TER 2021 – 073 - Développement

Just drag and drop

Etudiant(s) : Nicolas Demolin (MAM5-SD), Christel Ralalasoa (MAM5-SD), Ralph El Chalfoun (M2-WIA), Jérémy Hirth Daumas (M2-CASPAR)

Encadrant(s) : Peter Sander (sander@unice.fr)

# Résumé exécutif *(20 lignes, 300 mots max)*

[Insérer ici un résumé d’un maximum de 20 lignes décrivant succinctement l’objectif de votre projet et les résultats que vous espérez fournir. Ce résumé doit être clair, concis et compréhensible par un non-spécialiste.]

Le but de ce projet est de développer une application web qui permettra à l’utilisateur d’analyser des bases de données. En effet, l’utilisateur pourra, grâce à l’implémentation de différents algorithmes de prédiction (Machine Learning), à travers l’application et en quelques clics, faire différentes analyses qui lui seront utiles (par exemple : visualisation des données, compréhension des corrélations entre les colonnes…) via des paramètres défini par le serveur, ou par l’utilisateur lui-même.

Une sécurité sera appliquée aux utilisateurs. Il faudra s’identifier pour accéder à l’application. Les données de ces utilisateurs tant bien leurs identifiants que les bases qu’ils analyseront devront être protégés.

Comme il s’agit d’une application web, certaines vérifications devront être faites. Nous devrons vérifier si les données entrées ne sont pas des attaques (SQL attack), mais aussi nous assurer que le serveur ne soit jamais surchargé afin qu’il ne rende pas l’utilisation de l’application inaccessible.

Les données qui pourront être entrées ne concerneront pas les images ou les audios. En effet, nous ne ferons que de la classification/prédiction sur des bases de données « textuelles ».

Enfin, l’utilisateur aura ses analyses disponibles dans son historique sur l’application, qu’il pourra télécharger sous forme de fichier .html pour les avoir en local.

# Description du projet

## Contexte technologique

* La data est au centre du monde du travail. En effet, toutes les entreprises possèdent des données et certaines ne sont pas exploitées. L’exploitation de données nécessite des développeurs, data scientists, data analysts… Tout ce processus est long et extrêmement coûteux.
* Il existe des méthodes de Machine Learning qui permettent d’exploiter les données afin de pouvoir en tirer une utilisation pertinente.
* Les applications Web sont des technologies très utilisées car simples d’utilisation et intuitives.

## Motivations

[Quels sont les problèmes à résoudre ? Quel pan de la science va avancer grâce à lui ? Quelle utilité a le projet ?]

* L’utilité du projet est de faciliter l’analyse de données et le travail des data scientists.
* Cette application permettrait de raccourcir la chaîne d’analyse de données. En effet, il ne serait plus nécessaire de faire intervenir 7 ou 8 personnes sur un projet d’analyse mais seulement 2 ou 3.
* Pour un étudiant qui s’intéresse à la science de données, ce type d’application est très intéressant. En effet, il pourra l’utiliser pour faciliter ses analyses et mieux comprendre les hyperparamètres d’un modèle de Machine Learning.

## Objectifs à atteindre

[Listez les objectifs que vous comptez atteindre, que ce soit des résultats de recherche ou des livrables.]

* Créer une application web d’analyse de données interactive.
* Assurer une protection contre les cyberattaques (notamment XSS, Session Management, SQL attack etc…)
* Une fois l’application réalisée, créer un réseau de neurones étant capable de ressortir à l’aide d’une vision simple d’une base de données, l’algorithme de Machine Learning à utiliser avec ses hyperparamètres.
* Le client peut télécharger ses analyses sous forme de fichier (par exemple : HTML) pour l’avoir en local.

## Risques identifiés (et contremesures)

[Listez les risques associés à l’obtention des résultats escomptés, et décrirez comment vous comptez les surmonter.]

* Les applications web sont une porte d’entrée pour tout type d’attaque. Une attention particulière devra être portée sur la sécurité, à la fois pour le serveur de l’application mais également pour les futurs utilisateurs qui feront confiance au site en uploadant leurs bases de données.
* La surcharge du serveur qui peut entraîner une inutilisation de l’application. En fonction de la puissance du serveur nous devront établir un plafond de requête par utilisateur.
* Le volume des bases de données. Une taille maximale de fichier devra être fixée afin dans un premier temps, d’éviter les temps trop longs de téléchargement et ensuite réduire le nombre de calculs. Trouver le système d’envois le plus adapté (compression, format, …).
* Une base de données trop compliquée à analyser pour un modèle de Machine Learning.

