Application data sous forme de data storytelling

Type :	**PROJET**
Formations :	Ynov Informatique
Promotions :	Bachelor 3
UF :	SPE Data & IA

CADRE DU PROJET

Ce projet permet d'évaluer les compétences acquises dans le cadre de l'UF « Spécialité Data ». Le projet consiste à réaliser une application data exploitant des analyses de données, synthétisées sous forme de visualisations pour raconter une histoire.

- Équipe : 2 personnes
- Technologies : Utilisation obligatoire de Python, complété par SQL, R, Javascript, etc.
- Bonus : Méthodes d'analyse ou de visualisation avancées, déploiement en ligne.
- Objectifs de Formation Visés : Exploration & Analyse de Données :

Si vous n'avez pas d'idée de projet, vous avez le choix parmi une liste de projets proposés dans la partie « Projets au choix ».

Un bonus sera apporté aux projets personnels et aux groupes qui se challengent en proposant des fonctionnalités plus poussées.

Il vous appartient d'effectuer les recherches par vous-même pour trouver les ressources à la réalisation du projet.

Date de début : …….

Date de rendu : …….

OBJECTIFS DE FORMATION VISÉS

devrez mettre en application les principales compétences acquises durant les différents modules de la formation, à savoir :

PRINCIPES DE L'EXPLORATION & ANALYSE DE DONNEES

- Savoir **acquérir et structurer** des données pertinentes en utilisant des données ouvertes (open data) et/ou des méthodes de web scraping
- Vous Réaliser une **préparation des données** (formatage des différents types de données, gestion des valeurs manquantes, des doublons, des valeurs aberrantes) afin de les rendre exploitables par des méthodes d'analyses.
- Réaliser une **analyse exploratoire de données** afin de mettre en évidence les informations contenues dans les données, savoir les synthétiser sous forme de graphiques et identifier des méthodes d'analyses supplémentaires à faire.
- Savoir **interpréter vos résultats** et les **synthétiser** en utilisant des graphiques appropriés et des méthodes de ***data story telling*** pour présenter de manière synthétique vos conclusions.

MATHEMATIQUES POUR LA DATA SCIENCE

- Utiliser des métriques pour **quantifier la quantité et la qualité de données** présente dans les échantillons que vous aurez sélectionnés (indicateurs statistiques descriptifs, statistique inférentielle)
- Utiliser vos connaissances en **statistiques** et **probabilités** afin d'établir des **indicateurs** univariés et multivariés, pour la préparation et les analyses de données à réaliser
- Utiliser des méthodes d'analyses de données vues en cours et savoir rechercher de nouvelles méthodes qui pourraient être appropriées à votre sujet (modèles statistiques, modèles de machine learning)
- Savoir quantifier la qualité de vos analyses et de vos modèles.

MACHINE LEARNING

- Savoir appliquer **au moins un modèle** de machine learning dans vos analyses en fonction de la catégorie de problème à résoudre (classification, régression, clustering)

PYTHON POUR LA DATA SCIENCE

- Savoir implémenter, structurer et documenter du code pour les différentes étapes du projet
- Savoir utiliser des librairies appropriées de l'écosystème python pour la data science.

Vous pouvez utiliser les librairies de votre choix, mais nous vous recommandons les librairies suivantes :

- Acquisition, structuration et analyse exploratoire de données : pandas, numpy
- Modélisation statistiques et machine learning : scikit-learn, statsmodels
- Exploration et visualisation de données : matplotlib et/ou seaborn
- Outils orientés visualisation de données : Dash (de la suite logicielle Plotly) , Panel (de la suite logicielleHoloviz), Bokeh
- Outils orientés ''*interface application data*'' :
  Streamlit, Anvil,

LIVRABLES

Pour chaque groupe, vous devrez livrer les éléments suivants :

- Un dépôt Git accessible en ligne (par exemple via Gitlab ou Github) contenant tout le code et la documentation produits pour le projet.
- Un document au format Jupyter Notebook (ou équivalent) retraçant votre démarche et les analyses exploratoires que vous aurez mises en place durant le projet.
- Votre application data finale, synthétisant sous forme graphique vos analyses. Vous devrez présenter cette application, déployée à minima en local sur votre machine, pendant la soutenance.

