Projet chef d’œuvre

Détection des troubles de la santé mentale et du besoin de traitement avec l'IA chez les gens qui travaillent dans le domaine de la technologie.

Geoffroy Daumer

Table des matières

[1. Introduction du projet 5](#_Toc161867694)

[2. Compréhension du besoin client 5](#_Toc161867695)

[3. Présentation des données 5](#_Toc161867696)

[4. Etat de l'art 6](#_Toc161867697)

[5. Choix techniques liés au projet 7](#_Toc161867698)

[5.1 Partie application 7](#_Toc161867699)

[5.2 Partie IA 9](#_Toc161867700)

[6. Mise en œuvre du projet 11](#_Toc161867701)

[6.1. L'organisation technique et l'environnement de développement tout au long de la production 11](#_Toc161867702)

[6.2. Gestion de projet 11](#_Toc161867703)

[7. Retours d'expérience sur les outils, techniques et compétences à l'œuvre tout au long du projet 13](#_Toc161867704)

[7.1 Datasets de santé mentale 13](#_Toc161867705)

[8. Bilan du projet et améliorations envisageables 14](#_Toc161867706)

[9. Conclusion 14](#_Toc161867707)

[Références 23](#_Toc161867708)

[Annexes 24](#_Toc161867709)

# 1. Introduction du projet

Nous vivons dans une société où les cas de problèmes de santé mentale sont en forte croissance, et ils restent à ce jour mal soignés.

Dans le cas des travailleurs du domaine de la technologie, on remarque qu'ils sont souvent assis, derrière des écrans. Ce manque d'activité physique, voir sociale, peut être un facteur d'un bien être diminuant.

Avec l'IA, on pourrait détecter plus facilement les troubles de la santé mentale, et ainsi les prendre en charge plus rapidement.

Un client m'a contacté pour développer une application qui intégrerait l'IA, cette application permettrait de détecter le besoin de traitement pour des problèmes de santé mentale chez les travailleurs du domaine de la technologie.

Ce client a mené une enquête auprès des travailleurs, leur posant un certain nombre de questions, à choix multiples, et recueillant les résultats sous forme d'un jeu de données.

Parmi les questions posées, une question "Avez-vous recherché un traitement pour des problèmes de santé mentale ?", c'est cette question qui servira de cible à l'algorithme d'IA, pour l'entraînement du modèle de prédiction.

La solution finale apportée est une application sur laquelle on peut répondre au questionnaire du client, et où un algorithme d'IA prédit en conséquence si nous devrions considérer le fait de prendre un traitement pour la santé mentale.

# 2. Présentation des données

Le jeu de données comprend 1259 lignes et 27 colonnes, il y a donc 1259 questionnaires remplis.

Il y a une colonne "Timestamp", une colonne "Age" (valeur continue), "Genre", "Country", "state" (pour les États américains), le reste des colonnes c'est des questions à choix multiples, allant de 2 à 5 environ. Il y a une colonne "Comments" pour les commentaires.

Une image contenant capture d’écran, texte, noir

Description générée automatiquement

# 4. Etat de l'art

Pour réaliser la veille sur les algorithmes à utiliser, il faut chercher du côté des modèles de classification, car la prédiction est une catégorie : besoin d'un traitement ou non.

De plus, on sait que notre jeu de données se compose essentiellement de variables catégoriques.

Voici une liste des modèles de classification (1) :

**La machine à vecteurs de support (SVM)** : Classificateur linéaire, sépare les données à travers des lignes (hyperplans). Fonctionne bien pour les identifier des classes simples (exemple deux sorties).

**Les arbres de décision** : Algorithme classifiant les données sous forme de branches. On part d'une racine, et chaque échantillon prend une direction en fonction de ses caractéristiques. Ce qui permet de prédire une variable de réponse.

**La répartition en K-moyennes (K-means)** : Trie les données en différents groupes en fonction de leurs caractéristiques. Etablit une moyenne de référence dans chaque groupe pour définir un profil type. Bonne précision

**Le KNN (K-nearest neighbors)** : Classifie les variables d'un jeu de données en analysant les similitudes entre elles. Utilise un graphique et calcule la différence entre les différents points, l'échantillon est enregistré à la même catégorie que ses n plus proches voisins.

**La régression logistique** : Effectue des corrélations simples entre les entrées et les sorties, pour un nombre fini de résultats.

**Les réseaux de neurones** (2): Modèle de classification très puissant inspiré par le fonctionnement du cerveau humain. Composés de plusieurs couches de neurones interconnectés.

Un réseau est composé d'une couche d'entrée, de couches cachées et d'une couche de sortie.

