Projet M1 Fouille de données

Contexte

Dans le cadre de ce projet, nous cherchons à aider à analyser certains films sortis ces 100 dernières années dans 66 pays.

Le jeu de données
- 28 variables: "movie_title" "color" "num_critic_for_reviews" "movie_facebook_likes" "duration" "director_name" "director_facebook_likes" "actor_3_name" "actor_3_facebook_likes" "actor_2_name" "actor_2_facebook_likes" "actor_1_name" "actor_1_facebook_likes" "gross" "genres" "num_voted_users" "cast_total_facebook_likes" "facenumber_in_poster" "plot_keywords" "movie_imdb_link" "num_user_for_reviews" "language" "country" "content_rating" "budget" "title_year" "imdb_score" "aspect_ratio"
- 5043 films

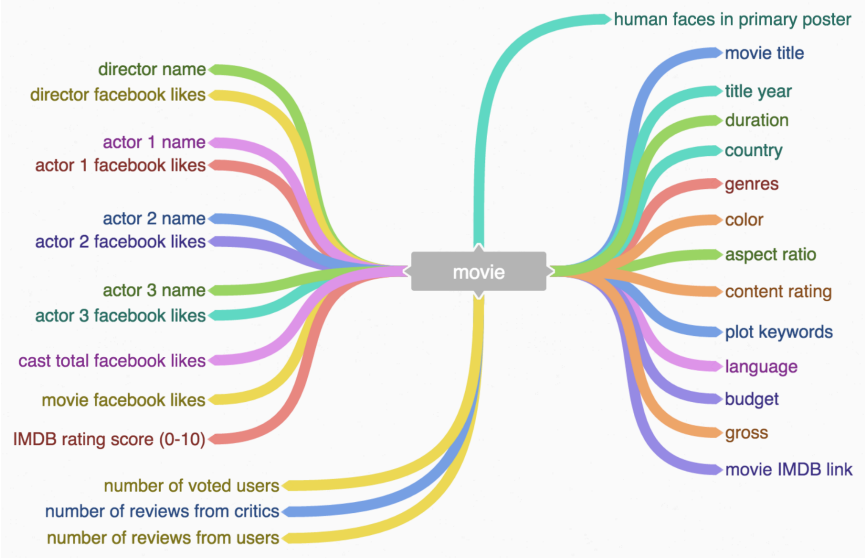

Figure 1 : les attributs disponibles

Quelques informations utiles sur les données :

- Il y a des données manquantes dans le fichier (présence de « 0 » pour certains attributs, il faudra les traiter quand nécessaire, il ne faut pas les supprimer d'office !).
- Le budget est en monnaie locale et correspond à la valeur du budget de l'époque, cette variable est donc à prendre avec précaution car la valeur est également soumise à l'inflation.
- Certains titres de film contiennent des virgules (et les champs sont séparés par des virgules !!!)

Logiciels : Vous utiliserez WEKA pour faire vos analyses.

Rendu : 2 rendus sont à faire : un rapport pour les parties 1 et 2, un deuxième rapport pour la partie 3. Ces rapports peuvent être réalisés en binôme.

Partie 1 – Analyse descriptive des données

Dans un premier temps, il convient de prendre en main les données. Vous vous servirez du support de cours, des visualisations offertes par WEKA ou votre tableur préféré. Votre rapport contiendra l'analyse à destination d'un multi-millionaire passionné de cinéma, celui-ci cherche à connaître mieux les films, leur impact, leur critère de succès afin de pouvoir à l'avenir financer des nouvelles propositions. Par la suite, nous appellerons cette personne le décideur. Vous l'aiderez à tirer quelques premières conclusions.

Création d'attributs

Le budget et ce que rapporte un film sont en monnaie locale et dépendent de l'année. Vous allez créer un attribut `ratio_rentabilité` qui est un ratio entre ce qu'a rapporté le film et ce qu'il a couté.
Créer un nouvel attribut « `ratio_rentabilitée_disc` » qui discrétise de façon pertinente le ratio de rentabilité précédemment calculé.
Créer un nouvel attribut « `score_IMBD_disc` » qui discrétisera le score IMDB.

Attributs et Statistique descriptive

Présentez les différents attributs (type, signification …), donnez quelques statistiques descriptives et quelques visualisation appropriées.

Importation dans Weka

L'importation dans Weka n'est pas automatique. En effet, le CSV contient des données que WEKA n'arrivera pas à transformer. Vous pouvez aller sur https://weka.wikispaces.com/ARFF+%28book+version%29 pour avoir un rappel des entêtes d'un fichier `arff`.
Ainsi il est rappelé qu'un fichier `arff` doit avoir une entête, qu'une donnée manquante est un `?`, que du texte de type `string` est entre `''` …

Partie 2 – Segmentation de films

On cherche à savoir si les films de la base présentent des similarités.

1. Mettez le fichier au bon format pour l'importer dans WEKA si ce n'est pas déjà fait
2. Proposer différentes segmentations avec seulement les attributs suivants
   - « country »
   - « duration »
   - « Movie year »
   - « Movie facebook popularity »
   - top 3 actors/actresses facebook popularity
   - « facenumber_in_poster »
   - « genre »
   - « budget »
   - « gross »
   - « ratio_rentabilité »
   - « ratio_rentabilitée_disc »
   - « score IMBD »

   Les segmentations proposées devront inclure des éléments permettant de les comparer, caractériser les groupes obtenus, voir si les films regroupés sont similaires en terme de « score IMBD », ou en « ratio_rentabilité ».
   N'oubliez pas que si vous voulez caractériser ces groupes en fonction d'un attribut celui-ci ne doit pas être inclus dans la segmentation.
   Vous indiquerez vos conclusions et conseils par rapport aux résultats obtenus pour le décideur.

Partie 3 – Prédiction

On cherche à savoir prédire 2 éléments : soit le score IMBD du film soit le ratio de rentabilité. Certains algorithmes de classification peuvent travailler directement avec des valeurs continues à prédire (algorithme de régression) d'autres ont besoin de données discrétisées (algorithme de classification).
L'objectif de cette partie est de trouver le meilleur algorithme pour trouver le score IMBD puis le meilleur pour prédire le ratio de rentabilité. Suivant les algorithmes, on utilisera la variable discrétisée ou non.
Utilisez votre TP de classification pour vous guider dans l'analyse et le rendu.

Prédiction du score IMDB

1. Appliquez différents algorithmes, présentez un tableau récapitulatif des résultats obtenus indiquant si les données ont due être discrétisées, la qualité obtenue (taux d'erreur, F-mesure …), le temps de création du modèle, la partition apprentissage/test utilisée … et tout ce que vous jugez utile de nous présenter
2. Sélectionnez quelques modèles donnés par les méthodes les plus performantes
3. Discutez ces modèles

Prédiction du ratio de rentabilité

1. Appliquez différents algorithmes, présentez un tableau récapitulatif des résultats obtenus indiquant si les données ont due être discrétisées, la qualité obtenue (taux d'erreur, F-mesure …), le temps de création du modèle, la partition apprentissage/test utilisée … et tout ce que vous jugez utile de nous présenter
2. Sélectionnez quelques modèles donnés par les méthodes les plus performantes
3. Discutez ces modèles