (notamment via web scraping).

J'ai également utilisé l'API de TMDb (The Movie Database)[7], une plateforme similaire à IMDb, qui se distingue par la qualité et la richesse de ses données, notamment au niveau de la localisation linguistique. Contrairement à IMDbPY, TMDb propose pour chaque série télévisée les noms des épisodes en français, ce qui représente un avantage fondamental si l'objectif est de réaliser une analyse textuelle adaptée. Cette possibilité pourrait permettre d'optimiser les modèles d'analyse prédictif.

TMDb propose une meilleure organisation sur la gestion des pays de production, la plateforme référence principalement les pays ayant participé à la production d'une œuvre, là où IMDb mélange parfois les pays de production, les lieux de tournage, voire inclut des informations incorrectes ou incohérentes. Par exemple, la série "Totally Spies" se voit attribuer la Corée du Nord comme pays de production sur IMDb, alors qu'il s'agit d'une erreur manifestement liée à un mauvais référencement des lieux de tournage.

Il est important de mentionner que l'API IMDbPY repose sur l'utilisation de la librairie Selenium pour collecter les informations. Cette particularité rend l'outil peu optimisé en termes de performances (l'objectif n'est pas de présenter en la librairie Selenium mais pour la faire courte, elle simule une utilisation humaine d'un navigateur web) l'exécution des algorithmes pour traiter 1 500 œuvres pouvait dépasser dix heures. Rapidement, il est apparu plus efficace de privilégier le web scraping direct lorsque des informations supplémentaires sur la plateforme IMDb étaient nécessaires. J'aborderai plus en détail ce choix et ses conséquences dans la section 7.4 consacrée à l'enrichissement du jeu de données.

Par ailleurs, l'API de Rotten Tomatoes a aussi été étudiée mais finalement écartée. En effet, cette plateforme s'est révélée beaucoup moins riche et exhaustive qu'IMDb, TMDb ou JustWatch, tant en termes de données disponibles que de diversité des informations.

Il est imortant de souligner que le passage d'une API à une autre pour la collecte de données s'est grandement facilité par l'utilisation des identifiants uniques des séries et films (IDs IMDb ou JustWatch). Ces identifiants ont permis la réalisation de jointures précises et fiables entre différentes sources permettant de correspondre parfaitement les œuvres et d'éviter les problèmes d'homonymie ou de changements de titres selon les pays et les plateformes. Cette approche aurait été bien moins efficace s'il avait fallu croiser les données uniquement à partir des titres des œuvres puisque cela aurait introduit de nombreuses ambiguïtés.

7. 4. Nettoyage, traitement et enrichissement des données.

L'objectif principal de cette partie est de présenter en détail les différentes variables incluses dans le jeu de données, et d'expliquer la logique intuitive qui a guidé leur sélection. Tout au long du rapport, nous utiliserons le terme "variables" pour désigner chacun des champs présents dans la base, afin de simplifier la lecture et l'analyse.

La sélection de ces variables ne s'est pas faite au hasard mais résulte d'une réflexion menée en amont pour optimiser au maximum l'efficacité des futurs modèles prédictifs. Certaines variables ont été choisies parce qu'elles constituaient des indicateurs qui étaient intuitivement déterminants dans la prédiction de la classification d'âge, d'autres (moins directement utiles pour l'analyse prédictive) jouent un rôle essentiel dans l'ajout ou le croisement de nouvelles informations et facilitent l'enrichissement du jeu de données.

Dans les paragraphes suivants, seront ainsi détaillées les différentes étapes de nettoyage, de traitement et d'enrichissement ayant permis de construire un jeu de données fiable, adapté à la modélisation et aisément extensible pour les besoins futurs du projet.

Certaines variables ont été ajoutées en amont au jeu de données initial tandis que d'autres ont été ajoutées progressivement dans le but d'améliorer les modèles prédictifs déjà existants au moment de l'intégration de ces variables. Ces ajouts successifs visaient à perfectionner ces modèles et à accroître leur performance.

La variable "idjw" correspond à l'identifiant unique d'une œuvre sur la plateforme JustWatch. Elle est obtenue par le procédé détaillé en section 7.2, qui consiste à effectuer une recherche par titre sur JustWatch puis à valider la correspondance avec l'identifiant IMDb de l'œuvre. Ce double contrôle garantit une identification précise et évite toute confusion, notamment en cas d'homonymie.

Chaque identifiant JustWatch est unique, il permet donc de retrouver facilement toutes les informations liées à une œuvre via l'API JustWatch présentée précédemment. L'intégration de cette variable répond à deux motivations principales. Tout d'abord, l'objectif initial était de récupérer les liens directs des œuvres sur les plateformes de SVOD afin d'en extraire la classification d'âge. Bien que cela puisse sembler superflu (puisqu'on pourrait se contenter d'extraire ces liens uniquement) mon expérience acquise au fil du projet m'a montré qu'il est préférable d'anticiper les besoins futurs. Inclure l'identifiant JustWatch laisse toujours une marge pour enrichir le jeu de données avec d'autres informations extraites ultérieurement grâce à cette API. Un accès facilité par identifiant est donc un véritable atout pour la flexibilité de l'enrichissement.