MODALITÉS D'ÉVALUATION DU PROJET

Vous serez évalué sur l'ensemble des productions. L'évaluation prendra aussi la forme d'une présentation orale de synthèse d'environ 15 minutes accompagnée d'un support de présentation et d'une démonstration des fonctionnalités du site mises en place.

Le jury sera composé d'une partie des intervenants des cours de l'UF « SPÉ Data ».

Un temps de questions-réponses d'une durée de 5 minutes sera prévu à l'issue des 15 minutes.

Des évaluations intermédiaires auront également lieu au cours du déroulement du Projet.

2 points bonus seront ajoutés à la note finale si vous choisissez de réaliser un projet personnel. Dans le cas d'un projet proposé, des points bonus seront accordés en fonction de la difficulté du projet choisi :

Difficulté : 1	*0 point bonus*
Difficulté : 2	*1 point bonus*
Difficulté : 3	*2 points bonus*

Toutefois, les points bonus ne seront accordés que si le projet est fonctionnel.

DESCRIPTIF DU PROJET

Vous avez la possibilité de choisir entre un projet personnel ou un projet proposé.

Vous trouverez ci-dessous une liste des différents sujets qui pourront être abordé dans les projets. Suivi de la liste de projets au choix, si vous n'avez pas d'idée :

LISTE DES PROJETS AU CHOIX :

1er PROJET PERSONNEL :

Le projet personnel devra être valider par l'établissement. Vous pouvez vous référer à la liste de thème énoncé ci-dessus ou à la liste de projet ci-dessous pour vous faire une idée.

Projet 2: Prédiction des Ventes pour Yshop (Difficulté : 2)

Entreprise fictive : Yshop

Présentation:

Yshop est une petite boutique en ligne spécialisée dans les produits artisanaux. En raison d'une forte croissance, l'entreprise souhaite utiliser des techniques de machine learning pour prédire les ventes futures de ses produits et optimiser ses stocks, réduisant ainsi les coûts et maximisant les profits.

Objectifs:

- Collecter et analyser les données de ventes historiques.
- Développer un modèle de prédiction des ventes.
- Visualiser les prédictions et les tendances de ventes.

Tâches à réaliser:

1. **Acquisition et préparation des données**
   - Collecter les données de ventes historiques via des sources open data (Kaggle Dataset).
   - Nettoyer les données (gestion des valeurs manquantes, des doublons, des valeurs aberrantes).
2. **Analyse exploratoire des données**
   - Réaliser des visualisations des ventes par période, catégorie de produit (Matplotlib, Seaborn).
   - Identifier les tendances et les anomalies dans les données.
3. **Modélisation prédictive**
   - Appliquer des modèles de machine learning simples (régression linéaire, forêts aléatoires) avec Scikit-learn.
   - Évaluer la performance des modèles (métriques de performance : RMSE, MAE).
4. **Data storytelling et visualisations**
   - Créer des graphiques interactifs avec Plotly ou Dash pour présenter les résultats des analyses et des prédictions.
   - Développer une interface utilisateur basique pour visualiser les tendances et les prédictions.
5. **Déploiement de l'application**
   - Déployer l'application localement.
   - Documenter le processus de déploiement.

Livrables:

- Dépôt Git avec tout le code et la documentation.
- Jupyter Notebook retraçant la démarche et les analyses.
- Application de data storytelling déployée localement.

Projet 3 : Détection des Tendances sur les Réseaux Sociaux avec TrendSpotter (Difficulté : 3)

Entreprise fictive : TrendSpotter

Présentation:

TrendSpotter est une entreprise fictive spécialisée dans l'analyse des réseaux sociaux pour détecter les tendances émergentes. L'objectif est d'analyser les données de Twitter pour comprendre les préférences des utilisateurs et anticiper les tendances futures.

