INTELLIGENCE ARTIFICIELLE I

IFT-2003

TRAVAIL PRATIQUE 3

PROBLÈME 3 : APPRENTISSAGE AUTOMATIQUE

ÉQUIPE NO 5

LANOIE, PATRICK

111 146 172

LAURIN, KAREN-FAE

111 104 290

LEBLANC, MÉLANIE

536 854 864

VARGAS, CYNTHIA-ALEXANDRA

537 178 151

TRAVAIL PRÉSENTÉ LE 18 DÉCEMBRE 2023 À

LAURENCE CAPUS

DÉPARTEMENT D'INFORMATIQUE ET DE GÉNIE LOGICIEL

UNIVERSITÉ LAVAL

AUTOMNE 2023

TABLE DES MATIÈRES

1. INTRODUCTION

Le présent rapport a pour but de résoudre le problème 3 : « Apprentissage automatique » qui est décrit à l'annexe 1 de l'énoncé des TP2 et TP3. Ce problème se résume à l'utilisation d'une technique d'apprentissage automatique pour mettre sur pied un moteur de recommandation de films. L'objectif de cet outil est de faire des recommandations basées sur les données disponibles dans le profil d'un utilisateur afin de lui proposer le film qu'il serait le plus susceptible d'aimer.

Pour résoudre ce problème, nous proposons l'utilisation d'un arbre de décision pour faire une recommandation à l'utilisateur en fonction des données collectées sur son profil. Plus spécifiquement, la construction de l'arbre se ferait à partir de l'historique de films écoutés par l'utilisateur. Ainsi, pour chaque film, plusieurs attributs sont répertoriés, à savoir le genre, la durée, la réception critique, l'année de sortie et la restriction d'âge. De plus, l'utilisateur devra attribuer une mention « aimé/pas aimé » pour chaque film visionné. Ces données nous permettent donc de construire un arbre pour cet utilisateur en fonction de ces attributs où la classe à prédire est « aimé/pas aimé ». De cette manière, pour un film quelconque du répertoire du moteur, l'arbre de décision permettra de prédire si l'utilisateur devrait aimer ce film et de lui faire des recommandations en conséquence.

Pour illustrer cette démarche, nous proposons 5 profils d'utilisateurs qui pourraient être utilisés pour générer un tel arbre de recherche. De plus, nous détaillons aussi la construction de l'un de ces arbres et explorons comment certains films seraient ou non proposés par le moteur.

On note ici que les données portant sur les films sont tirées des sites web IMBD et Rotten Tomatoes.

La prochaine section explore la solution au problème énoncé précédemment en discutant des données et profils pertinents, ainsi que de l'application d'un arbre de décision sur ces données.

2. DESCRIPTION DE LA SOLUTION

Comme abordé dans l'introduction, la solution que nous proposons comprend l'utilisation des données de l'utilisateur pour construire un arbre de décision en fonction de l'historique de l'utilisateur. Pour ce faire, nous devons discuter des données qui seront collectées, détailler la méthode employée pour construire l'arbre, présenter les profils fictifs et finalement expliquer l'application de notre solution sur cet ensemble de données.

2.1. Présentation des données pouvant être collectées

Nous supposons que notre moteur de recherche s'inscrit dans le cadre d'une plateforme de visionnement de films en diffusion. Ainsi, il est raisonnable de croire que chaque utilisateur de la plateforme détienne un profil qui enregistre son historique de visionnement. Pour chacun des films visionnés, les caractéristiques du film sont répertoriées dans une base de données et l'utilisateur doit indiquer s'il a aimé ou pas le film. On constate donc qu'au fur et à mesure que l'utilisateur visionne des films sur la plateforme et attribue des mentions à ces films, ces nouvelles données sont collectées et permettent de faire des recommandations de plus en plus précises à l'utilisateur.

Ainsi, pour chaque profil, les données récoltées sont les films écoutés et la mention attribuée par l'utilisateur. Pour chaque film, les attributs suivants sont répertoriés :

1 - **Genre :** Le genre correspond à la catégorie générale du film pour ce qui est du sujet et du type. Il s'agit d'une valeur nominale qui s'inscrit dans un grand ensemble de valeurs possibles. Par exemple, un film peut être du genre drame biographique, drame sentimental, drame judiciaire, comédie sportive, comédie sentimentale, suspense d'épouvante, etc. De plus, les bordures entre ces catégories sont subjectives et un film peut appartenir à plusieurs genres. Ainsi, aux fins du présent travail, nous nous limitons à un seul genre par film et nous utilisons seulement les grandes catégories pour chaque genre, p. ex. drame biographique et drame sentimental sont regroupés en « drame ».