Par ailleurs, la présence de "idjw" offre une grande aisance pour réaliser des jointures entre différentes sources d'information, si l'on possède l'identifiant IMDb d'une œuvre et que l'on souhaite récupérer des données disponibles sur JustWatch, le lien direct entre les identifiants uniques assure la fiabilité et la simplicité de ces rapprochements.
Ainsi, l'association systématique de chaque identifiant IMDb à son identifiant JustWatch dans les jeux de données constitue un choix structurant pour tout travail de croisement ou d'enrichissement à venir.

La variable "NbDiff" indique le nombre de diffuseurs auprès desquels une œuvre est disponible en France, (Rappel :il s'agit içi d'offre de vidéo à la demande par abonnement (VOD), leses options d'achat ou de location sont exclues).

D'un point de vue technique, il s'agit du nombre de fois qu'une même œuvre apparaît dans le jeu de données. Cette variable a principalement une vocation de contrôle qualité, elle sert de moyen de vérification pour limiter les risques d'erreurs lors des opérations de traitement ou de modification du jeu de données. Par exemple, elle permet de s'assurer qu'un film ou une série n'a pas été dupliqué à tort, ou qu'aucun diffuseur n'a été omis pendant le processus d'enrichissement.

Il n'y a pas d'intuition particulière concernant une éventuelle corrélation entre le nombre de diffuseurs et la classification d'âge attribuée à une œuvre, rien n'indique que plus une œuvre est diffusée, plus sa classification serait susceptible d'être restreinte. L'utilité de "NbDiff" réside donc essentiellement dans sa fonction de contrôle.

La variable "ClassifImdb" correspond à la classification d'âge attribuée en France à une œuvre sur la plateforme IMDb. Cette information résulte très souvent des contributions directes des utilisateurs et de la communauté d'IMDb. Cette information présente donc un intérêt particulier puiqu'elle reflète alors la perception et la sensibilité réelles du public vis-à-vis du contenu, ce qui peut constituer un indicateur intéressant à des fins prédictives.

Pour les films, le processus d'attribution de la classification est différent (ce que nous détaillerons pour les variables liées au CNC).
Un des principaux inconvénients de "ClassifImdb" est qu'elle n'est pas systématiquement renseignée. Il arrive fréquemment que certaines œuvres n'aient pas de classification attribuée pour la France, ce qui se traduit alors par la valeur "not rated" dans la variable correspondante.

Elle est initialement obtenue à partir de l'API IMDb mais rapidement remplacée par du web scraping classique pour des raisons d'optimisations (précédemment évoquées)

La variable “Diffuseur” correspond au lien direct vers la page de l’œuvre sur la plateforme de SVOD concernée. Cette variable est fondamentale, c’est grâce à elle que l’on peut accéder à la classification d’âge affectée par la plateforme à chaque œuvre, étape essentielle dans le cadre du projet. La présence du lien facilite également la récupération d’autres informations spécifiques à la plateforme lors d’éventuels enrichissements ultérieurs du jeu de données.
Elle est obtenue via l’API JustWatch, à partir de son identifiant sur la même plateforme.

La variable suivante, "SVOD", indique le nom du service de diffusion (Netflix, Prime Video, TF1+, etc.) chez lequel l’œuvre est disponible. Ce champ est essentiel pour la structuration et l’analyse du jeu de données, il permet de segmenter facilement les œuvres selon leur diffuseur, de comparer les pratiques de classification entre plateformes, et d’effectuer des analyses spécifiques à chaque service. Elle est obtenue simplement à partir de la variable “Diffuseur” (lien vers la plateforme de SVOD)

La variable "ClassificationSVOD" correspond à la classification d’âge attribuée par la plateforme SVOD (désignée dans la variable précédente) à une œuvre donnée. Cette information est centrale dans le projet puisqu'elle constitue le label cible des analyses prédictives, notamment dans la mise en œuvre d’algorithmes d’apprentissage supervisé visant à modéliser ou anticiper la logique de classification des différentes plateformes. D'un point de vue juridique, disposer d’une classification précise et fiable facilite la comparaison entre services, et permet de mieux comprendre les pratiques de protection des publics appliquées par chaque diffuseur.
Elle est obtenue à partir de web scraping sur les sites des plateformes de SVOD.

Toutes les variables présentées jusqu’ici ont été incluses de manière intuitive, en considérant leur utilité potentielle soit pour faciliter l’enrichissement et la modification du jeu de données, soit pour apporter une première compréhension lors des analyses prédictives.

La variable “Diffuseur” correspond au lien direct vers la page de l’œuvre sur la plateforme de SVOD concernée. Cette variable est fondamentale, c’est grâce à elle que l’on peut accéder à la classification d’âge affectée par la plateforme à chaque œuvre, étape essentielle dans le cadre du projet. La présence du lien facilite également la récupération d’autres informations spécifiques à la plateforme lors d’éventuels enrichissements ultérieurs du jeu de données.
Elle est obtenue via l’API JustWatch, à partir de son identifiant sur la même plateforme.

