4ERP-BI & 4SIGMA

2016-2017

Projet

Data Mining

Projet Complémentaire au module Data Mining
//40% de la note semestrielle

Enseignants intervenants :

M. Mohamed Heny SELMI
Technologue à ESPRIT – Data Scientist
medheny.Selmi@esprit.tn

Mme. Dorra Trabelsi
Technologue à ESPRIT – Data Scientist
dorra.trabelsi@esprit.tn

Mme. Wiem Trabelsi
A. Technologue à ESPRIT – Ing. R&D
wiem.trabelsi@esprit.tn

Le but de ce projet est de mettre en œuvre la démarche CRISP-DM d'un « Data Miner » qui doit fouiller dans un volume énorme de données hétérogènes,

et ce à l'aide des techniques du Data Mining pour en extraire des connaissances, des informations pertinentes, des décisions…

Méthodologie de travail :

Le projet se déroule selon la **méthodologie CRISP-DM**, il se compose en **six** étapes :

1. La compréhension de(s) problématique(s) métier

Dès que vous aurez choisi votre thématique, vous serez amenés à entamer la première partie qui consiste à bien comprendre les éléments métiers et les enjeux décisifs relatifs à la prise de décision. Pour ce faire vous devez formuler **de(s) problématique(s)** sous forme de plusieurs questions.

Une synthèse des recherches sur le métier sera très bénéfique pour les prochaines étapes du projet.

2. La collecte et la compréhension des données

Cette phase vise à déterminer précisément les données à analyser, à identifier la qualité des données retenues et à faire le lien entre les données et leur signification d'un point de vue métier.

3. La préparation des données

Cette phase regroupe les activités liées à la construction de l'ensemble précis des données à analyser. Elle inclut ainsi le classement des données en fonction des critères choisis, le nettoyage des données, et surtout leur recodage pour les rendre compatibles avec les algorithmes qui seront utilisés. Dans cette phase, vous êtes amenés à vous assurer de la fiabilité des données à traiter.

4. La modélisation

La modélisation comprend le choix, le paramétrage et le test de différentes méthodes ainsi que l'évaluation des modèles obtenus. Ce processus est d'abord descriptif pour générer de la connaissance, et ce en vous basant sur les indicateurs statistiques et les techniques de la visualisation de données. Il devient ensuite prédictif en expliquant ce qui va se passer, le comportement passé ou le comportement futur. Dans ce cas, vous devez utiliser tous les modèles possibles d'entamer l'étape suivante.

5. L'évaluation

L'évaluation vise à vérifier **le(s) modèle(s)** ou les connaissances obtenues afin de s'assurer qu'ils répondent aux objectifs formulés au début du projet. Elle contribue aussi à la décision de déploiement du modèle ou, si besoin, à son amélioration. A ce stade, on teste notamment la robustesse et la précision des modèles obtenus à l'aide d'outils dédiés.

6. Le déploiement

Il s'agit de l'étape finale du projet. Elle consiste en une mise en production pour les utilisateurs finaux des modèles obtenus. Son objectif : mettre la connaissance obtenue par la modélisation, dans une forme adaptée, et l'intégrer au processus de prise de décision.

Le déploiement peut ainsi aller, selon les objectifs, de la simple rédaction d'un rapport de synthèse décrivant les connaissances obtenues jusqu'à la mise en place d'une application, permettant l'utilisation du modèle obtenu, pour la prédiction de valeurs inconnues d'un élément d'intérêt.

Source de données :

De nombreux échantillons de données sont disponibles sur le Web. On pourra penser à plusieurs sources à savoir :

- Les sites de l'open data (données publiques tunisiennes, françaises, etc.) : http://data.industrie.gov.tn/, http://catalog.industrie.gov.tn/dataset, www.data.gouv.fr/, www.opendatafrance.net/, etc.
- La banque de données de l'INS (Institut National des Statistiques) : http://dataportal.ins.tn/
- Les sites des compétitions de type Data Science : www.kaggle.com, www.datascience.net, etc.
- Open API des réseaux sociaux : https://www.programmableweb.com/
- Les sites de stockage de données comme https://datamarket.azure.com/

NB : Quelques soit les données que vous utilisez, veuillez indiquer les sources. Vous pouvez également utiliser plus qu'une source et plus qu'un échantillon de données pour répondre à la problématique étudiée.