2 - **Restriction :** La restriction correspond à la cote attribuée au film qui indique à quel type d'auditoire il s'adresse. Aux fins du travail, on retient les 3 cotes suivantes : « PG », « 13+ » et « R ». Ces cotes visent respectivement le public général, les personnes de 13 ans et plus et les personnes et 18 ans et plus. Similairement à ce qui a été mentionné pour le genre, on

regroupe certaines restrictions spécifiques sous une bannière générale, p. ex. « 13+ Violence-Horreur » et « 13+ Érotisme » sont regroupés sous « 13+ ».

3 - **Durée :** La durée correspond au temps en minutes que dure le film. Cet attribut est un nombre entier. Afin de pouvoir représenter la durée dans l'arbre de décision, on catégorise la durée d'un film de la manière suivante, en supposant qu'un film dure régulièrement entre 90 et 135 minutes :

Durée du film	Classement
> 135 minutes	Long
Entre 90 et 135 minutes	Normal
< 90 minutes	Court

Figure 1 : Classement de la durée

4 - **Année de sortie :** L'année de sortie correspond à l'année où le film est paru au cinéma pour la première fois. Cet attribut est aussi un nombre entier. Afin de pouvoir représenter l'année de sortie dans l'arbre de décision, on catégorise un film de la manière suivante :

Année de sortie	Classement
> 2013	Récent
Entre 2003 et 2013	Moyen
< 2003	Vieux

Figure 2 : Classement de l'année de sortie

5 - **Appréciation par l'audience :** L'appréciation correspond à l'impression globale de l'audience quant au film. Cet attribut est répertorié sous la forme d'un score, mais, afin de pouvoir le représenter dans l'arbre de décision, on catégorise l'appréciation d'un film de la manière suivante :

Appréciation Rotten Tomatoes	*Classement*
> 85	Bon
Entre 60 et 85	Moyen
< 60	Faible

Figure 3 : Classement de l'appréciation de l'audience

En somme, pour chaque profil d'utilisateur, on peut enregistrer et représenter sous forme tabulaire tous les films visionnés par l'utilisateur, les attributs énumérés ci-haut pour chacun d'entre eux et finalement la mention « aimé/pas aimé » qui a été attribuée après le visionnement.

On aborde maintenant la technique d'apprentissage automatique qui sera appliquée sur les données récoltées pour effectuer les recommandations.

2.2. Présentation de la technique d'apprentissage automatique employée

À cette étape, on rappelle que l'objectif est de proposer un film à l'utilisateur en fonction des données récoltées sur son profil, c.-à-d. déterminer si un film particulier sera « aimé/pas aimé » en considérant l'historique des films qu'il a visionnés. Afin de résoudre ce problème, nous proposons l'utilisation d'un arbre de décision pour mettre en évidence des règles qui permettent de représenter les préférences de chaque utilisateur. Ainsi, on veut pouvoir classer un nouveau film selon si l'utilisateur va l'aimer ou pas. Pour ce faire, les propriétés utilisées pour construire l'arbre sont les attributs des films visionnés par cet utilisateur.

Lorsqu'on construit un arbre de décision, l'objectif est qu'il soit le plus petit possible afin d'être le plus facilement interprétable. Dans cet arbre, chaque nœud correspond à un attribut et chaque branche de ce nœud correspond à une différente valeur de cet attribut. Ainsi, pour atteindre cet objectif, les attributs sont sélectionnés en fonction de leur importance pour le profil en question. Plus précisément, à tout niveau de l'arbre, le choix de l'attribut doit se faire en fonction du gain d'information qui sera engendré par ce choix.

Le calcul du gain d'information permet de déterminer quel choix d'attribut diminue le plus l'entropie des sous-groupes, c.-à-d. quel attribut permet d'obtenir les sous-groupes les plus homogènes possible. Le calcul du gain d'information fait appel aux trois formules suivantes vues en classe :

$$Gain(P) = I(M) - E(P)$$

$$I(M) = \sum_{i=1}^{n} -p(m_i)\log_2(p(m_i))$$

$$E(P) = \sum_{i=1}^{n} |C_i|/|C| * I(C_i)$$

Figure 4 : Formules du calcul du gain d'information

En appliquant ces formules, on peut calculer à chaque étape l'attribut ayant le plus grand gain d'information et ainsi construire notre arbre de manière optimale. On procède ainsi jusqu'à ce que tous les sous-groupes soient homogènes, ou jusqu'à ce que tous les attributs soient épuisés. Ces formules sont appliquées pour construire l'arbre donné en exemple à la section 2.4. Toutefois, avant de procéder, on doit préalablement décrire les exemples de profils d'utilisateurs.

