Objectif

L'objectif principal est de résoudre une étude de cas en science des données.

Il s'agit d'un travail qui a pour objectif d'évaluer les compétences suivantes:

- Fondement de la programmation
- Statistiques et mathématiques
- Gestion des données
- Base de données et SQL
- Visualisation des données
- Fondements de l'apprentissage automatique
- Débrouillardise et autonomie
- Capacités de résolution de problèmes
- Compétences en communication / vulgarisation

❑ **Instructions :** Bienvenue à l'épreuve en Intelligence Artificielle! Ton prochain défi sera de nous démontrer certaines de tes compétences au moyen d'un court cas en science des données. Nous t'avons transmis un

jeu de données provenant de la plateforme Kickstarter (voir ficher « ks_dataset.csv »). En utilisant ces données, nous t'invitons à résoudre la problématique ci-dessous. Une fois résolue, renvoie-nous tes réponses aux questions ainsi que le code que tu as produit. Bonne chance!

Lien vers les données (mot de passe: moovai) **CLIQUEZ ICI**

❑ **Mise en situation :** Certains promoteurs de projet tentent de comprendre comment augmenter le taux de réussite de leurs futures campagnes. Ils ont à leur disposition des données historiques de campagnes annoncées sur la plateforme Kickstarter.

❑ **Objectif :** Développe en Python une approche ML (supervisée et/ou non supervisée) pour aider les promoteurs de projet à lancer des campagnes à fort potentiel de réussite.

Étude de Cas

- ❑ **Question #1 : Préparation des données (20 points)**
  - ❑ 1.1. Si tu rencontres des problèmes de qualité des données durant ta manipulation des données de Kickstarter que nous t'avons fourni (indice: nous te confirmons que les données ont été corrompues 😉), comment les as-tu résolus ? Précise les étapes spécifiques que tu as suivies pour préparer les données
  - ❑ 1.2. Fournis une visualisation (graphique) illustrant l'impact de la qualité des données sur les performances du modèle. Comment ces problèmes de qualité ont-ils été gérés ?

- ❑ **Question #2 : Insights et Caractéristiques (30 points)**
  - ❑ 2.1. Identifie trois "insights" pertinents liés au succès ou à l'échec des campagnes Kickstarter. Fournis une visualisation pour chaque insight.
  - ❑ 2.2. Discute des « variables de confusion (ou confondantes)[1] » qui pourraient affecter l'interprétation de ces observations. Comment les as-tu pris en compte dans ton analyse ?
  - ❑ 2.3. Comment ces "insights" pourraient être transformés en variables pour faciliter l'apprentissage d'un modèle ML ?

---

[1] *Les variables de confusion (ou confondantes) sont des facteurs ou des variables dans une étude de recherche qui sont liés à la fois à la variable indépendante (le facteur étudié) et à la variable dépendante (le résultat), ce qui peut conduire à une interprétation fausse ou trompeuse de la relation entre la variable indépendante et la variable dépendante. Les variables de confusion peuvent créer l'illusion d'une relation de cause à effet alors que, en réalité, l'association observée est influencée par un facteur externe.*

Étude de Cas

- ❑ **Question #3 : Modèle ML et Impact Commercial (30 points)**
  - ❑ 3.1. Propose une approche ML pour prédire le succès des campagnes Kickstarter en utilisant les données fournies. Explique les types de modèles, les hyperparamètres, et la validation croisée que tu utiliserais.
  - ❑ 3.2. Comment interprètes-tu les résultats produits par ta solution ML en termes de succès des campagnes ? Comment cette solution ajoute-t-elle de la valeur pour les promoteurs de projets sur Kickstarter ?
  - ❑ 3.3. Comment envisages-tu que les parties prenantes vont utiliser ta solution pour comprendre comment lancer des campagnes à haut taux de succès ? Fournis des exemples d'utilisation dans un contexte commercial.

- ❑ **Question #4 : Maintenance du Modèle (20 points)**
  - ❑ 4.1. Imaginons que ta solution est déployée et roule maintenant en production. Tu remarques que la performance de ton modèle se dégrade progressivement depuis les derniers mois. De plus, tu identifies également certaines variables dont les valeurs semblent avoir évolué durant la même période. Selon toi, quel serait une raison qui explique cette situation et comment la résoudrais-tu ?
- ❑ **Format et Soumission**
  - ❑ Renvoie tes réponses aux questions ainsi que le code que tu as produit dans un document PDF ou dans un Jupyter Notebook.

