I. Introduction

Contexte et objectifs

À l'ère du numérique, des nouvelles technologies, les entreprises sont confrontées à une augmentation exponentielle des informations à traiter. Devant cette quantité considérable de données, l'Intelligence Artificielle devient une solution indispensable pour analyser et exploiter ces informations de manière efficace.

L'intelligence artificielle offre la possibilité d'automatiser des tâches complexes tout en offrant des informations précieuses pour faciliter la prise de décision. C'est dans ce contexte que notre équipe HeticData intervient.

Notre startup se concentre sur l'assistance et le conseil en matière de stratégie d'entreprise axée sur les données. Grâce à notre savoir-faire, nous proposons des solutions complètes qui englobent tout le domaine de la data : de la gestion de projet et de l'analyse des données, à la création de modules d'Intelligence Artificielle, jusqu'à leur intégration dans des tableaux de bord de visualisation pour faciliter la prise de décision.

Notre client, une entreprise majeure dans le domaine du commerce en ligne, a pris conscience de l'importance grandissante de l'Intelligence Artificielle et cherche à optimiser l'utilisation des grandes quantités de données clients qu'elle a collectées. Afin de découvrir les opportunités offertes par ses données, il nous a demandé des conseils stratégiques. Il souhaite donc évaluer nos compétences dans sa division des ventes afin de comprendre comment nos solutions peuvent optimiser ses opérations et renforcer son avantage du marché.

Périmètre du projet

Inclus dans le projet

1. Analyse des Données Clients
2. Développement de Modules d'Intelligence Artificielle
3. Intégration et Datavisualisation
4. Assistance et conseil stratégique

Exclu dans le projet

1. Optimisation Globale de l'E-commerce (Analyse uniquement sur la vente des produits électroniques)

Description du Projet

Problématique

Mettre en place une stratégie qui permet la conversion des visiteurs en client, de fidéliser la clientèle au travers du ciblage publicitaire.

En jeux

Booster le trafic des utilisateurs sur le site ainsi que le temps passé dessus, augmenter le taux de vente de produit sur le site.
Faire des prédictions sur les ventes de certains produits.

Présentation des données

Préparation et preprocessing sur notre le data set

Environnement de traitement de notre donnée

Plusieurs choix d'environnements de traitement et de la visualisation de la donnée s'offrent à nous notamment Excel, Locker Studio, PowerBI, Tableau, Python, streamlit etc…

Pour ce projet nous avons choisi de travailler avec Python notamment Jupyter Notebook et google colab.

Chargement et visualisation de la donnée

Traitement des doublons

Une première analyse consistait à visualiser la structure de notre base de données. En effet, on a remarqué qu'il existait des doublons et procéder à leur suppression.

```
Dimension de dataset avant duplicated: (885129, 9)
Dimension de dataset après duplicated: (884474, 9)
```

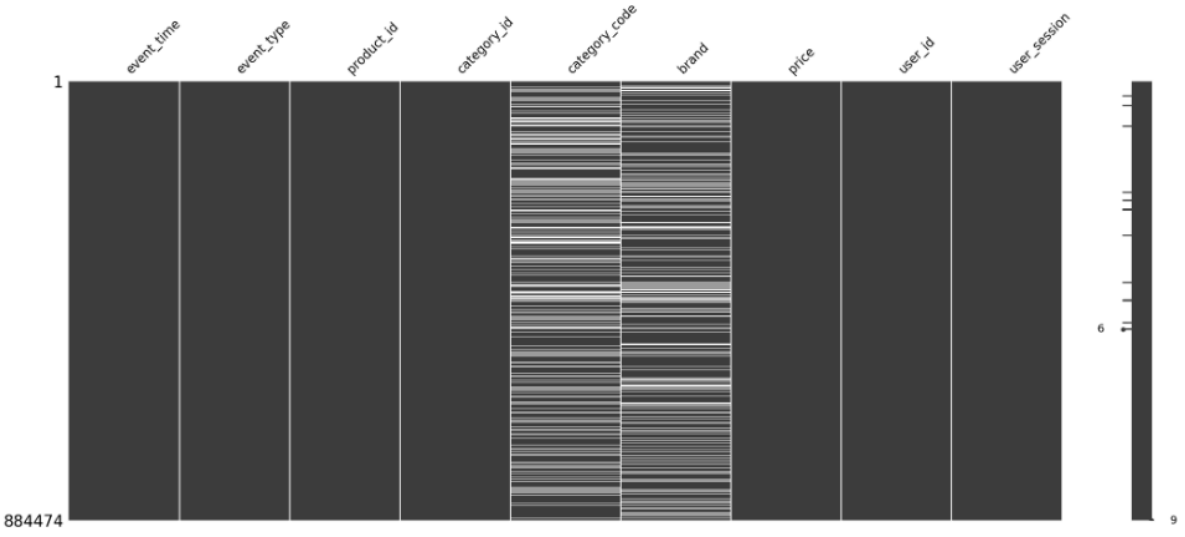
Traitement des valeurs manquantes

En statistique, on parle de valeurs manquantes lorsqu'on n'a pas d'observations pour une variable donnée pour un individu donné. Ces données ne peuvent pas être ignorées lors d'une analyse statistique. Les solutions possibles diffèrent selon la proportion et le type de ces valeurs. On pourra soit :