Objectifs:

- Collecter et analyser les données de Twitter.
- Développer des modèles de détection des tendances et d'analyse de sentiments.
- Visualiser les résultats de l'analyse.

Tâches à réaliser:

1. **Acquisition et préparation des données**

- Utiliser l'API Twitter pour collecter des données (tweets) sur des sujets spécifiques (Twitter API Documentation).
- Nettoyer et structurer les données pour les analyses.

2. **Analyse exploratoire et textuelle des données**

- Réaliser des visualisations des données collectées (Matplotlib, Seaborn).
- Utiliser des techniques de traitement du langage naturel (NLP) avec NLTK ou SpaCy pour analyser les sentiments.

3. **Modélisation prédictive**

- Appliquer des modèles de machine learning pour détecter les tendances (clustering avec K-means, régression logistique).
- Évaluer la performance des modèles.

4. **Data storytelling et visualisations**

- Créer des graphiques interactifs avec Plotly ou Dash pour présenter les résultats des analyses et des prédictions.
- Développer une interface utilisateur basique pour visualiser les tendances et les sentiments.

5. **Déploiement de l'application**

- Déployer l'application localement.
- Documenter le processus de déploiement.

Livrables:

- Dépôt Git avec tout le code et la documentation.
- Jupyter Notebook retraçant la démarche et les analyses.
- Application de data storytelling déployée localement.

Projet 4: Système de Recommandation pour YbookRecommender (Difficulté : 2)

Entreprise fictive : YbookRecommender

Présentation:

YBookRecommender est une bibliothèque en ligne fictive qui souhaite offrir une expérience personnalisée à ses utilisateurs en développant un système de recommandation de livres basé sur les préférences et l'historique de lecture.

Objectifs:

- Collecter et analyser les données d'utilisation et de notation des utilisateurs.
- Développer un système de recommandation personnalisé.
- Visualiser les recommandations et les comportements des utilisateurs.

Tâches à réaliser:

1. **Acquisition et préparation des données**

- Utiliser des données publiques de notation de livres disponibles en ligne (Goodreads Dataset).
- Nettoyer et structurer les données pour les analyses.

2. **Analyse exploratoire des données**

- Réaliser des visualisations des comportements des utilisateurs (Matplotlib, Seaborn).
- Identifier les patterns et les préférences des utilisateurs.

3. **Développement du système de recommandation**

- Appliquer des techniques de filtrage collaboratif (utiliser les bibliothèques comme Surprise).
- Évaluer la performance des modèles de recommandation.

4. **Data storytelling et visualisations**

- Créer des graphiques interactifs avec Plotly ou Dash pour présenter les résultats des recommandations.
- Développer une interface utilisateur basique pour visualiser les recommandations personnalisées.

5. **Déploiement de l'application**

- Déployer l'application localement.

- Documenter le processus de déploiement.

Livrables :

- Dépôt Git avec tout le code et la documentation.
- Jupyter Notebook retraçant la démarche et les analyses.
- Application de data storytelling déployée localement.

Projet 5: Prévisions Météorologiques avec WeatherForYnov (Difficulté : 2)

Entreprise fictive : WeatherForYnov

Présentation:

WeatherForYnov, est une entreprise fictive qui utilise des techniques de machine learning pour analyser les données climatiques et fournir des prévisions météorologiques précises.

Objectifs:

- Collecter et analyser les données climatiques historiques.
- Développer des modèles de prévisions météorologiques.
- Visualiser les résultats des analyses et des prévisions.

Tâches à réaliser:

1. **Acquisition et préparation des données**

- Utiliser des données climatiques publiques disponibles en ligne (NOAA Climate Data)
- Nettoyer et structurer les données pour les analyses.

2. **Analyse exploratoire des données**

- Réaliser des visualisations des données climatiques (Matplotlib, Seaborn).
- Identifier les tendances et les patterns dans les données.

3. **Modélisation prédictive**

- Appliquer des modèles de machine learning pour la prévision météorologique (régression linéaire, séries temporelles avec ARIMA).
- Évaluer la performance des modèles.