## Scenarios

[Décrivez 2 à 3 scénarios d’utilisation de votre projet. Ces scénarios doivent être montrés du point de vue des utilisateurs du système que vous construirez. Pour chaque scénario, vous soulignerez les critères d’acceptation, qui servent à prouver que le système permet l’exécution de ces scénarios. Maximum deux pages.]

1er scenario :

Un étudiant doit effectuer une analyse de données sur une base de données pour un TP/projet. Notre application pourra l’aider à faire une analyse rapide et effective de sa base. De plus, d’un point de vue éducatif, cette web app est très intéressante. En effet, si notre étudiant veut comprendre l’influence de certains hyperparamètres sur son algorithme de Machine Learning : il pourra le faire ! Imaginons que l’étudiant ne comprenne pas bien l’intérêt du learning rate dans un algorithme de Gradient Boosting, alors il pourra le modifier aisément sans écrire plusieurs lignes de code et observer différents phénomènes comme le gradient vanishing (learning rate trop petit). Enfin, il pourra télécharger un fichier (par ex : HTML) pour récupérer son travail.

*Exemples de scenario à partir du scenario 1 :*

**Scenario**: Je suis connecté et je charge ma base de données sur l’application web

**Étant donné**que je suis un utilisateur connecté et que je suis sur la page d’accueil

**Lorsque** je clique sur « Import a file » et que je choisis un fichier (du type accepté) et que je clique sur « Load »

**Alors** le système charge ma base de données

2ème scenario :

Le propriétaire d’une épicerie souhaite savoir quels sont les produits qu’il vend le plus et quels sont ceux qu’il ne vend pas assez voire pas ou plus du tout afin de réorganiser le stock de son magasin. Grâce à l’application web, après avoir en amont récupéré les données de son stock avec certains paramètres (par ex : la période, le prix), il pourra faire une analyse des produits qui sont le plus/moins vendus suivant une période (graphe/histogramme), s’il y a un lien entre la proportion de vente de certains produits avec la modification de leur prix (matrice de corrélation), etc. De ce fait, il aura une analyse plus claire de la gestion de ce magasin sans pour autant nécessiter de connaissances préalables en analyse de données.

*Exemples de scenario à partir du scenario 2 :*

**Scenario**: L'utilisateur affiche l’histogramme des ventes de bouteilles de soda suivant les 4 saisons de l’année

**Étant donné**que je suis un utilisateur connecté, que je suis sur la page d’analyse après avoir chargé mon fichier contenant ma base de données,

**Lorsque** que je vais sur la partie « histogramme », que je vais sur « axe des ordonnées », que je tape « soda » puis appuie sur « entrer », que je vais sur « axe des abscisses », que je tape « printemps » puis appuie sur « entrer », que je tape « été » puis appuie sur « entrer », que je tape sur « automne » et appuie sur « entrer » et que je tape « hiver » et appuie sur « entrer »

**Alors** le système m’affiche un histogramme avec en ordonnées le nombre de bouteilles de soda vendues et en abscisses les 4 saisons de l’année.

3ème scenario :

**Scenario**: Je suis un étudiant apprenant les analyses sur les bases de données. Je souhaite réaliser une étude rapide.

**Étant donné** que je ne sais pas quels hyperparamètres choisir

**Lorsque** que j’upload ma base en choisissant l’option paramètre automatique

**Alors** l’application web calcul par deep learning les hyperparamètres optimaux

# Mise en en œuvre *(quels paragraphes, utilisez des bulles)*

Liste d'activités déjà réalisés avant les semaines à plein temps :

* Répartition des tâches et création du GitHub
* Mise au point des objectifs et de la réalisation avec notre encadrant (P.Sander)
* Prototypage de l’application web via IHM
* Décision des langages et outils à utiliser (Typescript/Javascript – React – Python – Chart.js – Node – Yarn – Scikit-Learn -TensorFlow - Burp)

Listes d'activités prévues pour chaque semaine à plein temps

* Un quart du taf
* Un autre quart du taf
* Le troisieme quart du taf
* Le dernier quart du taf LOLL

Organisation du travail (répartition de l'équipe)

- Ralph El Chalfoun s’occupe de la partie développement de l’application web.

- Jérémy Hirth Daumas s’occupe de la partie sécurité

- Nicolas Demolin et Christel Ralalasoa s’occupent de la partie implémentation des méthodes de Machine Learning