# 5. Choix techniques liés au projet

Voici un schéma décrivant l'ensemble de la stratégie adoptée pour le fonctionnement de l'application :

Une image contenant texte, capture d’écran, diagramme, Police

Description générée automatiquement

Figure 1 : Schéma fonctionnement application

### 5.1 Partie application

#### 5.1.1 Front-end

Pour réaliser ce projet, j'ai choisi de réaliser une application streamlit, pour sa rapidité de développement. L'utilisateur devra donc répondre au questionnaire sur une interface streamlit pour avoir la prédiction de son état de santé mentale.

#### 5.1.2 Back-end

J'ai utilisé une API, FastAPI, pour le développement des fonctionnalités de l'application.

Pour le tracking des performances, étant en local, j'ai fait un script python qui se lance en même temps que l'application. Le script envoie des requêtes toutes les 5 secondes à l'API et écrit sur le terminal la bonne éxécution des requêtes, s'il y a une erreur, on voit l'erreur en direct et je recevrais un mail.

Mon application se lance avec l'éxécution d'un script .bat, qui lance trois terminaux, un pour l'API, un pour l'application streamlit, un pour le tracking.

Je n'ai pas déployé mon appli sur internet parce que j'ai des problèmes de connexion avec Heroku, pour les autres plateformes de déploiement, je n'ai rien trouvé de simple et gratuit d'utilisation.

#### 5.1.3 Base de données

Les questionnaires et les prédictions sont enregistrés sur une base de données relationnelle SQLite

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Figure 2 : Modèle relationnel des données de la base de donnée SQLite de l'application

### 5.2 Partie IA

J'ai entrainé le modèle d'IA à partir de la librairie python "Scikit-learn", elle regroupe l'essentiel des modèles de machine learning. Les modèles ont été entraînés sur une interface jupyter notebook.

Prétraitement des données :

- Standardisation des colonnes continues : Age

- Encodage One Hot des colonnes discrètes : Les autres questions

Plusieurs modèles ont été comparés :

- LogisticRegression

- KNeighbrosClassifier

- DecisionTreeClassifier

- AdaBoostClassifier

- GradientBoostingClassifier

- RandomForestClassifier

Pour déterminer le meilleur modèle, la précision, le recall et le f1 score des modèles ont été comparés



Le modèle qui a obtenu les meilleures performances est la régression logistiques, avec un f1 score de 0.83

# 6. Mise en œuvre du projet

### 6.1. L'organisation technique et l'environnement de développement tout au long de la production

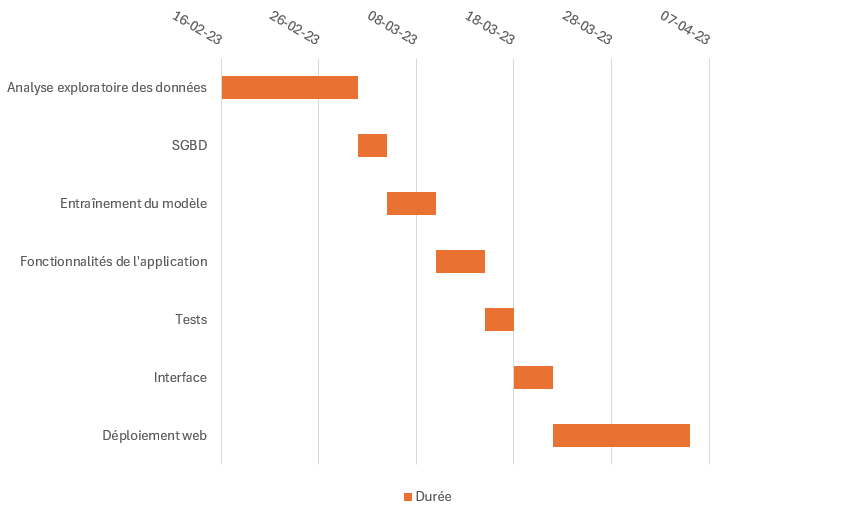
Après avoir développé mon API, il fallait que j'y intègre une base de données pour les compétences de la certification. J'ai donc regardé ce qui se faisait en ligne, je me suis rapidement aperçu que j'étais bloqué parce que c'était payant (Google Cloud Plateform notamment). Nous avions des accès gratuits à Heroku, mais dans mon cas il y avait un problème avec la double authentification.

J'ai eu des problèmes de compatibilité avec Docker et Windows, des packages d'environnements virtuels spécifiques à Windows (Jupyter notebook) faisaient buguer Docker. Au lieu d'ignorer ces packages, j'ai décidé de muter mon environnement de développement sur WSL, ce fut une erreur.

J'ai installé la version GNU/Linux Debian, et j'ai eu des problèmes ingérables de python, sql, etc...

Je suis donc retourné sur Windows pour résoudre mes bugs, et j'ai laissé tomber Docker, ayant ciblé un déploiement local et non web.

