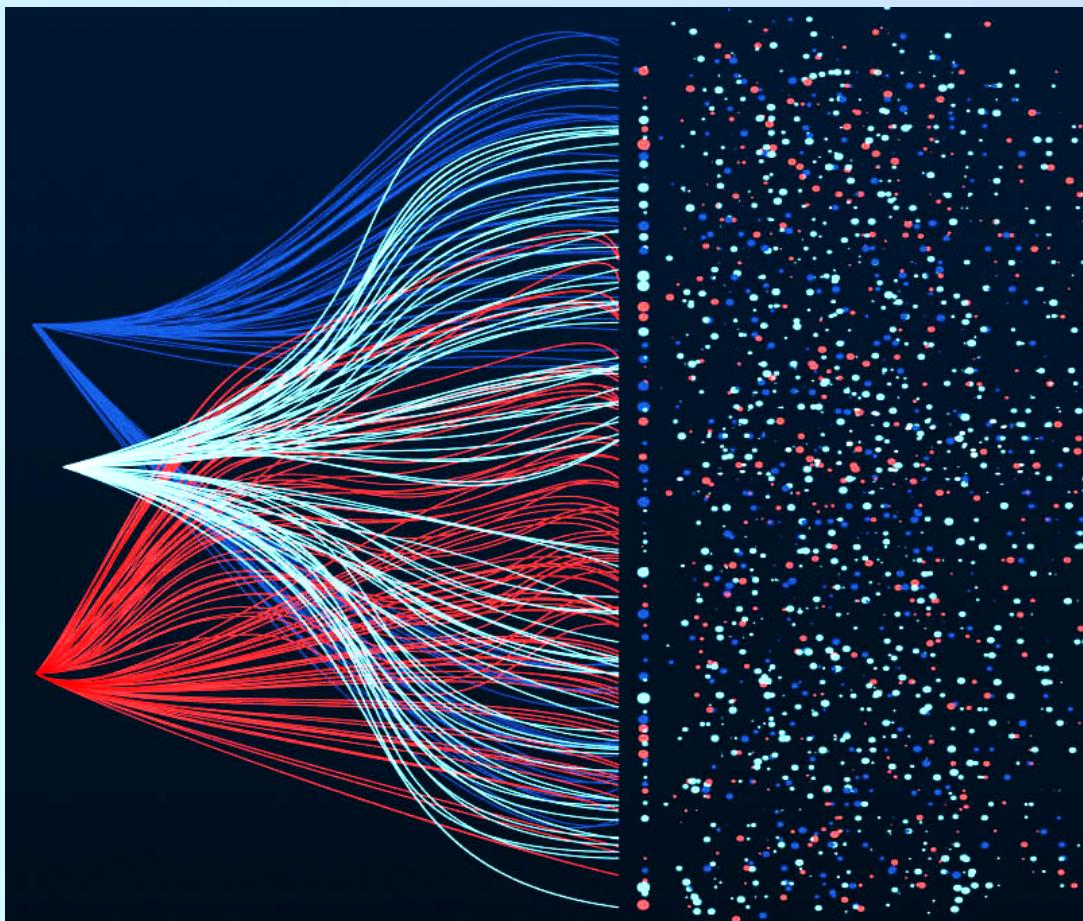


CAHIER DES CHARGES

**LA DATA AUGMENTATION AU
SERVICE DU MACHINE LEARNING
L3Q1-2022**



CAHIER DES CHARGES

PROJET DATA AUGMENTATION

Identification du document

Référence du document : L3Q1.A

Version du document : 3.0

Date du document : 15/02/2023

Auteurs :

Abou Assaf Mahyr-Florian

Daumont-Ouk Ilan'

Ngo Alexandre

Chemla Ethan

Sommaire

I - Introduction	4
1.1. Augmentation de la donnée	4
1.2. Concepts de base	5
1.3. Historique	6
1.4. Contexte	7
II - Description de la demande	8
2.1. Les objectifs	8
2.2. Produit du projet	8
2.3. Fonctionnalités essentielles	9
2.4. Fonctionnalités optionnelles	11
2.5. Fonctionnalités imaginées	11
2.6. Maquette du projet	12
2.7. Critères d'acceptabilité et de réception	14
III - Contraintes	15
3.1. Contraintes de coûts	15
3.2. Contraintes de langage	15
3.3. Contraintes d'environnement	17
3.4. Contrainte de délais	17
3.5. Contraintes de qualité	17
V. Organisation	17
5.1. Les phases du projet	18
VI. Glossaire	18
6.1. Définition	19
6.2. Références bibliographiques	19

