UNIVERSITÉ CÔTE D'AZUR

POLYTECH NICE-SOPHIA

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)

1. Résumé exécutif

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 définis 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 (identifiants, bases) devront être protégées.

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 (*annexe*), qu'il pourra télécharger sous forme de fichier .html pour les avoir en local.

2. 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.

Motivations

- 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

- Créer une application web interactive d'analyse de données.
- 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.

- Pour pouvoir enrichir notre algorithme de deep learning en data l'application web devra ressortir les paramètres descriptifs des bases et des tests réalisés par les utilisateurs.
- Le client peut télécharger ses analyses sous forme de fichier (par exemple : HTML) pour l'avoir en local.

Risques identifiés (et contremesures)

- 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.
- 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, ensuite réduire le nombre de calculs et éviter la surcharge du serveur. Trouver le système d'envoi le plus adapté (compression, format, …).

Scenarios

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 (*annexe*)

Alors l'application web calcule par deep learning les hyperparamètres optimaux

Scenario : Je suis gérant d'un magasin et je souhaite afficher 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.

3. Mise en en œuvre

Liste d'activités déjà réalisées 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 (*annexe*)
- 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

- Une première esquisse de l'application web, un client et un serveur communiquant ensemble, priorité au système de login, afin qu'une sécurité puisse être mise en place dès le début.
- Ajout de la fonction d'upload de base à envoyer au serveur. Réception et analyse basique coté serveur. Interface d'affichage des résultats. Protection des potentielles nouvelles attaques.

- Implémentation des historiques : interface et stockage des analyses sur le serveur. Implémentation d'un système de cap pour le nombre de requête et taille du fichier grâce à des observations de performances. Avancement des algorithmes de machine et deep learning.
- Refonte graphique potentielle une fois tous les éléments implémentés. Correction/amélioration des algorithmes. Tests/Scan des vulnérabilités encore présentes et protection adaptée.
- Dernières améliorations en cas de retard. Rédaction du rapport + Confection du diaporama pour la soutenance + Poster + vidéo

Organisation du travail (répartition de l'équipe)

- Ralph El Chalfoun s'occupe de la partie développement de l'application web :

- Une application web en TypeScript/React côté client et Node côté serveur
- L'application proposera les fonctionnalités suivantes :
    - Un système de compte avec login et d'upload de base de données en drag and drop
    - Une interface pour gérer les paramètres des analyses
    - Une interface pour visualiser de manière claire les résultats fournis par le serveur (à l'aide de chart.js)
    - Un système d'historique d'analyses
    - Un moyen de télécharger au format HTML les analyses pour une utilisation externe
- Le client et le serveur s'échangeront notamment les bases et les résultats
- Le serveur devra gérer plusieurs connexions et requêtes des clients et réaliser des appels pythons pour l'analyse des bases.

- Jérémy Hirth Daumas s'occupe de la partie sécurité :

- Lutter contre les attaques XSS, injections NoSQL, et éventuellement les attaque XXE.
- Contribution du développement de la partie backend :
    - Mise en œuvre d'une authentification JWT (JSON Web Token).
    - Sécurisation des sessions.
    - Communication avec la base de données mongoDB (pour les identifiants des utilisateurs).
    - Mise en place d'une restriction par utilisateur (et/ou par IP) pour un nombre maximal d'analyse.
    - Implémentation d'un captcha pour éviter les robots.
    - Double authentification via Email.
- Lorsque l'application sera fonctionnelle, une analyse de celle-ci sera faite avec des outils de sécurité tel que Burp, Nmap, SonarQube afin de détecter d'éventuelle faille et les corriger.

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

- Implémenter un maximum d'indicateurs pour aider à l'analyse des bases de données
- Implémenter un maximum d'algorithmes de machine learning à l'aide de la librairie sklearn afin de réaliser les prédictions sur les bases des utilisateurs
- Créer un algorithme de deep learning capable de ressortir l'algorithme et les hyperparamètres à utiliser pour réaliser les prédictions
- Aider sur la partie visualisation pour faire comprendre la data à l'utilisateur

Annexes

Templates de l'application :