2.3. Exemples de profils

Conformément à l'énoncé, nous proposons 5 exemples de profils d'utilisateurs différents qui permettraient d'effectuer des recommandations de films en fonction de leurs données récoltées. Pour chacun d'entre eux, nous présentons sous forme tabulaire les données collectées, à savoir les films qu'il a visionnés, les attributs de chacun de ces films et la mention que l'utilisateur leur a attribuée. Ces profils servent à entraîner le système, ils ont donc certaines caractéristiques prédictibles qui devraient ressortir lors de la création des arbres de décision.

2.3.1. Profil n° 1

Ce profil est celui d'un jeune adulte qui a un penchant pour les animations et les films dramatiques, majoritairement des films classés « 13+ ». Voici les données collectées pour ce profil.

Entrées	Nom du film	Année de sortie	Appréciation	Durée	Restriction	Genre	Aimé
1	Harry Potter and the Deathly Hallows : Part 2	Moyen	Bon	Normal	13+	Fantaisie	FAUX
2	Spider-Man	Vieux	Moyen	Normal	13+	Action	VRAI
3	Transformers	Moyen	Moyen	Long	13+	Action	VRAI
4	Ghost Rider	Moyen	Faible	Normal	13+	Fantaisie	VRAI
5	Gardians of the Galaxy	Récent	Bon	Normal	13+	Aventure	VRAI
6	Ant-Man	Récent	Moyen	Normal	13+	Action	VRAI
7	Ant-Man and the Wasp	Récent	Moyen	Normal	13+	Aventure	VRAI
8	The Day After Tomorrow	Moyen	Faible	Normal	13+	Action	VRAI
9	In this Corner of the World	Récent	Bon	Normal	13+	Drame	VRAI
10	Goblin Slayer: Goblin's Crown	Récent	Bon	Court	R	Animation	VRAI
11	The Last : Naruto the Movie	Récent	Moyen	Normal	13+	Aventure	VRAI
12	Violet Evergarden : The Movie	Récent	Bon	Long	13+	Drame	VRAI
13	Your Name	Récent	Bon	Normal	PG	Drame	VRAI
14	Words Bubble Up Like Soda Pop	Récent	Bon	Court	PG	Animation	VRAI
15	A Wisker Away	Récent	Moyen	Normal	PG	Drame	VRAI
16	Bubble	Récent	Moyen	Normal	R	Animation	VRAI

17	Drifting Home	Récent	Moyen	Normal	PG	Aventure	VRAI
18	Spirited Away	Vieux	Bon	Normal	PG	Fantaisie	FAUX
19	Space Jam	Vieux	Moyen	Court	PG	Comédie	FAUX
20	Dawn of the Planet of the Apes	Récent	Bon	Normal	13+	Action	FAUX
21	Zombieland	Moyen	Bon	Court	R	Comédie	FAUX
22	Percy Jackson & the Olympians : The Lightning Thief	Moyen	Faible	Normal	PG	Aventure	FAUX

Figure 5 : Données collectées pour le profil n° 1

2.3.2. Profil n° 2

Le profil 2 est celui des membres d'une jeune famille. Les films d'animation et de famille sont préférés. Ils n'apprécient pas les titres « 13+ » et « R » puisqu'ils ont de jeunes enfants. De plus, avec la vie mouvementée, ils n'ont pas de patience pour des films de longue durée. Voici les données collectées pour ce profil.

Entrées	Nom du film	Année de sortie	Appréciation	Durée	Restriction	Genre	Aimé
1	The Lion King	Vieux	Bon	Court	PG	Animation	VRAI
2	Balto	Vieux	Moyen	Court	PG	Animation	VRAI
3	Princess Mononoke	Vieux	Bon	Normal	13+	Animation	VRAI
4	My Neighbor Totoro	Vieux	Bon	Court	PG	Famille	VRAI
5	Eight Below	Moyen	Moyen	Normal	PG	Drame	VRAI
6	Ice Age	Vieux	Moyen	Court	PG	Comédie	VRAI
7	Step Up: All In	Récent	Faible	Normal	13+	Drame	VRAI
8	Soul Surfer	Moyen	Moyen	Normal	PG	Biographie	VRAI
9	Madagascar	Moyen	Moyen	Court	PG	Aventure	VRAI
10	Despicable Me	Moyen	Moyen	Normal	PG	Animation	VRAI
11	Wall-E	Moyen	Bon	Normal	PG	Famille	VRAI
12	Inside Out	Récent	Bon	Normal	PG	Animation	VRAI
13	E.T. The Extra-Terrestrial	Vieux	Moyen	Normal	PG	Famille	VRAI
14	Moana	Récent	Bon	Normal	PG	Musicale	VRAI
15	Despicable Me 2	Moyen	Moyen	Normal	PG	Comédie	VRAI
16	A Wrinkle In Time	Récent	Faible	Normal	PG	Aventure	FAUX
17	Night at the Museum	Moyen	Moyen	Normal	PG	Famille	VRAI
18	Night at the Museum : secret of the tomb	Récent	Faible	Normal	PG	Famille	VRAI
19	Final Destination	Vieux	Moyen	Normal	R	Suspense	FAUX
20	The Mask	Vieux	Moyen	Normal	13+	Comédie	FAUX
21	The Grinch	Vieux	Faible	Normal	PG	Saisonnier	FAUX
22	Mamma Mia!	Moyen	Moyen	Normal	13+	Musicale	FAUX
23	The Eye	Moyen	Faible	Normal	13+	Mystère	FAUX