I - Introduction

1.1. Augmentation de la donnée

L'augmentation de la donnée est un processus de collecte et d'enrichissement des données réelles, provenant de diverses sources afin de créer des informations plus globales, utiles et importantes pour la prise de décision. En augmentant les ensembles de données existants avec de nouvelles informations les entreprises peuvent mieux comprendre leurs clients, leurs marchés, leurs concurrents et leurs opérations.

Cela peut également s'inscrire dans une optique de donnée dirigée pour améliorer la qualité des résultats des modèles d'apprentissage automatique, en l'occurrence cela sera la direction prise pour notre projet.

L'augmentation de la donnée permet aux entreprises d'accroître leur capacité à extraire et intégrer efficacement des volumes importants de données à partir de sources variées afin d'obtenir une image pertinente et proche du monde réel pour alimenter leurs projets d'intelligence artificielle.

Il est important de noter que ces données obtenues ne sont pas réelles, il s'agit seulement de données synthétiques.

Dans un contexte où les projets d'intelligence artificielle et d'apprentissage automatique sont en pleine expansion, l'augmentation de la donnée est devenue une solution de plus en plus utilisée pour surmonter la rareté et la spécificité des données.

1.2. Concepts de base

L'intelligence artificielle est un domaine de l'informatique qui tente de reproduire le comportement humain. Elle utilise des algorithmes et des logiciels pour prendre des décisions, apprendre d'expériences passées et interagir avec des données. Les applications d'IA sont déjà largement utilisées dans les services informatiques tels que la reconnaissance vocale et image, l'analyse prédictive et la robotique.

Rareté et spécificité des données : il est souvent difficile et coûteux de collecter des données à grande échelle, ce qui peut entraver les projets d'IA. Les données rares et spécifiques sont particulièrement importantes dans les secteurs tels que la finance, l'automobile et le médical.

L'augmentation des données implique l'ajout d'informations supplémentaires à un ensemble de données existant afin d'améliorer sa qualité et sa précision.

Machine Learning : cette technique est utilisée pour amener les modèles à apprendre automatiquement en modifiant légèrement les exemples de données chaque fois que le modèle les traite.

L'application web et la base de données sont des termes liés étroitement qui décrivent un système informatique permettant aux utilisateurs d'accéder à des données à partir d'un site web. Les applications Web sont généralement construites sur une plateforme client-serveur, ce qui signifie que le client, où l'utilisateur final, interagit avec le serveur via une interface Web. Le serveur stocke et exécute tout le code, tandis que le client reçoit simplement les résultats. La base de données est une partie importante du système car elle stocke les informations accessibles par l'application web et permet aux utilisateurs de les consulter. Les bases de données peuvent prendre plusieurs formes, mais elles sont typiquement organisées en tables qui répertorient les différents champs et autres éléments contenus dans la base de données.

1.3. Historique

Le projet d'augmentation de la donnée s'inscrit dans un contexte historique plus large, celui du Big Data (ou mégadonnées). Le Big Data est une technologie qui permet aux organisations de collecter et d'analyser des données à grande échelle afin de prendre des décisions plus efficaces.

L'augmentation de la donnée a été mise en œuvre pour résoudre les problèmes liés à la rareté et à la spécificité des données qui sont des obstacles majeurs pour les projets d'intelligence artificielle. Grâce au Big data, les entreprises peuvent recueillir et traiter de vastes ensembles de données pour identifier les tendances, optimiser leurs performances et prendre des décisions plus judicieuses. L'augmentation de la donnée permet aux organisations d'accroître leur jeu de données en créant des données synthétiques qui peuvent être utilisées pour améliorer leurs modèles. De plus, cette technologie offre aux entreprises une meilleure compréhension du comportement humain et une meilleure prise en compte des facteurs contextuels lorsqu'elles prennent leurs décisions.

