Application data sous forme de data storytelling

Type : PROJET

Formations : Ynov Informatique

Promotions : Bachelor 3

UF : SPECIALITE DATA

---

1. CADRE DU PROJET

Ce projet permet l'évaluation des compétences acquises grâce aux modules de l'UF

« Spécialité Data».

Le projet consiste à réaliser une application data exploitant des analyses de données synthétisées sous forme de visualisations organisée pour raconter une histoire.

Vous devrez réaliser ce projet par équipe de deux.

Vous pouvez soumettre un projet personnel ou un projet à choisir parmi la liste proposée dans la section ***Descriptif du projet***

Dans tous les cas, le contenu et les fonctionnalités devront respecter des conditions décrites dans cette section. Les projets personnels devront être validés au préalable par le responsable de la formation.

Un **bonus** sera apporté aux projets qui proposeront des **méthodes d'analyses** ou de **visualisation** plus **poussées**, ainsi que pour les projets qui auront été **déployés en ligne**.

Vous devrez utiliser **obligatoirement du code en python** que vous pouvez compléter par d'autres langages si vous le souhaitez (par exemple : SQL, R, Javascript, ...)

2. OBJECTIFS DE FORMATION VISES

Vous 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

- 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

Interface graphique pour la création d'une application data :

Outils orientés visualisation de données :

Dash (de la suite logicielle Plotly), Panel (de la suite logicielleHoloviz), Bokeh

Outils orientés "*interface application data*" :

Streamlit, Anvil,

3. PREREQUIS

Afin de pouvoir réaliser le projet avec les compétences nécessaires, vous aurez besoin d'avoir suivi des modules d'enseignement concernant les thématiques suivantes :

- Gestion de projet avec Git

- Mathématiques pour la data science

- Python pour la data science

- Exploration et analyse de données

- Machine learning

4. 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.

5 MODALITES D'EVALUATION DU PROJET

Vous serez évalués par **une ou plusieurs notes** portant sur l'**ensemble des livrables** spécifiés à la section 4.

Vous serez également évalué par **une note de soutenance orale** de votre projet :
Vous présenterez votre application data au cours d'un créneau de 15 minutes suivies d'un temps de question et réponses d'une durée de 5 minutes.

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

Des points d'étapes intermédiaires auront également lieu au cours du déroulement du Projet.

6. DESCRIPTIF DU PROJET

Application data

Le projet consiste à réaliser une **application data** sur une thématique donnée que vous choisirez vous même (projet personnel) ou vous choisirez parmi les sujets proposés à la section 6.2

Votre note sera calculée en fonction des point de difficulté que vous pouvez récolter en remplissant les fonctionnalités obligatoires, ou celles en bonus, pour l'ensemble des livrables.

Fonctionnalités obligatoires

- Exploiter des données dont en décrivant comment vous les avez acquis (open data, scraping, APIs ...) **6 points de difficulté**

- Utiliser des méthodes d'analyses de données appropriées aux données recueillies **12 points de difficulté**

- Utiliser au moins un modèle de machine learning **8 points de difficulté**

- Organisez vos résultats sous forme de graphiques synthétiques pour raconter une histoire **10 points de difficulté**

- En utilisant des principes de data storytelling, présenter les analyses et conclusions que vous avez menées pour répondre au problème posé **7 points de difficulté**

- Déployer votre application localement et documenter son utilisation **4 points de difficulté**

- Proposer un dépôt Git structuré et documenté **6 points de difficulté**

Total des points de difficultés (sans bonus) : 41 points de difficulté

Bonus

Un bonus sera accordé aux projets qui seront innovants, proposeront des méthodes d'analyse ou de visualisation plus avancées, un data storytelling efficace.

Par exemple, les points suivants vous apporteront des bonus, mais les professeurs chargés de votre évaluation pourront vous accorder d'autres points de bonus.

- Proposer et se faire valider un sujet personnel **3 points de difficulté**

- Exploiter de manière pertinente des données cartographiques **6 points de difficulté**

- Exploiter de manière pertinente des données textuelles **6 points de difficulté**

- Réaliser des analyses poussées : clustering, régressions non linéaires, modèles prédictifs **8 points de difficulté**

- Ajouter une fonctionnalité à votre application permettant la mise à jour automatique des données et analyses présentées dans l'application **4 points de difficulté**

- Déployer votre application sur un serveur distant permettant d'accéder à votre application en ligne via une url **4 points de difficulté**

Total des points de difficultés (avec bonus) : 72 points de difficulté

Attribution de point de bonus sur votre note finale

En fonction des points de difficulté que vous aurez validés, vous obtiendrez des points bonus sur votre note finale :

Points de difficultés inférieurs à 30 : 0 points de bonus

Points de difficultés entre 35 et 50 : 1 points de bonus

Points de difficultés inférieurs entre 50 et 60 : 2 points de bonus

Points de difficultés supérieurs à 60 : 3 points de bonus

Document expliquant votre démarche

Vous devrez également livrer un document au format jupyter notebook (ou équivalent) expliquant de manière **synthétique** :

- la démarche que vous avez adoptée pour la réalisation du projet (acquisition des données, répartition des tâches, choix des méthodes d'analyses, choix des outils de développement, ...)

- les analyses préparatoires et exploratoires réalisées pendant le projet: vous pouvez y inclure les analyses que vous aurez choisi de ne pas conserver pour la présentation finale.

Dépot Git

Vous devrez également livrer un dépôt Git accessible en ligne contenant à la fois tout le code utilisé pour réaliser le projet ainsi que la documentation appropriée.
Vous inclurez dans votre dépôt une section spécifiant les étapes permettant à un utilisateur d'installer votre application en local sur sa machine (non nécessaire si vous avez déployé l'application sur un serveur distant)

6.2 Liste des projets

Projets personnel

Vous pouvez choisir une thématique de votre choix à traiter pour réaliser votre application data. Vous soumettrez au préalable la thématique au formateur afin que celui-ci puisse la valider.

Liste de projets

Si vous n'avez pas d'idée de projets sur lesquels vous pourriez trouvez suffisamment de données, vous pouvez choisir parmi les problématiques suivantes :

Attention, les sujets proposés sont volontairement très généraux, il vous appartiendra de trouver un axe précis d'analyse et d'histoire à raconter. Essayez d'être originaux !

Projet 1

Raconter une histoire concernant l'évolution de la pandémie covid-19 dans le monde

Projet 2

Construisez un résumé des conclusions de ces 5 dernières années concernant le changement climatique ainsi que des projections de scenarii possibles

Projet 3

Dressez un bilan des principales migrations dans le monde durant ces 5 dernières années

7. 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