Figure 6 : Données collectées pour le profil n° 2

2.3.3. Profil n° 3

Cet utilisateur aime les grands titres d'aventure et de sci-fi/fantastique. Il ne participe pas aux célébrations des fêtes saisonnières. Cela devrait donc se refléter dans ses préférences de film. Voici les données collectées pour ce profil.

Entrées	Nom du film	Année de sortie	Appréciation	Durée	Restriction	Genre	Aimé
1	Enemy at the Gates	Vieux	Moyen	Normal	R	Guerre	VRAI
2	Mission : Impossible Rogue Nation	Récent	Bon	Normal	13+	Action	VRAI
3	Mission : Impossible — Dead Reckoning, Part One	Récent	Bon	Long	13+	Action	VRAI
4	Men in Black	Vieux	Moyen	Normal	13+	Action	VRAI
5	Men in Black II	Vieux	Faible	Court	13+	Action	VRAI
6	Riddick	Moyen	Faible	Normal	R	Aventure	VRAI
7	The Chronicles of Riddick	Moyen	Moyen	Normal	13+	Sci-fi	VRAI
8	Pirates of the Caribbean : The curse of the Black Pearl	Moyen	Bon	Long	13+	Fantaisie	VRAI
9	Pirates of the Caribbean : At World'd End	Moyen	Moyen	Long	13+	Fantaisie	VRAI
10	Taken	Moyen	Moyen	Normal	13+	Crime	VRAI
11	Taken 2	Moyen	Faible	Normal	13+	Crime	VRAI
12	The Book of Eli	Moyen	Moyen	Normal	R	Aventure	VRAI
13	Inception	Moyen	Bon	Long	13+	Sci-fi	VRAI
14	John Wick	Récent	Moyen	Normal	R	Action	VRAI
15	Indiana Jones and the Last Crusade	Vieux	Bon	Normal	13+	Aventure	VRAI
16	Barbie	Récent	Moyen	Normal	13+	Comédie	VRAI
17	Home Alone	Vieux	Moyen	Normal	PG	Saisonnier	FAUX
18	Elf	Moyen	Moyen	Normal	PG	Saisonnier	FAUX
19	Charlie's Angels	Vieux	Faible	Normal	13+	Action	FAUX
20	The Girl with the Dragon Tattoo	Moyen	Bon	Long	R	Crime	FAUX
21	Lucy	Récent	Faible	Court	R	Sci-fi	FAUX
22	Harry Potter and the Prisoner of Azkaban	Moyen	Bon	Long	PG	Fantaisie	VRAI

Figure 7 : Données collectées pour le profil n° 3

2.3.4. Profil n° 4

Ce profil représente un fanatique de film d'horreur qui apprécie aussi les actions/aventures. Ce sont donc en général des films « 13+ » ou « R ». Voici les données collectées pour ce profil.

Entrées	Nom du film	Année de sortie	Appréciation	Durée	Restriction	Genre	Aimé
1	Tomb raider	Récent	Faible	Normal	13+	Aventure	VRAI
2	Lara Croft : Tomb raider	Vieux	Faible	Normal	13+	Aventure	VRAI
3	Alien	Vieux	Bon	Normal	R	Horreur	VRAI
4	Aliens	Vieux	Bon	Long	R	Sci-Fi	VRAI
5	The Silence of the Lambs	Vieux	Bon	Normal	R	Crime	VRAI
6	Hannibal	Vieux	Moyen	Normal	R	Crime	VRAI
7	Blade	Vieux	Moyen	Normal	R	Action	VRAI
8	Resident Evil	Vieux	Moyen	Normal	R	Horreur	VRAI
9	Resident Evil : Extinction	Moyen	Faible	Normal	R	Action	VRAI
10	Resident Evil: The Final Chapter	Récent	Faible	Normal	R	Action	VRAI
11	Jurassic Park	Vieux	Bon	Normal	13+	Aventure	VRAI
12	War of the Worlds	Moyen	Faible	Normal	13+	Sci-Fi	VRAI
13	Sinister	Moyen	Moyen	Normal	R	Horreur	VRAI
14	Sinister 2	Récent	Faible	Normal	R	Horreur	VRAI
15	30 Days of Night	Moyen	Faible	Normal	R	Suspense	VRAI
16	Insidious	Moyen	Moyen	Normal	13+	Horreur	VRAI
17	Insidious : Chapter 2	Moyen	Faible	Normal	13+	Suspense	VRAI
18	Insidious : Chapter 3	Récent	Faible	Normal	13+	Horreur	VRAI
19	Frozen	Moyen	Moyen	Normal	PG	Animation	FAUX
20	Austin Powers: International Man of Mystery	Vieux	Moyen	Court	13+	Comédie	FAUX
21	The Strangers	Moyen	Faible	Court	R	Mystère	FAUX
22	The Chronicles of Narnia : Prince Caspian	Moyen	Moyen	Long	PG	Fantaisie	FAUX
23	Paranormal Activity	Moyen	Faible	Court	R	Horreur	FAUX