Ainsi, il est clair que l'augmentation de la donnée joue un rôle crucial dans le développement et l'amélioration continue des systèmes intelligents. Cette technologie permet ainsi la résolution de problèmes liés à la rareté et à la spécificité des données, qui sont des obstacles majeurs pour les projets d'intelligence artificielle.

1.4. Contexte

Les données sont devenues un élément essentiel pour les entreprises et les organisations. Cependant, l'augmentation des données peut entraîner des problèmes de qualité de données. Les principaux problèmes liés à la qualité des données comprennent les erreurs d'entrée manuelle, l'absence d'informations complètes, la duplication de données et des formats de données inconsistants. Ces problèmes peuvent avoir un impact négatif sur la prise de décision et le fonctionnement opérationnel. Par conséquent, il est important que les entreprises prennent des mesures pour garantir la qualité des données afin d'améliorer leur prise de décision et leur fonctionnement opérationnel. À noter également l'arrivée de nouvelles technologies d'intelligence artificielle telles que DALL-E 2, OpenAI, JasperAI, WriteSonic, etc... qui insufflent un nouvel élan à l'industrie de l'IA.

Notre projet s'inscrit donc dans ce contexte en proposant une solution qui permettra aux entreprises d'améliorer leurs modèles d'IA et de réduire les coûts de collecte de données. Nous travaillerons à la mise en place d'une application web qui sera facilement accessible aux utilisateurs, et qui sera dotée d'algorithmes performants pour assurer la qualité des données générées. Les données prises en compte seront des images. De plus, nous explorerons les différentes techniques d'augmentation de la donnée pour permettre à nos utilisateurs d'obtenir des données modifiées de haute qualité.

II - Description de la demande

2.1. Les objectifs

Dans le cadre de ce projet nous proposons de développer une application web capable de collecter un ensemble de données d'images, puis de générer des données légèrement modifiées de manière quantitative. Ces méthodes de génération de données se feront au moyen de l'implémentation d'algorithmes d'augmentation de la donnée du côté back-end tels que le bruit, la luminosité, le contraste, le recadrage, le flopping, le zoom. L'objectif est de retourner à l'utilisateur un ensemble de données augmenté d'images, ainsi beaucoup plus important que celui initialement fourni à l'application. Notre projet vise ainsi à aider les entreprises à améliorer leurs modèles d'IA en leur fournissant un dataset plus significatif et en réduisant les coûts liés à la collecte de données.

2.2. Produit du projet

L'application offrira une plateforme simple d'utilisation et rapide à prendre en main pour ses utilisateurs. Ce produit logiciel s'accompagne d'une base de données temporelle permettant l'authentification de l'utilisateur et le stockage de ses données. En effet, la récupération des données avant l'augmentation est primordiale afin que l'utilisateur puisse comparer le nouveau jeu de données avec celui fourni initialement.

Enfin, des outils de visualisation des données seront également présents pour s'assurer de l'intégrité des données. Nous souhaitons adapter notre produit logiciel dans un environnement Linux et sur Google Chrome dans un premier temps. De plus, le format de donnée augmentée sera les données de type png / jpg, qui sont des formats d'images idéals pour une utilisation sur le web.

2.3. Fonctionnalités essentielles

Voici les fonctionnalités que notre produit logiciel fournira :

La création d'un compte utilisateur :

Le processus de création de compte sera simple et rapide, en remplissant un formulaire d'authentification. Une fois que le compte sera créé, chaque utilisateur pourra se connecter à son propre espace personnel afin de garantir la sécurité et la confidentialité des données de chacun.

L'authentification :

L'authentification peut se faire par le biais d'un nom d'utilisateur et d'un mot de passe, ou par d'autres méthodes telles que l'authentification à deux facteurs. L'authentification est une étape importante pour garantir la sécurité et la protection des données. Elle permet d'assurer que seules les personnes autorisées ont accès aux données sensibles de l'application.

Le dépôt de dataset utilisateurs :

Le dépôt de dataset est une fonctionnalité clé de notre application web d'augmentation de la donnée. Une fois que l'utilisateur s'est authentifié, il pourra déposer son jeu de données en utilisant un bouton spécifique disponible dans son espace personnel. Ce bouton permettra à l'utilisateur de télécharger le fichier d'images depuis son ordinateur et de le stocker sur notre serveur.