Méthodes et aspect multidisciplinaire du projet :

L'étude de ce projet implique nécessairement plusieurs méthodes ou combinaisons de méthodes du Data Mining.

Cependant, la nature des données mises à votre disposition, peut faire penser à l'implication de nombreuses autres disciplines à savoir le Text Mining, le Web Mining, les Séries Temporelles, l'analyse statistique décisionnelle, les techniques ou les outils de visualisation de données, l'analyse des réseaux sociaux, et le et le Computer Vision (Signal Processing et Pattern Recognition).

Dans ces cas, vous êtes amenés à étudier en profondeur le contenu des données étudiées souvent riche en texte, images et vidéos. Ici, il est intéressant de puiser dans les techniques propres du domaine dédiées à l'extraction de connaissances à partir des contenus riches en images et en vidéos.

Environnements et outils de développement :

Vos développements **doivent être réalisés** avec :

un environnement *R*

(Rcran, RHadoop, Rspark, Rstudio, R shiny, Rattle, Microsoft R, Revolution R Enterprise etc.)

NB : L'utilisation d'autres environnements de développement tels que : Python (Python basique, Anaconda Spider, Anaconda Jupyter Notebook, IPython, PyCharm, etc.), SAS9, Oracle Data Miner, Mahout de Hadoop, Spark MLib sera **très appréciée.**

Délivrables :

Les documents qui doivent être remis le jour de la soutenance :

- **Tous les codes (R et autres)** sur un support numérique bien commentés.
- **Un imprimé du rapport** décrivant tout le travail effectué.
- Une **présentation** (max 20 slides) décrivant les détails de toute étape de CRISP-DM par les points suivants :
  - ✓ **Présentation du métier.**
  - ✓ **Problématique étudiée et questions posées.**
  - ✓ **Compréhension et préparation des données.**
  - ✓ **Analyses et modélisations effectuées, avec argumentations des choix.**
  - ✓ **Présentation et interprétation des résultats obtenus.**
  - ✓ **Déploiement et perspectives du travail effectué.**

NB : Le tout sera sauvegardé sur un CD + une version papier du rapport à transmettre à votre enseignant au plus tard 48h avant la date de la soutenance finale.

Les séances consacrées au projet DM :

1. **Présentation du projet DM et choix des thématiques des projets** [Séance 2/17]
2. **Avancée sur le projet : Début du sketching et présentation de l'état de l'art** [Séance 6/17]

- Présentation de l'état d'avancement de chacun des groupes selon la CRISP-DM.
- Présentation du métier.
- Business objectives (sous formes de questions).
- Compréhension des données.
- Préparation des données.

3. **Atelier suivi continu du projet** [séance 10/17]

- Présentation de l'état d'avancement.
- Exposition des difficultés rencontrées.
- Description de la démarche d'analyse proposée et des résultats attendus.

4. **Atelier DataViz** [séance 15/17]

- Visualisation des résultats.

5. **Soutenance Finale du projet Data Mining** [séance 16/17]

Evaluation :

Ci-joint la fiche d'évaluation qui sera adoptée tout au long ce projet :

Membres de groupe :

Nom	**Prénom**	***Note individuelle – Appréciation Globale***
		Portant sur la discipline, **La présence,** **L'assiduité,** **Le travail fourni au cours des séances (Cours, TP ou projet)** **L'implication dans l'équipe,** **Etc.** **/20**

Livrables :

Rapport (Français ou Anglais)	**/2**
Codes commentés	**/3**
Présentation (Anglais)	**/2**

Critères d'évaluation :

Thématiques d'évaluation	***Note***	***Note Soutenance***
La pertinence de la problématique étudiée.	**/2**	
Le choix des datasets en fonction des données. *La qualité des données retenues.*	**/2**	
Les méthodes décisionnelles mises en œuvre.	**/5**	
La fiabilité des résultats obtenus. *La qualité de présentation des résultats.* *La pertinence des interprétations et commentaires déduits.*	**4/**	

Note finale du Projet Data Mining (note CC du module Data Mining de chaque étudiant)

$$= \frac{\mathbf{2 * Note\ Soutenance\ +\ Note\ Individuelle}}{\mathbf{3}}$$