Figure 8 : Données collectées pour le profil n° 4

2.3.5. Profil n° 5

Cet utilisateur représente un profil d'utilisateur plus réaliste. Il est plus nuancé dans ses préférences de films pour un même genre. Cependant, il a un penchant pour les films de crime. Contrairement aux autres profils, on s'attendrait donc à un arbre de décision plus profond pour aider à distinguer ces nuances. Voici les données collectées pour ce profil.

Entrées	Nom du film	Année de sortie	Appréciation	Durée	Restriction	Genre	Aimé
1	The Silence of the Lambs	Vieux	Bon	Normal	R	Crime	VRAI
2	Frozen	Moyen	Moyen	Normal	PG	Animation	FAUX
3	P.S. I Love You	Moyen	Moyen	Normal	13+	Romance	FAUX
4	Eragon	Moyen	Faible	Normal	PG	Fantaisie	FAUX
5	The Lion King	Vieux	Bon	Court	PG	Animation	VRAI

6	E.T. The Extra-Terrestrial	Vieux	Moyen	Normal	PG	Famille	FAUX
7	Seven	Vieux	Bon	Normal	R	Crime	VRAI
8	The Mask	Vieux	Moyen	Normal	13+	Comédie	VRAI
9	The Grinch	Vieux	Faible	Normal	PG	Saisonnier	VRAI
10	Enemy at the Gates	Vieux	Moyen	Normal	R	Guerre	VRAI
11	The Fast and The Furious	Vieux	Moyen	Normal	13+	Action	FAUX
12	Pirates of the Caribbean : The curse of the Black Pearl	Moyen	Bon	Long	13+	Fantaisie	VRAI
13	300	Moyen	Bon	Normal	R	Drame	VRAI
14	Taken	Moyen	Moyen	Normal	13+	Crime	VRAI
15	Taken 2	Moyen	Faible	Normal	13+	Crime	VRAI
16	Venom	Récent	Moyen	Normal	13+	Action	VRAI
17	John Wick	Récent	Moyen	Normal	R	Action	VRAI
18	Transformers	Moyen	Moyen	Long	13+	Action	VRAI
19	The Dark Knight	Moyen	Bon	Long	13+	Crime	VRAI
20	Battleship	Moyen	Faible	Normal	13+	Action	FAUX
21	Deadpool	Récent	Bon	Normal	R	Comédie	VRAI
22	The Lord of the Rings: The Fellowship of the Ring	Vieux	Bon	Long	13+	Fantaisie	VRAI
23	Zombieland	Moyen	Bon	Court	R	Comédie	FAUX

Figure 9 : Données collectées pour le profil n° 5

Bref, ces 5 profils différents permettent de produire un arbre de décision propre à chaque profil. Cet arbre peut alors être utilisé pour faire des suggestions de films à cet utilisateur. L'application de cette technique d'apprentissage automatique à nos données est décrite à la section suivante. On inclut aussi une démonstration de cette technique basée sur les données du profil n° 1.

2.4. Application de la technique d'apprentissage automatique sur le profil n° 1.

Soit l'ensemble M, l'ensemble qui est constitué des 22 données collectées du profil de l'utilisateur n° 1 sur les films aimés ou non, soit 16 films aimés (« Vrai ») et 6 films non aimés (« Faux »). Calculons l'entropie de M :

$$I(M) = - (p(\text{vrai})*\log_2(p(\text{vrai})) + p(\text{faux})*\log_2(p(\text{faux})))$$

$$I(M) = - ((16/22)*\log_2(16/22) + (6/22)*\log_2(6/22))$$

$$I(M) = 0{,}845$$

À titre d'exemple, le calcul du gain d'information de l'attribut « Genre » est démontré. Le « Genre » est constitué des valeurs « Action », « Animation », « Aventure », « Comédie », « Drame » et « Fantaisie ». Chacun possède respectivement 5, 3, 5, 2, 4 et 3 valeurs.