La variable suivante, "SVOD", indique le nom du service de diffusion (Netflix, Prime Video, TF1+, etc.) chez lequel l’œuvre est disponible. Ce champ est essentiel pour la structuration et l’analyse du jeu de données, il permet de segmenter facilement les œuvres selon leur diffuseur, de comparer les pratiques de classification entre plateformes, et d’effectuer des analyses spécifiques à chaque service. Elle est obtenue simplement à partir de la variable “Diffuseur” (lien vers la plateforme de SVOD)

La variable "ClassificationSVOD" correspond à la classification d’âge attribuée par la plateforme SVOD (désignée dans la variable précédente) à une œuvre donnée. Cette information est centrale dans le projet puisqu'elle constitue le label cible des analyses prédictives, notamment dans la mise en œuvre d’algorithmes d’apprentissage supervisé visant à modéliser ou anticiper la logique de classification des différentes plateformes. D'un point de vue juridique, disposer d’une classification précise et fiable facilite la comparaison entre services, et permet de mieux comprendre les pratiques de protection des publics appliquées par chaque diffuseur.
Elle est obtenue à partir de web scraping sur les sites des plateformes de SVOD.

Toutes les variables présentées jusqu’ici ont été incluses de manière intuitive, en considérant leur utilité potentielle soit pour faciliter l’enrichissement et la modification du jeu de données, soit pour apporter une première compréhension lors des analyses prédictives.

Nous allons maintenant aborder les autres variables qui ont été ajoutées progressivement au fur et à mesure de l'avancement du projet (ces variables seront détaillées dans la suite, tandis que les développements des modèles eux-mêmes seront présentés plus en détail dans la section 8 dédiée à la modélisation et à la prédiction des classifications d'âge).

Les variables "visaCNC" et "classifCNC" concernent uniquement le jeu de données des films. Lors de nos échanges avec Mme Laurence Calandri du laboratoire IDETCOM, cette dernière a suggéré qu'il serait intéressant d'inclure dans le jeu de données les classifications d'âge émises par le CNC pour les films ayant été diffusés en salle. Cette inclusion paraît particulièrement pertinente pour analyser si les recommandations officielles du CNC vis-à-vis de la diffusion en salle sont respectées par les plateformes de SVOD. En termes plus techniques, il s'agit de voir si les classifications d'âge prédites par les modèles sont largement influencées par la variable "classifCNC".

L'ajout de ces variables a été probablement la tâche la plus fastidieuse de la partie d'enrichissement des données, car elles sont spécifiques au contexte français et donc pas référencées sur les sites tels qu'IMDb ou JustWatch, ce qui complique grandement l'automatisation de leur extraction.

Le site officiel du CNC, utilisé pour la recherche des visas, est très mal conçu du point de vue de la recherche, la plupart des titres originaux, souvent en anglais, ne sont pas reconnus correctement. De plus, le système de recherche souffre d'une reconnaissance parfois arbitraire des articles (grammaticalement parlant), et il est impossible de faire des recherches directement à partir des identifiants IMDb ou JustWatch. Par conséquent, la recherche doit obligatoirement se faire par titre, ce qui est loin d'être optimal en raison des homonymies et du fait que les titres sur la plateforme CNC sont en français, tandis que ceux du jeu de données sont en version or iginale.

Au fil de mes recherches j'ai identifié une alternative via la plateforme AlloCiné, qui référence parfois (mais pas systématiquement) les visas CNC, d'où la présence distincte de la variable "visaCNC" dans le jeu de données. Contrairement au site du CNC, la barre de recherche d'AlloCiné est beaucoup plus robuste, elle gère sans problème les titres originaux et les articles présents ou non dans les titres des films. Cette alternative m'a permis de renseigner une majorité des classifications CNC manquantes grâce à du web scraping, avec une personnalisation des URLs pour la recherche par titre et une vérification systématique par année de sortie afin de limiter les erreurs liées aux homonymes.

Cependant il restait une partie non négligeable de films pour lesquels les informations n'avaient pas pu être trouvées via AlloCiné. J'ai donc dû revenir au site du CNC pour tenter de compléter les données, toujours en effectuant une double vérification basée sur l'année de sortie, également via web scraping. Malgré ces efforts, une centaine de films restaient encore sans classification CNC dans le jeu de données. Pour ces cas, une intervention manuelle a été nécessaire. Ces films n'avaient pas été renseignés soit parce qu'ils n'étaient pas sortis en salles, et donc ne disposaient pas de visa CNC (notamment les créations originales Netflix), soit parce que la page web correspondante ne se chargeait tout simplement pas sur le site CNC. Pour ces derniers, j'ai consulté la page de résultats de recherche Google pour trouver les informations manquantes et compléter manuellement le jeu de données.

En réalité pour la grande majorité des œuvres de mon jeu de donnée, les classifications CNC correspondent aux classifications présentes sur IMDb (variable "ClassifImdb). Cette concordance m'a permis de mettre en place une vérification supplémentaire quant à l'exactitude des informations issues du CNC, car je pouvais ainsi garantir avec une certitude absolue que la variable "ClassifImdb" ne comportait aucune valeur erronée, étant donné que sa source principale est l'identifiant IMDb des œuvres, extrait directement du jeu de données source d'IMDb.
Néanmoins cette similitude implique également que l'utilité de la variable "classifCNC" reste marginale dans un cadre d'analyse prédictive, sauf pour quelques rares œuvres (négligable donc à l'échelle du jeu de donnée entier), elle constitue essentiellement un doublon de la classification IMDb, sans apporter d'information supplémentaire significative.