Cette fonctionnalité permettra aux utilisateurs de stocker leurs données sur notre plateforme, de les utiliser pour les projets d'augmentation de données et de les récupérer facilement en cas de besoin.

Configuration de l'augmentation des données :

Pour offrir aux utilisateurs plus de flexibilité dans la modification de leurs images, l'application permettra la configuration de l'augmentation des données. Cela permettra aux utilisateurs de régler les paramètres tels que le bruitage, le contraste, la saturation, le recadrage et la luminosité pour produire des données augmentées qui répondent à leurs besoins spécifiques.

L'augmentation de données :

A partir du jeu de données fourni par l'utilisateur, l'application retournera à l'utilisateur un jeu de données issu du jeu de données initial et de la génération de données synthétiques à l'aide des configurations d'augmentation de la donnée données précédemment (plus précisément des données d'images).

Conservation des données :

La fonctionnalité de conservation des données permettra à nos utilisateurs de stocker leurs jeux de données modifiés ou augmentés en toute sécurité. Nous allons mettre en place une base de données robuste et performante pour stocker les données de manière organisée et cohérente.

Récupération des données augmentées :

Notre fonctionnalité de récupération des données augmentées permettra à l'utilisateur de facilement télécharger ses données depuis son espace personnel vers son ordinateur. Pour ce faire, nous mettrons en place une interface intuitive avec un bouton de téléchargement bien visible.

2.4. Fonctionnalités optionnelles

Historique des données augmentées :

Cette fonctionnalité permettra aux utilisateurs de visualiser toutes les données qu'ils ont générées et d'accéder à l'historique de leurs activités. Les utilisateurs pourront également utiliser de nouveau les données générées précédemment et les télécharger à tout moment. Cette fonctionnalité permettra de faciliter la gestion et l'organisation des données augmentées, et de mieux suivre les activités des utilisateurs.

Web responsive :

Il est prévu que l'application soit conçue de manière à être adaptée à toutes les tailles d'écran et à tous les appareils, ce qui permettra aux utilisateurs de profiter d'une expérience de qualité, quel que soit le dispositif utilisé.

2.5. Fonctionnalités imaginées

Voici les fonctionnalités envisagées pour le développement du projet dans une perspective post-licence :

Exigences de sécurité :

L'authentification à elle seule ne suffit pas à garantir la protection des données. D'autres mesures de sécurité, telles que le chiffrement des données, la gestion des autorisations d'accès et la surveillance des activités d'utilisation de l'application, doivent également être mises en place pour garantir la protection des données de manière globale.

Données dirigées:

Nous essaierons éventuellement de fournir les images augmentées à un modèle de machine learning et jugerons arbitrairement si les images générées à l'aide de l'IA, sont satisfaisantes ou non.

2.4. Maquette du projet

Page d'accueil :

DataFlow AI Amplifiez vos données, découvrez vos performances

Bienvenue sur DataFlow AI, votre solution pour augmenter vos jeux de données rapidement et efficacement. Avec notre application web, vous pouvez facilement étendre la quantité de données disponibles pour votre modèle d'IA, ce qui améliorera les résultats de vos outils de Deep Learning / Machine Learning. Rejoignez la révolution du Deep Learning avec DataFlow AI.

S'inscrire Ou Se connecter

Pourquoi utiliser DataFlow AI ?

Le monde de l'intelligence artificielle évolue rapidement et le Machine Learning / Deep Learning joue un rôle clé dans cette transformation. Cependant, pour obtenir des résultats précis et fiables avec ces techniques, il est souvent nécessaire d'avoir un grand jeu de données. C'est là que notre application web entre en jeu.

Notre application a été conçue pour aider les utilisateurs à augmenter leur jeu de données de manière simple et efficace, sans avoir besoin de connaissances en intelligence artificielle ou en développement informatique. Nous avons identifié les méthodes les plus avancées d'augmentation de la donnée et les avons intégrées dans notre application pour offrir un outil facile à utiliser pour nos utilisateurs.

Des Algorithmes Performants

Grâce à l'utilisation de techniques avancées d'augmentation de la donnée et à la mise en œuvre d'algorithmes performants, cette application garantit des résultats précis et fiables. Les algorithmes sont continuellement mis à jour pour s'assurer de leur pertinence et de leur efficacité. Les utilisateurs peuvent donc être sûrs que leur jeu de données sera traité de manière optimale, ce qui les aidera à atteindre leurs objectifs. En choisissant cette application, les utilisateurs peuvent être certains d'obtenir des résultats rapides, fiables et pertinents.