$$\text{Gain(genre)} = I(M) - E(\text{genre})$$

$$E(\text{genre}) = p(ac)*I(ac) + p(an)*I(an) + p(av)*I(av) + p(c)*I(c) + p(d)*I(d) + p(f)*I(f)$$

$$E(\text{genre}) = (5/22)*I(ac) + (3/22)*I(an) + (5/22)*I(av) + (2/22)*I(c) + (4/22)*I(d) + (3/22)*I(f)$$

Par souci de la concision du rapport, seul le calcul de l'entropie d'« Action » sera pris en exemple. Celui-ci a 4 valeurs affirmatives contre 1 valeur négative.

$$I(ac) = -\left(p(\text{vrai})*\log_2(p(\text{vrai})) + p(\text{faux})*\log_2(p(\text{faux}))\right)$$

$$I(ac) = -\left((4/5)*\log_2(4/5) + (1/5)*\log_2(1/5)\right)$$

$$I(ac) = 0{,}722$$

Au final, en appliquant la même démarche pour calculer l'entropie des autres genres, nous obtenons E(genre) = 0,453 et donc un gain d'information de 0,392 pour le genre. Le calcul du gain de tous les autres attributs se fait de la même façon et les résultats sont présentés dans le tableau suivant :

Attributs	Gain Information
Appréciation	0.073
Durée	0.074
Genre	0.392
Année	0.216
Restriction	0.052

Figure 10 : Gain d'information des différents attributs

Puisque l'attribut « Genre » a la valeur la plus élevée en gain d'information, il sera choisi comme étant la racine de l'arbre. À cette étape, nous pouvons remarquer que nous avons atteint les feuilles de l'arbre pour les genres « Animation », « Comédie » et « Drame » puisque leurs valeurs « Aimé » sont identiques (voir figure 12).

Pour les autres catégories de genre, l'arbre décisionnel doit être continué. Le calcul reste sensiblement le même, cependant l'ensemble de départ pour le calcul est limité aux films de la branche. Prenons comme exemple l'attribut « Appréciation » pour la branche où l'attribut

« Genre » correspond à « Action ». Comme nous avons mentionné plus haut, cet ensemble est constitué de 4 valeurs affirmatives contre 1 valeur négative, ce qui constitue le nouvel ensemble M.

$$I(M) = - (p(\text{vrai})*\log_2(p(\text{vrai})) + p(\text{faux})*\log_2(p(\text{faux})))$$
$$I(M) = - ((4/5)*\log_2(4/5) + (1/5)*\log_2(1/5))$$
$$I(M) = 0{,}722$$

Maintenant, calculons le gain d'information de l'attribut « Appréciation », où A = {moyen, faible, moyen, bon, moyen}.

$$\text{Gain}(A) = I(M) - E(A)$$
$$E(A) = p(\text{faible})*I(\text{faible}) + p(\text{moyen})*I(\text{moyen}) + p(\text{bon})*I(\text{bon})$$
$$E(A) = (1/5)*I(\text{faible}) + (3/5)*I(\text{moyen}) + (1/5)*I(\text{bon})$$
$$\text{Où } I(\text{faible}) = I([\text{vrai}]) = 0,$$
$$I(\text{moyen}) = I([\text{vrai,vrai,vrai}]) = 0,$$
$$I(\text{bon}) = I([\text{faux}]) = 0$$

Ainsi, Gain(A) = 0,722. On effectue alors ces calculs pour tous les attributs restants et l'on obtient le tableau présenté à la figure 11. On constate que l'attribut « Appréciation » sera choisi comme attribut au deuxième niveau de la branche « Action » puisque c'est celui avec le gain d'information le plus élevé.

Attributs	Gain Information
Appréciation	0.722
Durée	0.073
Année	0.322
Restriction	0

Figure 11 : Gain d'information des différents attributs sachant que le genre est « Action »

Cette démarche est appliquée à l'ensemble de l'arbre. Une fois terminée, on obtient l'arbre suivant pour le profil n° 1.