Les variables "ValNudite", "ScoreNudite", "ValViolence", "ScoreViolence", "ValVulgarite", "ScoreVulgarite", "ValStupefiants", "ScoreStupefiants", "ValIntensite" et "ScoreIntensite" proviennent de la section « guide parental » des œuvres sur la plateforme IMDb. Ces données ont été extraites assez simplement grâce à un web scraping classique automatisé en s'appuyant sur les identifiants IMDb de chaque œuvre.
Comme souvent avec IMDb ces informations sont issues des contributions des utilisateurs eux-mêmes, offrant ainsi une vision plus pragmatique, plus réaliste des perceptions des spectateurs concernant les thématiques sensibles.

D'un point de vue métier et juridique il aurait été particulièrement intéressant d'obtenir des informations supplémentaires sur le profil des contributeurs ayant voté, notamment leur âge, sexe, pays d'origine, etc. Ces données auraient en effet permis d'affiner la compréhension contextuelle des avis recueillis, en appréciant par exemple la diversité des sensibilités selon les groupes démographiques, ou en évaluant la représentativité des votes en fonction de critères socioculturels.

Les variables de score ("ScoreX") sont toutes de nature catégorielle avec quatre niveaux possibles : Aucun, Faible, Modéré et Élevé.
Elles reflètent l'intensité perçue de chaque thématique (Scènes à caractère sexuel/nudité, Violence et horreur, Propos injurieux ou vulgaires, Consommation d'alcool, de drogues et tabagisme, Scènes effrayantes et de grande intensité) telle qu'évaluée par les contributeurs. Quant aux variables ValX, elles correspondent à un pourcentage exprimant le degré d'accord des utilisateurs concernant l'attribution d'un score particulier à la variable associée.
Techniquement, ce pourcentage est calculé comme le ratio entre le nombre de votes pour la valeur donnée et le nombre total de votes exprimés sur cette thématique. Plus cette valeur est élevée, plus il est probable que le score associé soit correct d'un point de vue objectif.

Passons maintenant aux variables relatives au genre des œuvres, mais cette fois-ci sous une forme encodée.
Les modèles d'apprentissage automatique n'acceptant que des données numériques il est indispensable de transformer les variables qualitatives en valeurs numériques exploitables.

Pour les variables catégorielles simples comme les variables "ScoreX" (qui ne prennent que quatre valeurs distinctes), l'encodage n'a pas été réalisé avec OneHotEncoder car il suffit de remplacer chaque catégorie par un chiffre associé. Cela permet d'indexer chaque modalité par une valeur numérique.

En revanche pour les variables liées aux genres un encodage plus systématique a été mis en œuvre avec l'outil OneHotEncoder.
Cette technique consiste à créer pour chaque genre une variable distincte prenant la valeur 1 si ce genre est associé à une œuvre donnée et 0 sinon. Cette représentation binaire multiple est une pratique courante et bien établie en ingénierie des données, cela permet de préserver l'information catégorielle en rendant les variables compatibles avec les algorithmes d'apprentissage supervisé.

Dans la même logique d'encodage des variables qualitatives en données numériques exploitables par les modèles d'apprentissage, nous avons introduit les variables "CP_X", correspondant aux codes des pays de production des œuvres ("CP_us" pour les États-Unis, "CP_fr" pour la France, etc..). Chaque variable prend la valeur 1 si l'œuvre a été produite par le pays correspondant, et 0 sinon.

Les informations sur les pays de production ont été obtenues via l'API TMDb, à laquelle nous avons accédé grâce à une jointure effectuée depuis l'API JustWatch qui référence l'identifiant TMDb des œuvres.

La variable "Duree" (pour le jeu de données des films) correspond à la durée de chaque film en minutes.
La variable "DureeEp" est son équivalent pour le jeu de données des séries, elle représente la durée moyenne en minutes des épisodes d'une série.
Ces informations ont été obtenues via l'API JustWatch.

La variable “classifJW” correspond à la classification d’âge attribuée à une œuvre sur la plateforme JustWatch. À ce stade du projet, mes modèles semblaient avoir atteint leurs limites techniques, mais je souhaitais néanmoins explorer la possibilité de les améliorer en intégrant de nouvelles variables susceptibles d’en accroître la précision. L’ajout de la classification JustWatch faisait partie de ces pistes, dans l’optique de tirer parti d’une autre source de données de référence, basée sur les pratiques et règles propres à cette plateforme.
Cette information a été obtenue via l’API JustWatch.

La dernière variable, “ScoreParental”, est distincte des variables de contenu parental présentes sur IMDb. Il s’agit d’un score sur 100 représentant un indice composite lié à la violence, au sexe et à la drogue, calculé à partir d’une liste de mots-clés associées à chaque œuvre depuis la plateforme IMDb.

Mon idée initiale était de réaliser une analyse de texte sur les mots-clés d’une œuvre (depuis la section “mots-clés de l’intrigue” sur IMDb).
Au départ j’envisageais d’utiliser des modèles pré-entraînés classiques mais ceux-ci étaient soit principalement adaptés à l’analyse de texte complet et non à une simple liste de mots-clés soit non-performants après tests.
Je me suis donc rabattu sur l’API proposée par MistralAI que j’ai configurée afin de retourner un score parental associé à une liste de mots-clés spécifique.

7. 5. Structuration par plateforme SVOD.