4. **Data storytelling et visualisations**

- Créer des graphiques interactifs avec Plotly ou Dash pour présenter les résultats des analyses et des prévisions.
- Développer une interface utilisateur basique pour visualiser les prévisions météorologiques.

5. **Déploiement de l'application**

- Déployer l'application localement.
- Documenter le processus de déploiement.

Livrables:

- Dépôt Git avec tout le code et la documentation.
- Jupyter Notebook retraçant la démarche et les analyses.
- Application de data storytelling déployée localement.

Projet 6: ChatBot Intelligent pour Support Client avec Ysupport (Difficulté : 3)

Entreprise fictive : Ysupport

Présentation:

Ysupport est une entreprise fictive qui souhaite développer un chatbot intelligent capable de fournir un support client automatisé en utilisant des modèles de langage avancés comme GPT-4.

Objectifs :

- Collecter et analyser les données des interactions clients.
- Développer un chatbot intelligent pour répondre aux questions des clients.
- Intégrer des fonctionnalités de machine learning et IA pour améliorer les réponses.

Tâches à réaliser:

1. **Acquisition et préparation des données**

- Collecter les données des interactions clients via des logs de chat ou des tickets de support.
- Nettoyer et structurer les données pour les analyses.

2. **Analyse exploratoire des données**

- Réaliser des visualisations des questions fréquentes et des types de réponses (Matplotlib, Seaborn).
- Identifier les patterns et les tendances dans les interactions.

3. **Développement du chatbot**

- Utiliser un modèle de langage avancé (GPT-4) pour développer le chatbot.
- Intégrer des fonctionnalités de machine learning pour améliorer les réponses basées sur les interactions passées.

4. **Data storytelling et visualisations**

- Créer des graphiques interactifs avec Plotly ou Dash pour présenter les résultats des analyses et des performances du chatbot.
- Développer une interface utilisateur basique pour visualiser les interactions et les performances du chatbot.

5. **Déploiement de l'application**

- Déployer l'application en ligne pour un accès en temps réel.
- Documenter le processus de déploiement.

Livrables:

- Dépôt Git avec tout le code et la documentation.
- Jupyter Notebook retraçant la démarche et les analyses.
- Application de data storytelling déployée en ligne.

Barème commun

Documentation de l'architecture	**4**	
	Clarté et précision de la documentation	2
	Pertinence des choix technologiques et justification	2
Mise en œuvre technique	**6**	
	Fonctionnalité des configurations et déploiements	3
	Innovation et complexité technique	3
Sécurité	**4**	
	Adéquation des mesures de sécurité mises en place	2
	Conformité avec les normes et régulations	2
Livrables et présentations	**4**	
	Qualité et utilité des livrables	2
	Clarté et professionnalisme de la présentation orale et de la démonstration	2
Gestion de projet	**2**	
	Respect des délais et des étapes du projet	1
	Qualité de la collaboration et de la communication au sein de l'équipe	1

RESSOURCES complémentaires

Voici quelques exemples (non exhaustifs) de réalisations qui pourront vous aider à cerner le travail attendu et aussi vous inspirer :

Pour le travail d'analyse exploratoire :

- Ce notebook en ligne, sur l'analyse de communautés est un très bon exemple d'analyse bien menée

- Ce notebook en ligne, propose des idées d'analyse et de graphiques utilisables pour l'analyse de données sur le thème du covid 19

Plus généralement, vous trouverez de nombreux exemples d'analyse exploratoire sur le site de kaggle

Pour le rendu final sous forme de data storytelling :

- Voici un exemple de data storytelling sur le sujet des élections américaines qui vous donnera une idée de ce que vous pourriez réaliser
- Le site OurWorldinData donne des exemples intéressants de visualisations de données sur des sujets utilisant des données ouvertes, en particulier sur le covid-19

Autres ressources d'inspiration :

- Le blog storytellingwithdata est souvent cité comme référence concernant les bonnes pratiques du data storytelling
- Certains journaux comme le New York Times et le Washington publient régulièrement des articles sous forme de data storytelling, comme par exemple cet article