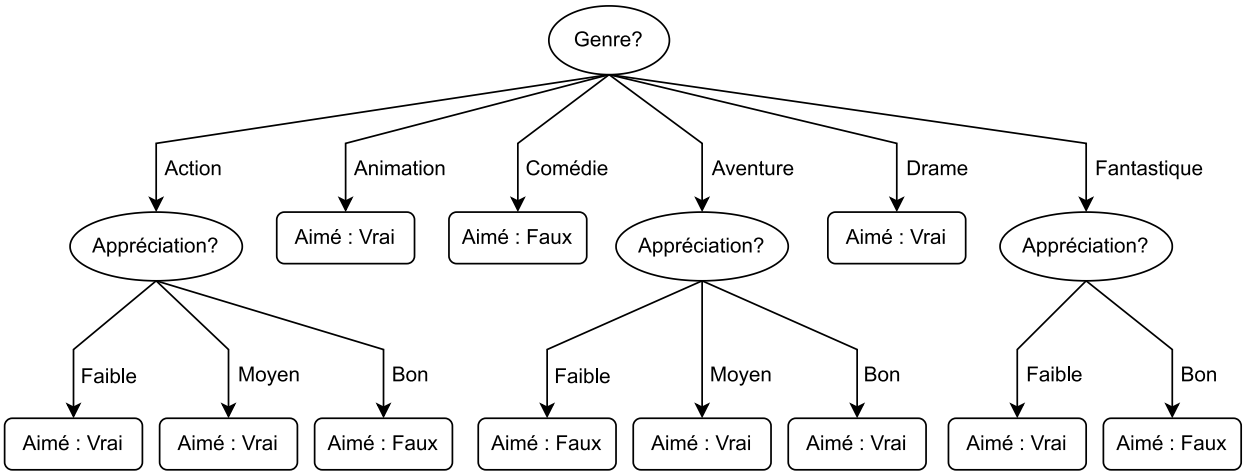

Figure 12 : Arbre de décision du profil n° 1

Depuis la figure de l'arbre de décision ci-dessus, il est possible de traduire celui-ci par les 11 règles de décision suivantes :

1. SI Genre = Action ET Appréciation = Faible ALORS Aimé = Vrai
2. SI Genre = Action ET Appréciation = Moyen ALORS Aimé = Vrai
3. SI Genre = Action ET Appréciation = Bon ALORS Aimé = Faux
4. SI Genre = Animation ALORS Aimé = Vrai
5. SI Genre = Comédie ALORS Aimé = Faux
6. SI Genre = Aventure ET Appréciation = Faible ALORS Aimé = Faux
7. SI Genre = Aventure ET Appréciation = Moyen ALORS Aimé = Vrai
8. SI Genre = Aventure ET Appréciation = Bon ALORS Aimé = Vrai
9. SI Genre = Drame ALORS Aimé = Vrai
10. SI Genre = Fantastique ET Appréciation = Faible ALORS Aimé = Vrai
11. SI Genre = Fantastique ET Appréciation = Bon ALORS Aimé = Faux

Pour décider de la recommandation, chaque film non vu par l'utilisateur est passé au travers de l'arbre de décision. Ils commencent tous à la racine de l'arbre, puis descendent dans les branches inférieures tout en respectant les conditions de branchement basées sur les données du film considéré. Lorsqu'un film arrive à une feuille, celle-ci contient la réponse, à savoir s'il devrait être aimé et donc recommandé à l'utilisateur.

À titre d'exemple, considérons l'ensemble présenté à la figure 13, soit 3 films pour lesquels on veut déterminer s'ils devraient être recommandés à l'utilisateur du profil n° 1.

Nom du film	Année de sortie	Appréciation	Durée	Restriction	Genre
Deadpool	Récent	Bon	Normal	R	Comédie
Titanic	Vieux	Moyen	Long	13+	Drame
Sharknado	Moyen	Faible	Court	13+	Action

Figure 1314 : Données des films dont la recommandation est à déterminer

Les deux premiers films « Deadpool » et « Titanic » ont atteint une décision dès le premier niveau de l'arbre de décision avec respectivement Aimé = **FAUX** et Aimé = **VRAI** en sortie. Ces résultats proviennent des règles 5 et 9 qui indiquent respectivement que les comédies ne sont pas aimées alors que les drames sont aimés. Pour ce qui est du film « Sharknado », c'est aussi Aimé = **VRAI**. La décision provient cependant du niveau le plus profond de l'arbre. On constate qu'elle concorde avec la règle 1, puisque c'est un film d'action avec une appréciation faible.

3. CONCLUSION

En conclusion, le présent rapport détaille la résolution du problème 3, soit l'utilisation d'une technique d'apprentissage automatique pour développer un moteur de recommandation de films sur une plateforme de visionnement en ligne. Pour résoudre ce problème, nous abordons premièrement les données qui pourraient être collectées pour chaque profil. Ensuite, nous détaillons notre approche pour construire un arbre de décision à partir de ces données qui permet de prédire si un utilisateur donné est susceptible, ou non, d'aimer un film particulier. Finalement, nous présentons des profils fictifs et illustrons l'application de notre technique sur l'un d'entre eux.