- Retirer les variables ou les individus présentant des données manquantes
- Imputer les valeurs

Notre jeu de données présente des valeurs manquantes sur 9 variables (voir graphe ci-dessous) avec un taux près de 50% dont.

```
category_code    26.687522
brand            23.992435
user_session      0.018641
event_time        0.000000
event_type        0.000000
product_id        0.000000
category_id       0.000000
price             0.000000
user_id           0.000000
dtype: float64%
```



Ce graphique ci-dessus représente les données manquantes des différentes variables de notre dataset.Les bandes noires traduisent les lignes contenant des données et celles grises l'absence d'observations. Comme on peut le voir, ce sont les variables category_code, brand et user_session contenant des valeurs manquantes.

Fonctionnalités et Exigences

Exigences Fonctionnelles

Les exigences fonctionnelles décrivent ce que le système doit faire.

Nous allons les structurer comme suit :

Tableaux de Bord (Dashboards) : en créant des tableaux de bord interactifs permettant de visualiser les données grâce au dataset pour affichager les données de manière dynamique sous différentes formes (mettre le nom de nos datas viz).

Rapports : Générer des rapports détaillés sur les données collectées grâce à l'interprétation de nos tableaux de bord en personnalisation les rapports en fonction des besoins (sections, graphiques, annotations).

Modèles de Prédiction : Implémenter des modèles de machine learning pour prévoir les tendances futures grâce à des modèles de régression et ou de classification. De mettre à jour régulière des modèles avec de nouvelles données.

Exigences Non-Fonctionnelles

Les exigences non-fonctionnelles définissent comment le système doit fonctionner. Pour cela on va parler de :

Performance : c'est-à-dire que le système doit être rapide et réactif ; les tableaux de bord interactifs et dynamiques et avoir la capacité de traiter les données de manière précises et compréhensible.

Sécurité : Garantir la confidentialité et l'intégrité des données.

Scalabilité : Le système doit pouvoir évoluer avec la croissance des données et des utilisateurs.

Exigences : l'architecture doit faciliter l'ajout de nouvelles fonctionnalités.

Critères de Réussite

Les critères de réussite sont des indicateurs mesurables pour évaluer le succès du projet comme les KIP suivants :

- Précision des Modèles de Prédiction
- Taux de Satisfaction

- Disponibilité du Système
- La latence

Technologie et Architecture

Environnement technique

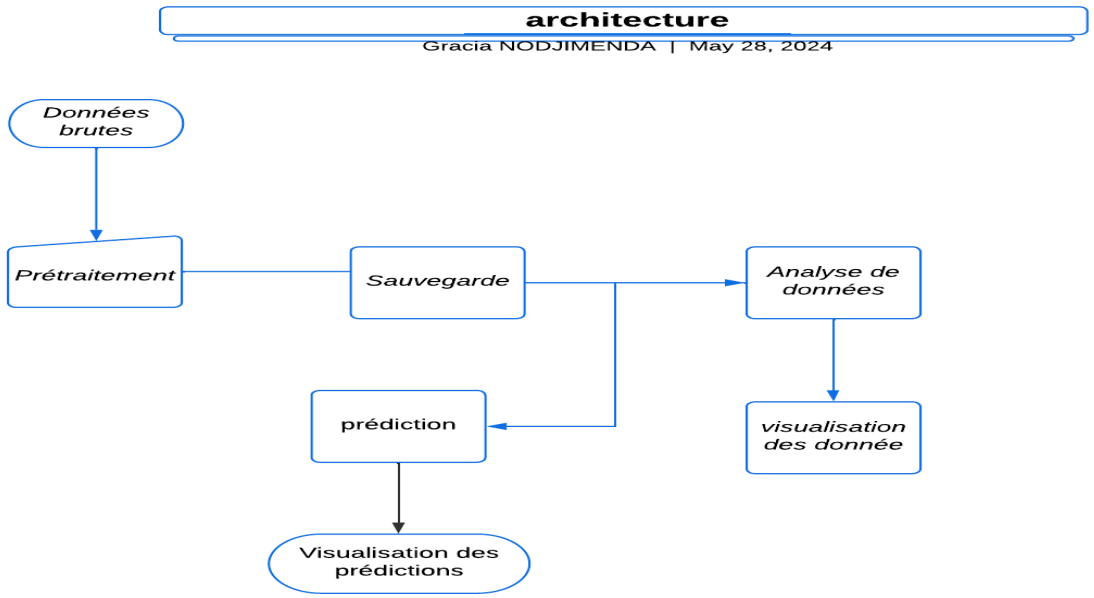
Nous analysons nos données en utilisant python sur un notebook. Nous avons choisi ces outils parce qu'ils offrent une flexibilité, une interactivité et une facilité de documentation qui sont particulièrement bien adaptées aux tâches d'analyse de données.