Nous savons qu'il est essentiel de segmenter les analyses prédictives par plateforme de SVOD car chaque service utilise ses propres critères de classification.
Créer un jeu de données distinct pour chaque association SVOD/type d’œuvre aurait rendu les tâches d’ingénierie des données inutilement complexes.

Il a alors été choisi de dupliquer chaque œuvre autant de fois qu’elle dispose de diffuseurs différents, en adaptant à chaque duplication le lien vers la plateforme SVOD, le nom du SVOD et la classification d’âge propre à cette plateforme. Cette solution permet de conserver la souplesse d’un jeu de donnée unique pour chaque type d’œuvre.

8. Modélisation et prédiction de la classification d'âge.

8. 1. Choix des variables explicatives.

Dans la phase de modélisation de la classification d'âge certaines variables nominatives ou non numériques ne sont pas utilisées comme variables explicatives.
Cela concerne les identifiants IMDb, JustWatch et CNC ("visaCNC" pour les films) mais aussi le type d'œuvre (film ou série, qui reste constant au sein d'un même jeu de données), le titre, le lien de l'œuvre vers la plateforme de SVOD (variable "Diffuseur") ainsi que le nom du SVOD (la sélection du diffuseur ayant été effectuée en amont de l'analyse).
Ces variables servent surtout aux tâches d'ingénierie des données et au filtrage préalable. Elles n'apportent pas d'information utile à la prédiction de la classification d'âge.

8. 2. Présentation des modèles de machine learning utilisés.

Pour aborder la prédiction de la classification d'âge, plusieurs modèles de machine learning ont été testés de façon comparative.
Tout d'abord, le modèle Random Forest, il s'appuie sur une multitude d'arbres de décision et effectue une agrégation des prédictions, ce qui lui confère une bonne robustesse mais il nécessite en contre-partie de grands échantillons de données.

Ensuite deux autres modèles de gradient boosting ont été intégrés à l'analyse, il s'agit de CatBoost et XGBoost[8].
Ces algorithmes sont réputés pour leur efficacité et leur capacité à traiter les variables catégorielles ou codées (majorité des variables de mes jeux de données), tout en maîtrisant le sur-apprentissage, notamment dans le cadre de jeux de données avec des structures complexes et peu volumineux.

Enfin, le modèle SVM (Support Vector Machine) a également été testé.
Celui-ci cherche à séparer les données dans l'espace par des frontières optimales, mais il s'est rapidement avéré inadapté au problème puisque les résultats obtenus étaient médiocres.
Le SVM a donc été très vite écarté au profit des autres modèles plus performants.

8. 3. Évaluation des modèles: métriques et exemples concrets.

Nous présentons ici quelques modèles initiaux, sans optimisation des hyperparamètres, ni ingénierie des variables, ni intégration des nouvelles variables issues du nettoyage et enrichissement présentés précédemment. Ces modèles permettent d'illustrer l'évolution globale au cours du projet, en montrant des exemples sans analyse détaillée pour chaque SVOD ou type d'œuvre.

Pour le premier modèle test sur les séries Netflix (M1) avec Random Forest, l'accuracy obtenu est de 74,6%.
En comparaison un modèle aléatoire, basé uniquement sur le hasard, aurait une précision d'environ 16,7% (soit 100/6, 6 étant le nombre de classes disponibles pour Netflix).
Ce résultat initial peut donc paraître satisfaisant.

Cependant en analysant la précision par classe, on observe que si les classifications 13+ et 16+ sont bien prédites, la précision est nulle pour les classes 18+ et Tous publics (la matrice de confusion le confirme). Ce déséquilibre pose un problème de généralisation du modèle, dû au fait que ces classes minoritaires ne disposent pratiquement d'aucun support d'apprentissage (support = 0).

Trois voies sont envisageables pour résoudre ce problème, forcer l'apprentissage sur les classes minoritaires, diversifier et élargir le jeu de données, ou combiner les deux approches.

Pour le modèle initial sur les séries AppleTV+ (M2), l'accuracy obtenue est de 62,5%. La précision de certaines classes reste médiocre, reflétant également un déséquilibre dans les données et les difficultés à bien prédire les classifications minoritaires.
Les autres modèles ne seront pas abordés puisque les problèmes restent les mêmes, avec des précisions trop faibles sur certaines classes malgré des modèles prometteurs.
Il a donc été décidé d'élargir les jeux de données à 5 000 films et 5 000 séries (toujours selon la popularité sur IMDb et conditionné au fait qu'il y ait au moins un diffuseur en France).

8. 4. Résultats améliorés des modèles.

Pour suivre l'évolution des performances, nous reprenons les mêmes plateformes que dans les résultats précédemment présentés, à savoir Netflix et AppleTV+. Cette fois-ci les modèles ne se basent plus sur des arbres de décision mais intègrent les approches de gradiant boosting expliquées en section 8.2.

Les hyperparamètres techniques des modèles ont été optimisés grâce à des travaux techniques de recherche d'optimums, qui ne seront pas détaillés ici, car ce n'est pas l'objectif de ce rapport.
La sélection des variables explicatives a été affinée[9] et les jeux de données considérablement élargis, ce qui en théorie permet d'améliorer la précision des prédictions.