En ce qui concerne l'utilisation de l'IA pour résoudre ce problème, les avantages sont multiples. Pour débuter, comme nous utilisons un arbre de décision pour résoudre ce problème, nous sommes en mesure de mettre en évidence des règles exprimables quant aux préférences de l'utilisateur. Ces règles permettent non seulement de guider les recommandations du moteur, mais aussi d'identifier certaines tendances du profil que l'utilisateur lui-même pourrait ne pas être en mesure de verbaliser. De plus, en appliquant la solution proposée, nous sommes en mesure d'effectuer des suggestions pertinentes pour l'utilisateur sans son intervention, c'est-à-dire sans qu'il soit nécessaire de lui demander explicitement ses préférences. On se base uniquement sur l'historique des films qu'il visionne et la seule action requise de la part de l'utilisateur est d'attribuer la mention « aimé/pas aimé ». Finalement, l'utilisation de l'IA dans ce cas permet de produire des résultats très efficacement. En effet, une fois que l'arbre de décision est généré pour un profil, il peut être parcouru très rapidement pour évaluer la pertinence d'une recommandation. Ainsi, à partir d'une banque de nouveaux films, on peut facilement générer un ensemble de recommandations appropriées en utilisant l'arbre pour évaluer chaque film de la banque envisagée.

Toutefois, bien que nous soyons de l'avis que notre solution adresse la problématique posée, il existe tout de même plusieurs ajouts ou améliorations qui pourraient être inclus au projet pour le rendre encore plus intéressant.

D'une part, à propos des attributs des films qui sont répertoriés, il va sans dire qu'on pourrait augmenter le discernement de l'arbre en augmentant le nombre d'attributs considérés. On peut citer par exemple la provenance du film ou le nombre d'acclamations remportées comme d'autres attributs intéressants. De plus, toujours à propos des attributs, le genre d'un film pourrait être relaxé

pour inclure plusieurs valeurs. En effet, lorsque nous avons récolté les données pour créer les 5 exemples de profil, nous avons été confrontés à la difficulté de la représentation du genre d'un film. La très grande majorité des films nécessitent 2 à 3 genres pour être représentés correctement, alors que nous avons décidé de n'en inclure qu'un seul. Nous avons donc dû faire un choix parmi les genres indiqués, ce qui s'avérait parfois difficile lorsque des genres très différents (par exemple « comédie » et « horreur ») représentaient un même film. L'approche d'inclure plusieurs valeurs pour le genre correspondraient donc mieux à la réalité, puisqu'un film peut rarement être correctement décrit par un seul genre.

D'autre part, en ce qui concerne la mention « aimé/pas aimé », plusieurs améliorations pourraient être apportées. En premier lieu, la pertinence de l'arbre dépend grandement du nombre de visionnements qui ont été effectués sur la plateforme. Cet aspect est particulièrement évident lorsqu'un nouveau profil est créé. Dans ce cas, aucune recommandation ne peut être faite, puisque l'utilisateur n'a visionné aucun film. Pour pallier ce problème, nous pourrions demander à un nouvel utilisateur d'identifier parmi une liste de films populaires lesquels il a « aimé/pas aimé » afin d'obtenir un premier échantillon de données pour construire un arbre initial. Dans le même ordre d'idées, puisque l'attribution de la mention « aimé/pas aimé » est cruciale au bon fonctionnement du moteur de recommandations, nous sommes à la merci de l'utilisateur sur ce point. Ainsi, pour s'assurer de la pertinence des données récoltées, on pourrait obliger l'utilisateur à attribuer une mention avant d'écouter un autre film. Autrement, on pourrait utiliser des indicateurs implicites pour attribuer automatiquement la mention, p. ex. si l'utilisateur ne finit pas un film avant d'en commencer un autre, probablement qu'il n'a pas aimé le premier. Finalement, en ce qui concerne la mention elle-même, la catégorisation binaire employée connait des limites, surtout lorsqu'on discute de quelque chose d'aussi nuancé que l'appréciation d'un film. Une amélioration pourrait donc être d'utiliser un score sur une échelle de 1 à 10. Cette méthode donnerait plus de flexibilité à l'utilisateur pour exprimer son appréciation et nous pourrions développer l'arbre de décision basé sur un seuil ou une classement de ce qui est suffisamment « bon » pour être recommandé.

Enfin, en ce qui concerne l'expérience de résoudre ce problème, nous sommes satisfaits de la solution proposée. Toutefois, nous ne pouvons prétendre y être arrivées facilement. En effet, bien que le consensus ait facilement été atteint quant au choix de la technique d'apprentissage

automatique utilisée, le choix de comment l'appliquer a fait couler bien plus d'encre. Plus précisément, l'interprétation de l'énoncé a fait l'objet de plusieurs discussions animées. Nous avons oscillé entre une approche par profil et une approche globale. De plus, plusieurs attributs ont été successivement considérés, écartés et reconsidérés, le tout dans le but de déterminer les données qui sont les plus pertinentes pour prédire la qualité d'une recommandation. En fin de compte, la solution proposée est le fruit de multiples changements de position et concessions faites dans le but de répondre au problème posé au meilleur de nos capacités. Bref, bien qu'instructive, cette expérience fût un frappant rappel de la complexité de développer une bonne solution en intelligence artificielle.