Amplifiez vos données, découvrez vos performances.
www.DataFlowAI.com
DataFlowAI

UNIVERSITÉ PARIS DESCARTES

Université de Paris

Page d'identification :

<- Retour à la page d'accueil

Identification

Adresse Mail

Mot de Passe

Se souvenir de moi

Se connecter

[Mot de passe oublié ?](#)

Vous n'avez pas de compte ? [S'inscrire](#)

Page d'inscription :

<- Retour à la page d'accueil

Inscrivez vous dès maintenant



Prénom:

Nom:

Email:

Mot de passe:

Confirmer le mot de passe:

Pays:

Genre:

J'ai lu et j'accepte les conditions générales d'utilisation

Créer mon compte

Page espace personnel :

The screenshot shows a web application interface titled "Data Augmentation". At the top right is a magnifying glass icon. On the left, a sidebar contains a "Se déconnecter" button and a list of links: "Présentation", "Mon compte", "Mes dépôts", and "Histoire de la Data Augmentation". Below this is a section for "Déposer votre dataset", followed by "Visualiser mon dataset" and "Supprimer mon dataset".

2.5. Critères d'acceptabilité et de réception

Pour mesurer si les objectifs de qualité sont atteints concernant une application web proposant des services d'augmentation de données, il convient d'utiliser des indicateurs précis tels que la fiabilité et la non spécificité des données générées (globalité des données). La fiabilité est utilisée pour déterminer l'exactitude des informations produites par l'application, la globalité des données quant à elle est nécessaire au bon fonctionnement des modèles d'apprentissage.

Ces indicateurs peuvent également être le nombre de plaintes des clients, leur satisfaction ou le taux d'erreur lors des transformations des données.

Les tests de performance peuvent aussi être mis en place pour vérifier si les performances attendues sont obtenues. Par exemple, un test peut être effectué pour mesurer le temps nécessaire à l'application pour répondre aux demandes entrantes ou le temps qu'il faut à l'application pour traiter les données.

Enfin nous essaierons éventuellement de fournir les images augmentées à un modèle de machine learning et jugerons arbitrairement si les images générées à l'aide de l'IA, sont satisfaisantes ou non.

III - Contraintes

3.1. Contraintes de coûts

- Les coûts d'infrastructure : il est nécessaire d'avoir un serveur pour héberger les données et pour effectuer toutes les opérations en back-end.
- Le coût d'hébergement : 0€ si l'on se contente de l'utilisation d'un serveur local Flask sinon pour le déploiement et l'hébergement de l'application web compter entre 70~80€ par an.

De plus, il est important de prévoir des coûts pour la sauvegarde des données, la maintenance et les mises à jour du serveur.

3.2. Contraintes de langage

Pour le développement du *Back-end* de l'application Web, nous utiliserons :

- *Flask* : Framework web Python qui fournit des outils et des fonctionnalités utiles pour créer des applications web en Python. Flask permet de se concentrer sur ce que les utilisateurs demandent et quelle sorte de réponse ils attendent, ce qui rend le processus de conception d'une application web plus simple. Nous avons décidé d'utiliser Flask plutôt que le framework Django car dans le contexte actuel, l'application web Flask équivalente est plus explicite.
Avec Flask, il est possible de créer des applications Web complètes avec des comptes utilisateurs, des modèles et des fichiers statiques.
- *Python* : Langage de programmation open source qui peut être utilisé pour construire des applications Web, des logiciels et même des systèmes intégrés. En outre, la communauté Python est très active et ses modules, frameworks et packages constituent une riche source d'informations sur l'utilisation de diverses technologies dont notamment le développement d'applications web avec *Flask*. Les principales caractéristiques de *Python* comprennent sa portabilité, son interprétation interactive, ses types dynamiques, sa gestion automatique de la mémoire et son support multi-plateforme.
Il est le langage principal dans le développement d'intelligences artificielles.