Pour le modèle finale des séries Netflix (M3) l'accuracy globale a légèrement augmenté atteignant 76%.
Cette augmentation est négligeable (~1.5%) mais la différence la plus notable se situe au niveau de la précision par classe puisqu'aucune n'a désormais de précision nulle ce qui améliore considérablement la perception globale du modèle. La précision reste néanmoins encore trop faible pour certaines classes ce qui indique que le modèle reste perfectible.

Pour AppleTV+ (M4) les résultats sont nettement meilleurs avec une précision globale de 82,5%.
Plus important encore la précision par classe est très satisfaisante, les valeurs les plus basses atteignant 75%, tandis que les meilleures atteignent 100%.
La matrice de confusion suit étroitement la diagonale témoignant d'un modèle fiable, bien calibré et équilibré.

8. 5. Limites des modèles et déséquilibre des classes.

La plupart des difficultés rencontrées lors de la modélisation proviennent du déséquilibre entre les classes dans le jeu de données. Si l'on reprend la lecture comparative basée sur le hasard pour évaluer le résultat final sur Netflix, on constate que le modèle final reste nettement supérieur à un modèle de prédiction aléatoire (76 % contre 100/6% = 16.7%).

Cependant la classe 16+ représente à elle seule 45 % des échantillons du jeu de données, il serait ainsi possible d'atteindre une accuracy globale de 45 % simplement en prédisant systématiquement cette classe bien que cela serait totalement inefficace pour toutes les autres clases.

Cette limitation potentielle reste bien contrôlée ici puisque :

- L'accuracy globale du modèle dépasse largement la proportion de la classe majoritaire (45%), ce qui témoigne de sa capacité à différencier les classes.
- Des techniques de sur-échantillonnage[10] ont été réalisées (privilégiées au sous-échantillonnage, plus adapté aux très grands jeux de données) afin de standardiser le nombre d'exemples pour chaque classe. Cela a limité l'impact du déséquilibre et n'a pas affecté significativement la performance globale du modèle, ce qui confirme que le problème lié au déséquilibre des classes n'en est pas un pour Netflix.

9. Compétences techniques développées.

9. 1. Compétences techniques développées.

Au cours de ce stage j'ai pu développer de solides compétences en ingénierie des données, notamment en automatisant l'extraction d'informations sur de nombreux sites via des techniques avancées et diverses de web scraping depuis différentes sources (une dizaine) et la mise en place de pipelines de données robustes.
Ce projet m'a également permis de renforcer mon expérience dans la création et la gestion de bases de données, avec l'élaboration de trois jeux de données structurés : séries, films et épisodes (ce dernier n'ayant pas pu être pleinement exploité durant le stage par manque de temps).

Cette expérience m'a également initié au machine learning aussi bien sur les aspects théoriques que pratiques comme le feature engineering, l'optimisation des hyperparamètres, ainsi que l'utilisation et la comparaison de différents modèles prédictifs adaptés au contexte des données traitées.

9. 2. Apports méthodologiques et interdisciplinaires.

La rigueur a été essentielle tout au long du projet, ce dernier a exigé une grande autonomie de ma part et un important travail de recherche personnelle. Il a ainsi fallu anticiper dès la construction des jeux de données leur possible modification ou enrichissement futur, afin de garantir la flexibilité du pipeline de données.

Pour la partie machine learning, une organisation structurée était indispensable, avec 6 plateformes SVOD et 2 types d'œuvres, cela représentait la gestion de 12 modèles différents, auxquels s'ajoutaient de nombreuses variantes selon le choix des modèles, la sélection des variables explicatives, et l'optimisation des paramètres.

La vulgarisation et l'explicabilité des résultats ont été fondamentales.
Il a souvent été nécessaire d'expliquer des analyses informatiques et des résultats de modèles statistiques à des personnes expertes dans le domaine juridique, mais non familiarisées avec les concepts du machine learning.

10. Conclusion.

Ce projet a pour moi été une expérience riche et formatrice car il mêle une de mes passion personnelle (séries) et un apprentissages techniques approfondis. Travailler sur la classification d'âge des œuvres audiovisuelles disponibles sur différentes plateformes SVOD exige une démarche de traitement complexe des données et d'élaboration de modèles prédictifs adaptés toujours dans une vision collaborative Droit-Informatique.

L'ingénierie des données a constitué une part centrale de mon travail. Grâce à l'automatisation du webscraping et au développement de pipelines fiables il a été possible de collecter et structurer des jeux de données cohérents (un pour les films, séries et épisodes.)
La structure des œuvres par plateforme permet de faciliter la maintenance (évolutions futures) et la segmentation de l'analyse prédictive.

La modélisation a fait appel à diverses approches de machine learning. J'ai travaillé sur le réglage des hyperparamètres, la sélection des variables explicatives et l'enrichissement des données et ces efforts ont conduit à une amélioration sensible des performances
Les résultats montrent notamment une meilleure précision par classe et globalement plus équilibrée même si certaines limites liées au déséquilibre des classes persistent.

Au-delà de l'aspect technique cette expérience a renforcé mon intérêt pour les projets qui relient analyse de données et secteur culturel. Elle m'a aussi préparé à évoluer dans des environnements exigeant autonomie et polyvalence.