Pour la visualisation des données, notre choix est porté sur powerBI ;

Power BI est un outil versatile et puissant qui aide les entreprises à transformer leurs données en informations exploitables grâce à des visualisations interactives, une analyse en temps réel, et des capacités de collaboration avancées.

« A compléter avec les outils d'apprentissage »

Architecture



Sécurité et Confidentialité

Mesures de Sécurité

1. Politiques de Sécurité

Contrôle d'accès basé sur les rôles (RBAC) : Définir des rôles spécifiques avec des permissions adaptées pour chaque utilisateur ou groupe d'utilisateurs afin de restreindre l'accès aux données sensibles.

Authentification Multi-Facteurs (MFA) : Imposer l'utilisation de MFA pour l'accès aux systèmes Power BI pour renforcer la sécurité des comptes utilisateurs.

Politiques de sécurité des Informations : Élaborer des politiques écrites qui détaillent les pratiques de sécurité des données, la gestion des incidents, et la protection de la vie privée des utilisateurs.

2. Cryptage des Données

Cryptage au Repos : Utiliser des technologies de cryptage pour protéger les données stockées dans les bases de données et les entrepôts de données (ex. Azure SQL Database, Azure Data Lake).

Cryptage en Transit : Protéger les données en transit entre les utilisateurs et les services Power BI en utilisant des protocoles de sécurité tels que TLS (Transport Layer Security).

3. Accès Restreint

Contrôle d'accès granulaire : Implémenter des contrôles d'accès précis pour limiter l'accès aux données en fonction des besoins spécifiques des utilisateurs.

Groupes de Sécurité et Permissions : Utiliser des groupes de sécurité et des configurations de permissions dans Power BI pour gérer l'accès aux tableaux de bord et aux rapports.

Surveillance et Audit : Mettre en place des systèmes de journalisation et d'audit pour surveiller les accès aux données et détecter les activités suspectes.

Conformité Réglementaire

Réglementation Générale sur la Protection des Données (RGPD)

Consentement des Utilisateurs : Assurer que le consentement explicite des utilisateurs est obtenu avant de collecter et traiter leurs données personnelles.

Droit à l'Oubli : Implémenter des processus pour permettre aux utilisateurs de demander la suppression de leurs données personnelles conformément aux exigences du RGPD.

Portabilité des Données : Permettre aux utilisateurs de télécharger leurs données personnelles dans un format structuré et lisible par machine.

Notification de Violation de Données : Mettre en place des procédures pour notifier les autorités de protection des données et les utilisateurs concernés en cas de violation de données.

Autres Régulations et Normes

PCI-DSS (Payment Card Industry Data Security Standard) : Si des données de paiement sont traitées, suivre les normes PCI-DSS pour la sécurité des informations de carte de crédit.

ISO/IEC 27001 : Adopter des pratiques conformes à la norme internationale ISO/IEC 27001 pour la gestion de la sécurité de l'information.

Organisation et Gestion du Projet

Composition de l'équipe, rôles et responsabilités

Data Analyst

Nom **:** Blondy ULYSSE

Rôle **:** Analyse des données

Responsabilité **:**

- Participe à la rédaction du cahier des charges
- Collecte, nettoyage et prétraitement des données
- Analyse exploratoire des données clients
- Visualisation des données

Data Analyst

Nom **:** HEUMOU OSCAR

Rôle **:** Consultant data

Responsabilité **:**

- Création du repository Git pour le versionning du code
- Participe à la rédaction du cahier des charges
- Preprocessing des données
- Participe à la mise en place de l'application pour la visualisation.

Data Scientist

Nom **: Nodjimenda Gracia**

Rôle **:** Construction des modèles

Responsabilité **:**

- Développement de module d'apprentissage
- Nettoyage et prétraitement des données
- Étude la performance des modèles
- Participe à la mise en place de l'application pour la prédiction.