Étude de Cas

- ❑ Assure-toi d'inclure des commentaires explicatifs dans ton code pour faciliter la compréhension et porte attention à ce que nous soyons en mesure de reproduire tes résultats. Le document doit être structuré de manière claire et inclure des graphiques pertinents.
- ❑ Nous évaluerons ta compréhension des concepts de science des données, ta capacité à résoudre des problèmes concrets et ta créativité dans la résolution de la problématique.
- ❑ Nous t'encourageons à aller valider directement sur le site web de **Kickstarter** ta compréhension de la signification des variables si besoin, en plus de te référer à l'annexe #1. Bonne chance !

Annexe #1 - Kickstarter

Kickstarter est une plateforme de financement participatif en ligne qui permet aux créateurs de projets de collecter des fonds auprès du grand public pour financer leurs idées, projets artistiques, innovations, produits ou initiatives. Elle a été fondée en 2009 à Brooklyn, New York, et est devenue l'une des plateformes de crowdfunding les plus populaires et les plus connues au monde.

Le fonctionnement de Kickstarter est le suivant :

- Les créateurs de projets proposent une description détaillée de leur idée ou projet sur la plateforme Kickstarter, y compris leurs objectifs financiers et le calendrier du projet.
- Les créateurs fixent un objectif de financement (un montant de fonds à atteindre) et une durée de campagne pendant laquelle les contributions peuvent être faites.
- Les personnes intéressées par le projet (les "backers") peuvent soutenir financièrement le projet en faisant des contributions, appelées "pledges". Ces contributions peuvent être de différentes natures, allant des montants symboliques aux investissements plus importants.
- Si le projet atteint ou dépasse son objectif de financement dans le délai imparti, les fonds collectés sont remis aux créateurs pour qu'ils puissent réaliser leur projet. Sinon, les contributions sont généralement remboursées aux "backers") .

Kickstarter est principalement utilisé pour financer des projets créatifs et artistiques tels que des films, de la musique, des jeux, des livres, des œuvres d'art, mais il est également utilisé pour des projets technologiques, des initiatives humanitaires, des

inventions, et plus encore. La plateforme a permis à de nombreuses idées innovantes de devenir réalité en permettant aux créateurs de trouver un financement direct auprès du public, contournant ainsi les canaux de financement traditionnels.

Annexe #2 - Documentation des Données

Le jeu de données (« ks_dataset.csv ») contient les informations suivantes

Variable	Description	Type
ID	L'identifiant unique du projet correspondant. Par exemple, "1000014025".	string
name	Le nom du projet correspondant. Par exemple, "Monarch Espresso Bar".	string
main_category	La catégorie principale dans laquelle le projet s'inscrit. Par exemple, "Poésie", "Alimentation", "Musique« , etc.	string
category	Une description plus précise de la catégorie principale. Sous-groupe de la catégorie principale (voir 2.). Par exemple, "Boissons" serait un sous-groupe de la catégorie "Alimentation" de l'attribut catégorie principale.	string
currency	La devise du projet (par exemple, USD ou GBP).	float
deadline	La date limite du projet.	time
goal	Montant en monnaie locale demandé initialement par le projet	float
launched	La date de lancement du projet.	time
pledged	Montant en monnaie locale que le projet a réalisé à la date limite.	float

state	Le projet a-t-il été couronné de succès à la fin de la journée ? L'état est une variable catégorielle divisée en niveaux : succès, échec, en cours, annulé, indéfini et suspendu.	string
backers	Le nombre de supporters qui ont investi dans le projet.	int
country	Pays d'origine du projet.	string
usd pledge	Montant en USD que le projet a réalisé à la date limite.	float

Critères d’évaluations

Aptitudes Techniques

- ❑ Traitement des données : Évalue la capacité de l'équipe à nettoyer, prétraiter et transformer efficacement les données brutes.
- ❑ Feature Engineering: Évalue la créativité et l'efficacité des techniques de Feature Engineering. utilisées pour améliorer les performances du modèle.
- ❑ Sélection du modèle : Considère la pertinence des modèles d'apprentissage automatique ou statistiques choisis pour le problème et les données.
- ❑ Performance du modèle : Évalue l'exactitude prédictive du modèle, la précision, le rappel, le score F1 ou d'autres mesures pertinentes.
- ❑ Visualisation des données : Évalue la clarté et la pertinence des visualisations des données pour transmettre des informations.
- ❑ Qualité du code : Organisation, lisibilité et documentation du code.

Résolution de Problèmes

- ❑ Compréhension du Problème : Évalue à quel point l'équipe comprend le problème, ses subtilités et sa pertinence pour les scénarios du monde réel.
- ❑ Créativité et Innovation : Évalue si l'équipe a proposé des solutions ou des approches innovantes au problème.

- ❑ Robustesse : Considère à quel point la solution s'adapte aux circonstances changeantes ou aux défis supplémentaires présentés pendant la compétition.
- ❑ Débrouillardise : Évalue à quel point les candidats s'adaptent aux changements inattendus ou aux défis présentés lors du concours

.