11. Bibliographie.

Dépôt Github du projet : *“https://github.com/Sofiane-h31/Stage”*

[1] Octopusteam, *“full-imdb-dataset”*
https://www.kaggle.com/datasets/octopusteam/full-imdb-dataset
(Indisponlble mais le jeu de données est sur le GitHub du projet : Datasets/data.csv)

[2] Ouest-France, *“Rhône. Les parents d’une élève « traumatisée » par un film réclament 11 000 € d’indemnités”*
https://www.ouest-france.fr/societe/justice/rhone-les-parents-d-une-eleve-traumatisee-par-un-film-reclament-11-000-euros-d-indemnites-7498998

[3] Arcom, *“Série "It’s a Sin" diffusée le 18 mars 2024 : France 2 mise en garde”*
https://www.arcom.fr/se-documenter/espace-juridique/decisions/serie-its-sin-diffusee-le-18-mars-2024-france-2-mise-en-garde

[4] Netlix, *“Catégories d'âge pour les séries et les films sur Netflix”*
https://help.netflix.com/fr/node/2064

[5] Electronic-Mango, *“Simple JustWatch Python API”*
https://github.com/Electronic-Mango/simple-justwatch-python-api
https://electronic-mango.github.io/simple-justwatch-python-api/

[6] Davide Alberani, H. Turgut Uyar, *“Cinemagoer Documentation”*
https://imdbpy.readthedocs.io/_/downloads/en/stable/pdf/

[7] TMDb, https://developer.themoviedb.org/docs/getting-started

[8] Catboost & XGBoost Documentations,
https://catboost.ai/
https://xgboost.readthedocs.io/en/stable/

[9] Aris,*"Understanding Wrapper Methods in Machine Learning"*
https://arismuhandisin.medium.com/understanding-wrapper-methods-in-machine-learning-a-guide-to-feature-selection-23f71059abf8

Machine Learnia, *"FEATURE SELECTION avec SKLEARN"*
https://www.youtube.com/watch?v=T4nZDuakYlU&t=1205s

[10] Mate Voros,*"Handling Imbalanced Datasets with XGBoost"*
https://medium.com/@mate.voros1998/handling-imbalanced-datasets-with-xgboost-optimizing-model-performance-with-smart-parameter-tuning-18568c7783cf

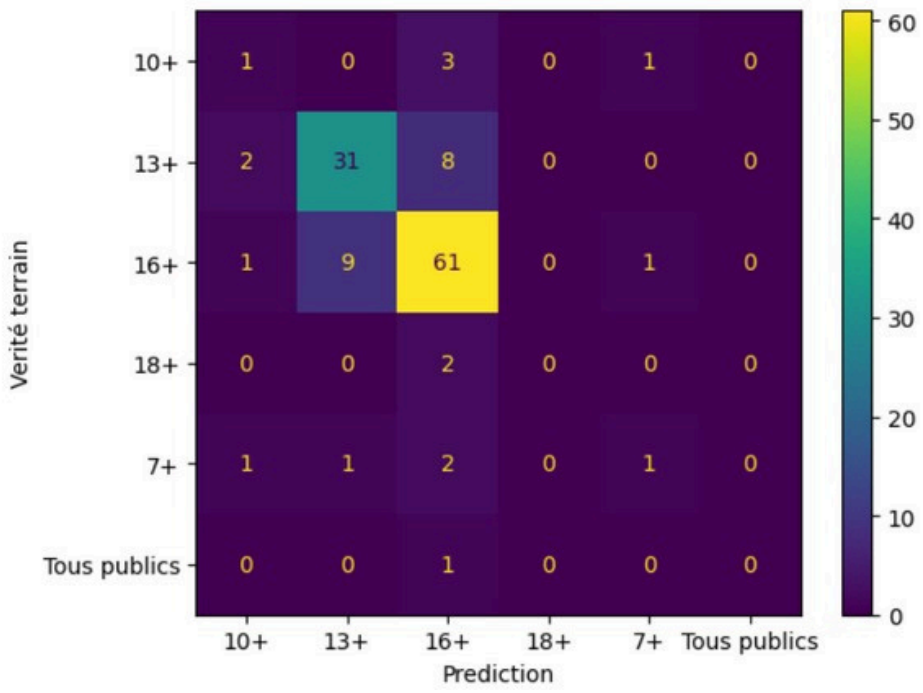
12. Annexes.

Modèle Netflix initial (M1).

Accuracy : 0.746031746031746

	precision	recall	f1-score	support
10+	0.20	0.20	0.20	5
13+	0.76	0.76	0.76	41
16+	0.85	0.79	0.82	77
18+	0.00	0.00	0.00	0
7+	0.20	0.33	0.25	3
Tous publics	0.00	0.00	0.00	0
accuracy			0.75	126
macro avg	0.33	0.35	0.34	126
weighted avg	0.78	0.75	0.76	126

Lecture : Le modèle a une précision de 76 % pour la classe “16+”.
Il s’est entraîné sur 77 données de la classe “16+” (support = 77).

