Projet de spécialité – Kaggle

Conduite de projet

BAILLET Valerian – BEAUPÈRE Matthias

FISCHMAN Adrien – MEURET Thibault

Plan :

0 – Introduction :

Dans le cadre du projet de spécialité à l'ENSIMAG notre groupe a choisi d'approfondir le cours de Fouille de Données en se mesurant à plusieurs challenges proposés par le site web **Kaggle**. Ce dernier propose de nombreux défis relatifs au **machine-learning**. Leur difficulté est très variable allant du niveau « introductif », avec des problèmes pour la plupart déjà résolus qui permettent de mieux comprendre les méthodes classiques, au niveau « avancé », avec des problèmes ouverts offrant généralement de belles récompenses pour ceux qui obtiennent les meilleurs résultats : jobs, cash-prize, etc.

Notre démarche consiste donc à **se confronter à un problème ouvert**. Nous pourrons ainsi découvrir en détail la démarche classique du machine learning : analyse des données, modélisation, expérimentation. De plus, en faisant ce choix, nous aurons l'occasion d' étoffer nos compétences au fil des discussions des **Kaggle Master** : déterminer pourquoi un modèle est adapté à un certain type de données ou au contraire pourquoi il ne l'est pas, comprendre comment représenter les données efficacement, savoir quand décider d'écarter certaines variables, etc.

I – Objectif :

Notre premier objectif est la **compréhension** des méthodes et des choix propres aux problèmes de machine-learning. A la fin de ce projet nous souhaitons être capables de **justifier le choix d'un algorithme** par rapport à un autre étant donné un problème fixé, ainsi qu'**estimer les gains** réalisés, ou réalisables, au vu des traitements appliqués aux données fournies en entrée

Néanmoins, nous estimons qu'une bonne compréhension de toutes ces méthodes devrait nous permettre de réaliser des algorithmes assez performants. Ainsi, nous serions satisfait de constater que notre démarche expérimentale nous permet de nous classer dans le **top 20 % du leaderboard** associé au problème choisi.

II – Organisation du projet :

1) Charte d'équipe :

Afin de mener à bien notre projet, nous avons mis en place un ensemble de règles et de principes à suivre. Ces derniers ont été décidés et validés par l'ensemble de l'équipe pour nous permettre de rentabiliser au mieux le temps qui nous est imposé dans le cadre du projet de spécialité.

i) Valeurs d’équipe :

Nous nous sommes tous mis d’accord sur un ensemble de **valeurs fondamentales** que nous devions respecter au sein de l’équipe afin de **travailler dans les meilleurs conditions possibles** :

- ✔ Esprit d'équipe et d'initiative
- ✔ Volonté d'entraide
- ✔ Écouter les autres
- ✔ Volonté de partage et d'écoute
- ✔ Assiduité et ponctualité
- ✔ Respect d'autrui

ii) Communication au sein de l’équipe :

Nous avons décidé d'utiliser en priorité la **communication orale** car elle permet d'avoir un retour immédiat sur les questions ou propositions que peuvent avoir les membres de l'équipe. Chaque membre de l'équipe doit pouvoir se rendre disponible pour expliquer ses choix ou résoudre un problème soulevé par un autre membre de l'équipe.

Tous les autres moyens de communications modernes (groupe de discussion instantanée, mail, SMS) seront utilisés **si la personne n'est pas joignable de façon orale**.

Nous utilisons **git** pour garder une trace des informations récoltées sur les différents algorithmes rencontrés ainsi que pour partager le code implémenté.

2) Fonctionnement de l’équipe :

*Afin de s’assurer de la **qualité** et de la **cohérence** globale de notre travail nous avons décidé de travailler en reprenant certains point des **méthodes agiles**, notamment le principe de **meeting quotidien** et de **rush**.*

Nous avons mis en place une **réunion quotidienne à 10h30** afin de discuter de ce qui a été fait, de ce qui reste à faire et de ce qui devrait être fait d’ici le lendemain. Ces meeting sont inclus au sein de **rush de 2-3 jours** qui correspondent à une étape de la **démarche expérimentale** décrite dans le point suivant.

Afin de travailler dans les meilleurs conditions, nous avons analysé les profils de chacun afin de savoir vers qui se diriger en cas de question spécifique ; ainsi, nous avons démarqué un groupe plutôt orienté **« théorie »** (statistiques, fouilles de données, etc.) constitué de **Adrien et Thibault**, et un groupe plutôt orienté **« implémentation et optimisation »** (implémentation des algorithmes identifiés comme les plus adaptés, amélioration du code afin de traiter plus de données, etc.) constitué de **Matthias et Valérian**. Cependant, puisque nous souhaitons tous comprendre l’ensemble du processus, nous réalisons des comptes-rendus, écrits et oraux, lorsque nous avons réalisé quelque chose de remarquable.

3) Méthodologie :

Afin de s'imprégner au mieux des méthodes propres aux problèmes de machine-learning, nous avons décidé de suivre une démarche expérimentale bien particulière. Ainsi, le problème est découpé en **4 étapes principales** :

- **Analyse :** il s'agit de comprendre au mieux les données. Pour cela, nous devrons les représenter graphiquement, analyser les dépendances et chercher à déterminer les variables les plus importantes à considérer.
- **Modélisation** : il s'agit de choisir le modèle statistique qui approxime au mieux le problème au vue de l'analyse des données.
- **Implémentation** : il s'agit de mettre en place les algorithmes retenus dans un langage adapté au machine-learning et à l'utilisation de très grosses données (plusieurs Go dans notre cas)
- **Expérimentation** : il s'agit de faire tourner nos algorithmes sur les données fournies et conclure sur la pertinence du modèle choisi et/ou de l'implémentation réalisée.

Tout au long du projet, l'ensemble des choix réalisés et des résultats obtenus sont tenus à jour dans deux fichiers sous la forme de **Markdown** et de **notebook Jupyter**.

RQ : Afin d'implémenter du code orienté machine-learning, nous avons décidé de coder l'ensemble de nos algorithme en R ou en Python.

III – Obstacles identifiés :

Nous avons identifiés **deux problèmes récurrents** à l'ensemble des challenge proposés par Kaggle. Le premier est de réussir à **représenter graphiquement des données multi-dimensionnelles** ; il va falloir faire des choix, au risque qu'ils soient mauvais, sur les variables à représenter. Le second est la **taille des données** à faire traiter par la machine ; les données pouvant faire plusieurs Go, il va falloir trouver un moyen d'optimiser tous les processus implémentés, sans quoi les calculs pourraient devenir démesurément longs.

IV – Ressources :

Afin de nous épauler, en terme de compétences techniques, nous pourrons compter sur l'aide de notre encadrant **Michael G. B. Blum**, spécialiste en statistiques et bio-informatique, ainsi que sur l'ensemble des **Kaggle Master** via les forums de discussion intégrés à Kaggle.

En ce qui concerne les ressources matérielles, nous disposons tous d'un ordinateur portable que nous pourrons apporter à l' ENSIMAG afin de pouvoir continuer à effectuer des recherches pendant que les données seront en traitement sur les machines.