### 6.2. Gestion de projet

Voici le diagramme de Gantt de suivi de réalisation du projet :

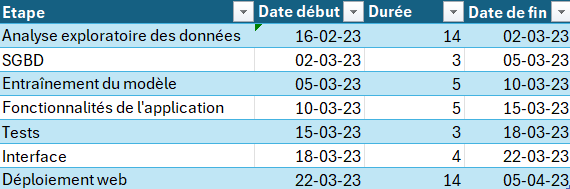


Figure 3 : Diagramme de Gantt

J'ai prévu plus de temps là où je savais que j'étais moins à l'aise, le planning est fait de sorte que je sois tout le temps en avance et qu'il n'y ai pas de problèmes avec les échéances, même en cas d'imprévus.

Pour faire un suivi de l'avancement de mon projet, nous communiquions quand c'était nécessaire avec mon client par mails et téléphone.

Nous avons fait 3 réunions :

- Une réunion le 29 février, pour présenter l'analyse exploratoire des données, et les méthodes de nettoyage choisies

- Une réunion le 7 mars, pour présenter l'entraînement du modèle, et les résultats obtenus

- Une réunion le 28 mars, pour présenter l'application fonctionnelle

Voir Annexe 1 : Rapports d'avancement pour plus de détails

# 7. Retours d'expérience sur les outils, techniques et compétences à l'œuvre tout au long du projet

Je n'avais pas trop l'habitude de l'analyse exploratoire des données mais je suis satisfait de mon travail, j'ai pu aller assez vite.

L'entraînement du modèle est allé assez vite aussi, j'ai utilisé des protocoles auxquels je suis habitué.

J'ai pu apprendre un peu mieux les API parce que je ne les ai pas encore beaucoup utilisées.

Je n'utiliserais pas de si tôt WSL, j'ai eu des bugs très compliqués que je n'ai pu gérer qu'en quittant / rebootant WSL.

Pour le côté base de données, SQLite fut très pratique pour mon cas d'usage : les databases sont sous forme de fichier et facilement exportables (pour les push et pull github par exemple).

Le côté déploiement web reste assez trouble pour moi, car étant payant. Aussi je n'ai pas développé dans Docker car je n'en ai pas eu l'intérêt direct.

### 7.1 Datasets de santé mentale

Je me suis renseigné sur les autres jeux de données de santé mentale en libre accès sur internet, j'ai été étonné du vide que j'y ai trouvé.

# 8. Bilan du projet et améliorations envisageables

Je suis content de moi, car j'ai parcouru les différentes compétences à valider et en étant en avance sur le planning, j'avais vu large. J'ai résolu les bugs un par un, sans me décourager.

Une amélioration envisageable pourrait être l'implémentation du niveau de certitude dans la prédiction de l'état de santé mentale. Ainsi l'utilisateur saurait si l'algorithme est sûr ou pas sûr

Je pense aussi que cette application est une bonne base pour développer de nouveaux outils pour prendre en charge la santé mentale des gens, en développant des chatbots par exemple.

# 9. Conclusion

J'ai beaucoup aimé la formation, et je suis très reconnaissant envers toute l'équipe qui m'a intégré dans l'école. Nous avons abordé beaucoup de choses durant la formation, même si parfois trop de choses ont été abordées et de manière trop brèves.

# Références

1. ***https://datascientest.com/algorithme-de-classification-definition-et-principaux-modeles.***

*2.* ***https://chat.openai.com/.***

# Annexes

[Annexe 1 : Rapports d'avancement 25](#_Toc161866885)

Annexe 1 : Rapports d'avancement

Les réunions se sont déroulées les jeudis, à 14 heures, avec le Pr René Gallimard et Pr Hervé Ducroc.

Rapport 1, 29 février

Présentation du fichier d'analyse exploratoire des données :

- Stratégie adoptée pour les valeurs manquantes, colonnes inutiles, lignes supprimées, simplification de la colonne "Genre".

- Graphiques de représentation des données et interprétation

- Résumé des stratégie adaptées (script python)

Nous nous sommes mis d'accord sur les stratégies adoptées de nettoyage de données, nous avons exploré ensemble les résultats de l'analyse exploratoire des données.

Rapport 2, 7 mars

Présentation de l'entraînement du modèle et des résultats :

- Vulgarisation du machine learning

- Présentation des résultats

- Présentation d'un test

Lors de l'analyse des résultats, les professeurs ont cherché à savoir le fonctionnement interne du meilleur modèle, à savoir régression logistique, je me suis donc informé et leur ai présenté son fonctionnement.

Rapport 3, 28 mars

Présentation de l'application fonctionnelle :

Nous avons vu ensemble si le produit fini était conforme aux attentes, nous avons discuté du fait d'intégrer un niveau de certitude dans les prédictions.