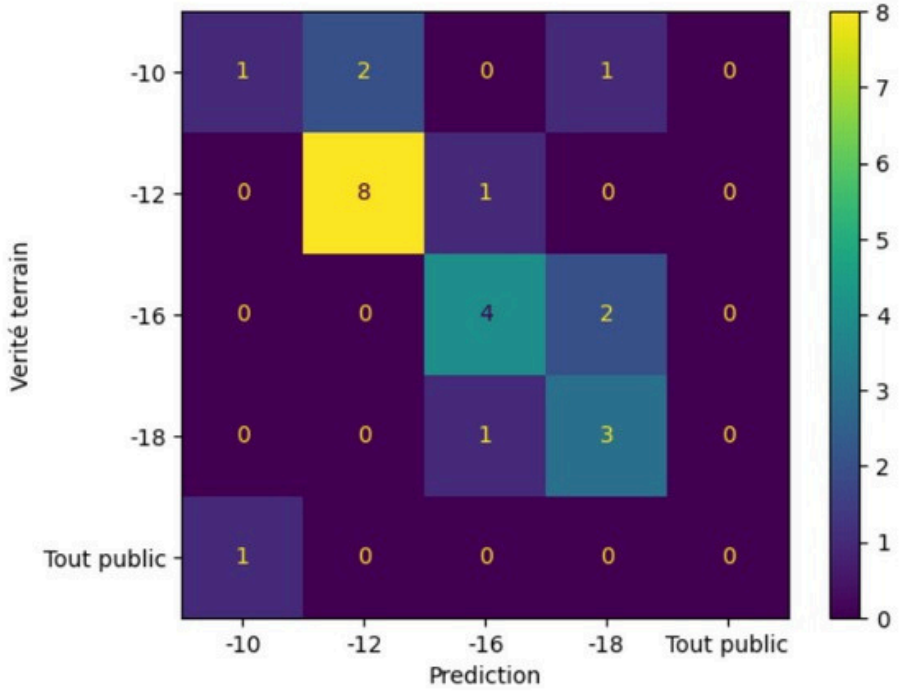


Lecture : Pour la classe “13+”, le modèle a correctement prédit 31 fois et s’est trompé 8 fois en prédisant “16+” et 2 fois en prédisant “10+”.

Modèle AppleTV+ initial (M2).

Accuracy : 0.6666666666666666

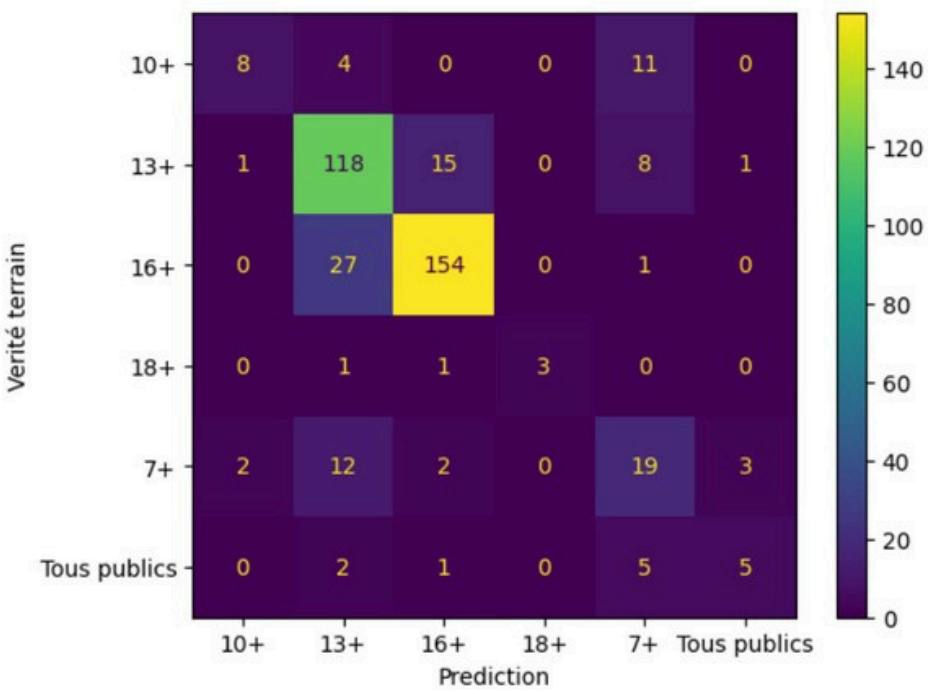
	precision	recall	f1-score	support
-10	0.25	0.50	0.33	2
-12	0.89	0.80	0.84	10
-16	0.67	0.67	0.67	6
-18	0.75	0.50	0.60	6
Tout public	0.00	0.00	0.00	0
accuracy			0.67	24
macro avg	0.51	0.49	0.49	24
weighted avg	0.75	0.67	0.70	24



Modèle Netflix final (M3).

Accuracy : 0.7599009900990099

	precision	recall	f1-score	support
0	0.35	0.73	0.47	11
1	0.83	0.72	0.77	164
2	0.85	0.89	0.87	173
3	0.60	1.00	0.75	3
4	0.50	0.43	0.46	44
5	0.38	0.56	0.45	9
accuracy			0.76	404
macro avg	0.58	0.72	0.63	404
weighted avg	0.77	0.76	0.76	404



Modèle AppleTV+ final (M4).

Accuracy : 0.825

	precision	recall	f1-score	support
0	0.83	0.83	0.83	6
1	0.75	0.80	0.77	15
2	0.75	0.75	0.75	8
3	1.00	1.00	1.00	4
4	1.00	0.86	0.92	7
accuracy			0.82	40
macro avg	0.87	0.85	0.86	40
weighted avg	0.83	0.82	0.83	40