Pour le développement du *Front-end* nous utiliserons :

- HTML : Inévitable pour quiconque souhaite développer son site web, HTML est un langage de balisage qui permet aux développeurs web d'utiliser des mots-clés pour aider les navigateurs à structurer et afficher le contenu d'une page Web. Il fournit une structure pour le contenu, ce qui donne l'apparence d'un site Web aux utilisateurs finaux.
- CSS : Langage de feuilles de style qui aide les développeurs web à améliorer la présentation et l'apparence des sites Web. Il permet aux développeurs de créer des designs attrayants en séparant le contenu HTML du design. Les feuilles de style CSS sont importantes car elles permettent au navigateur d'appliquer un même style à plusieurs pages, ce qui rend les choses plus simples pour le développeur. Les avantages principaux d'utiliser CSS dans le développement d'une application Web comprennent: une meilleure organisation, une meilleure flexibilité et l'habileté de gérer et modifier facilement le style visuel global des applications.

Etant donné le fait que nous développons une application destinée aux utilisateurs, il est important de rendre agréable l'expérience utilisateur.

Concernant le développement de la base de donnée :

- InfluxDB : Système de gestion de base de données orientée séries temporelles hautes performances, écrit avec le langage de programmation Go, et distribué sous licence MIT.

Afin d'évaluer le taux d'acceptabilité des données générées nous utiliserons :

- TensorFlow : *TensorFlow* est une plateforme open source pour le machine learning qui permet aux développeurs de créer des modèles et des applications d'apprentissage automatique à l'aide de bibliothèques, outils et opérations. Il propose une flexibilité et une puissance inégalées pour les développeurs professionnels et amateurs. Grâce à un large éventail d'outils intuitifs pour l'entraînement, la validation et le déploiement des modèles, *TensorFlow* simplifie considérablement le processus de développement du machine learning. Il est également compatible avec *Python*

3.3. Constraintes d'environnement

Le développement du site sera sous le système d'exploitation Linux, qui est très performant pour les projets d'envergures, dispose d'une sécurité élevée, en open-source donc gratuit et facilement personnalisable. De plus, 75% des serveurs web utilisent linux (source : w3tech).

3.4. Contrainte de délais

Le projet a débuté le 23/01/2023 et le produit doit être livré d'ici le 10/04/2023.

Rendu final du cahier des charges : 17/02/2023

Rendu cahier de recettes et cahier de test : 24/02/2023

Du 13/02/2023 au 27/03/2023 : Développement

Semaine du 03/04/2023 : Phase d'intégration

10/04/2023 : Livraison du produit

Un intervalle de 12 semaines nous est imposé à raison de 12h de travail par développeur soit un total de 576 heures au total.

3.5. Constraintes de qualité

Les données générées devront être suffisamment proches de la réalité pour être acceptées. L'application web et chacune des fonctionnalités proposées doivent être parfaitement fonctionnelles.