Data Scientist

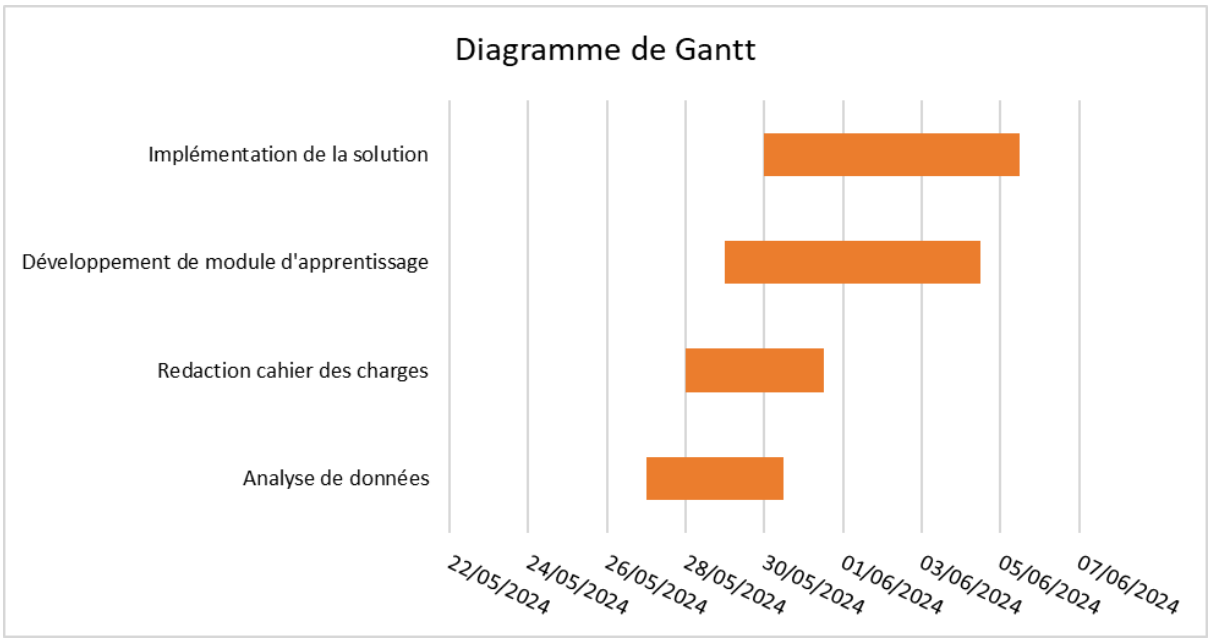
Nom **:Djibril DIOP**

Rôle **:** Construction des modèles

Responsabilité **:**

- Développement de module d'apprentissage
- Étude la performance des modèles
- Participe à la mise en place de l'application pour la prédiction.

Planning



Estimation des coûts et allocation des ressources.

Ressources (matériel et Logiciels)

- Ordinateurs et Matériel Informatique
    - Coût estimé : 0 € (Ressources personnelles existantes)
- Logiciels de Data Analysis (Python, Excel)
    - Coût estimé : Gratuit (open source)
- Logiciels de Visualisation de Données (Power BI, Streamlit)
    - Coût estimé : 0 € (Version étudiante gratuite, open source)

Risques et Contingences

Risque de qualité des données : Les données fournies par le client peuvent être incomplètes, inexactes ou non normalisées, ce qui pourrait fausser les résultats de l'analyse.

Plan de mitigation : Effectuer une analyse approfondie des données dès le début du projet pour identifier les lacunes et les erreurs. Travailler en étroite collaboration avec le client pour nettoyer et normaliser les données.

Risque de délai : Les délais peuvent être plus longs que prévu en raison de problèmes techniques ou de difficultés imprévues dans la mise en œuvre des modules d'apprentissage.

Plan de mitigation : Définir des étapes de projet claires avec des échéances intermédiaires pour suivre la progression.

Risque de conformité RGPD : Le traitement des données personnelles des clients doit être conforme au RGPD (Règlement général sur la protection des données).

Plan de mitigation : Implémenter des mesures de sécurité strictes pour protéger les données personnelles. Travailler avec des experts juridiques pour s'assurer que toutes les activités de traitement des données sont conformes aux réglementations en vigueur.

Risque de communication : Des malentendus ou des lacunes dans la communication avec le client peuvent entraîner des divergences par rapport aux attentes.

Plan de mitigation : Organiser des réunions régulières avec le client pour discuter de la progression du projet et des résultats intermédiaires. Utiliser des outils de gestion de projet pour suivre les tâches et les communications.

Risque de performance des modèles d'apprentissage : Les modèles d'apprentissage développés peuvent ne pas atteindre les performances souhaitées en raison de la complexité des données ou d'autres facteurs imprévus.

Plan de mitigation : Effectuer des tests rigoureux des modèles sur des ensembles de données de validation pour évaluer leurs performances. Être prêt à ajuster et à affiner les modèles en fonction des résultats des tests.