V. Organisation

5.1. Les phases du projet

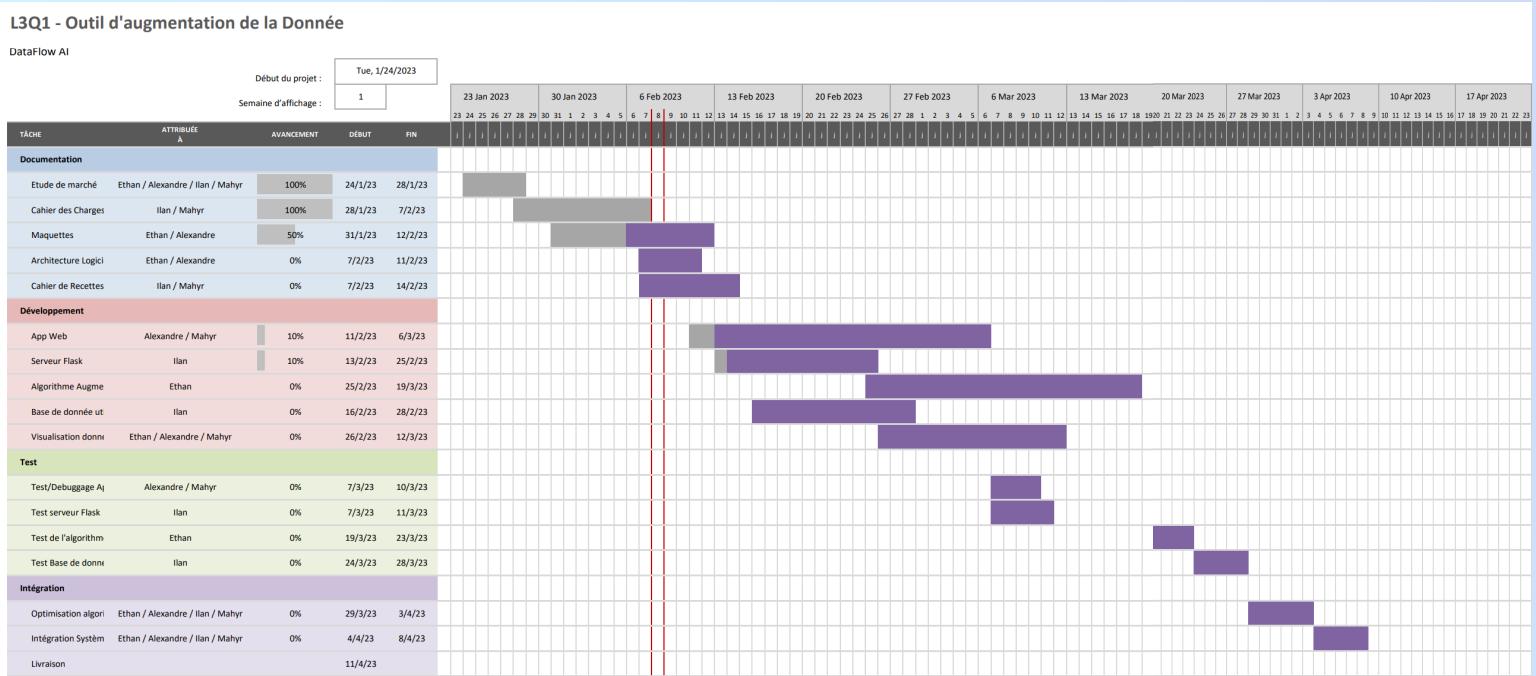


Diagramme de Gantt du projet

VI. Glossaire

6.1. Définition

Application web: Programme ou logiciel qui est accessible depuis différents navigateurs web.

Data Cleaning : Ensemble de processus ayant pour but d'améliorer la qualité des données en corrigeant les données inexactes dans un ensemble de données.

Dataset : Ensemble de données cohérentes qui peuvent se présenter sous différents formats (textes, chiffres, images, audios, vidéos...).

Données non structurées : Données non organisées en base de données, c'est-à-dire la messagerie, les images, les vidéos, etc...

Machine learning : Le Machine Learning ou apprentissage automatique est un domaine scientifique, et plus particulièrement **une sous-catégorie de l'intelligence artificielle**. Elle consiste à laisser des algorithmes découvrir des » patterns « , à savoir des motifs récurrents, dans les ensembles de données. Ces données peuvent être des chiffres, des mots, des images, des statistiques...

Deep learning : Il s'agit d'un type de machine learning, version plus améliorée. L'apprentissage profond utilise une technique lui conférant une aptitude supérieure à détecter les patterns même les plus subtiles.

Python : Langage de programmation orienté objet, notamment très utile dans le domaine de l'intelligence artificielle.

Flask : Un serveur Flask est un serveur web basé sur le framework Flask, qui permet de créer des applications web en Python. Flask est un framework web léger qui offre une grande flexibilité et une facilité de développement pour la création d'applications web.

6.2. Références bibliographiques

Introduction générale à la data augmentation :

<https://neovision.fr/data-augmentation-solutions-manque-donnees/>

https://en.wikipedia.org/wiki/Data_augmentation

<https://blog.datumize.com/the-five-most-common-data-quality-issues-and-how-to-overcome-them>

<https://www.cnil.fr/fr/definition/augmentation-de-donnees>

<https://lbourdois.github.io/blog/nlp/Data-augmentation-in-NLP/>

Présentation machine learning et deep learning :

<https://datascientest.com/machine-learning-tout-savoir>

https://inside-machinelearning.com/pourquoi-et-comment-normaliser-ces-donnees-pytorch-une-etape-essentielles-du-deep-learning-partie-1/#Normaliser_les_donnees