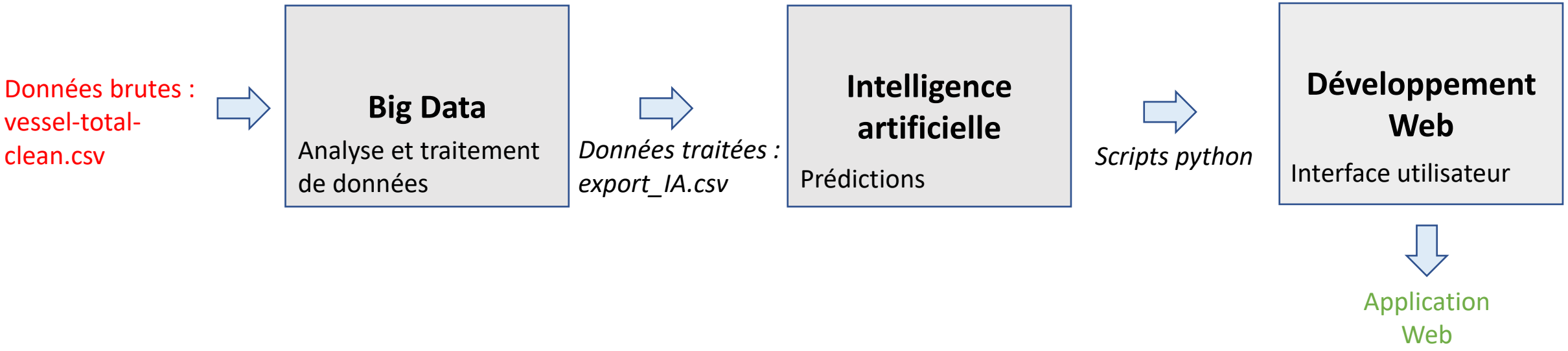
ISEN ALL IS DIGITAL! OUEST | yncréa

Projet Année 3
Big Data/IA/Web

Partie Big Data

Contexte du projet

Déroulement du projet



Sujet

Objectif

Concevoir et développer une application d'étude des données AIS

Approfondir les compétences acquises dans les modules *Big Data, Intelligence Artificielle, Développement Web et Base de Données* à travers une application complète de traitements et de visualisation de données concernant les données provenant du système d'identification automatique (AIS en anglais) des bateaux dans le golfe du Mexique.

Objectifs de la partie Big Data :

- Extraction des données : à partir d'un fichier, d'un site web...
- Visualisation d'un grand volume de données
- Nettoyage des données : suppression des données incomplètes, suppression des données erronées
- Application de modèles statistiques pour l'analyse des données recueillies : régression linéaire et/ou régression linéaire multiple et/ou régression logistique, corrélation entre les caractéristiques

Cahier des charges

Cahier des charges

5 parties principales sont attendues :

1. Exploration des données
2. Visualisation des données sur des graphiques
3. Visualisation des données sur une carte
4. Etude des corrélations
5. Prédiction de la variable « VesselType»

Fonctionnalité 1

Description et exploration des données:

- Description du jeu de données
- Statistiques descriptives univariées
- Nettoyage des données
  - Valeurs manquantes, valeurs aberrantes
  - Doublons

Fonctionnalité 2

Visualisation des données sur des graphiques :

- Créer des représentations graphiques
  - Exemple: répartition des bateaux suivant leur type

- Créer des histogrammes
  - Exemple: Différentes catégories de bateau, ports les plus utilisés.

Exporter et sauvegarder vos figures en png !

Fonctionnalité 3

Visualisation des données sur une carte :

- Construire des trajectoires de bateaux (grâce au latitude et longitude)
  - Exemple : Afficher toutes les trajectoires, la trajectoire d'un bateau par son nom.
  - En déduire les routes principales.

Fonctionnalité 4

Etude des corrélations entre variables :

- Quels sont les liens entre les variables ?
  - Exemple : si on veut estimer la variable type de bateau, quelles sont les variables importantes dans son estimation?
- Conduire des analyses bivariées
- Etude des relations entre variables qualitatives
  - Faire des tableaux croisés et des tests d'indépendance du chi2 sur les tableaux entre les différentes variables
  - Représenter graphiquement ces tableaux (mosaicplot) et les analyser

Fonctionnalité 5

Prédiction / Régression :

- Régression logistique :

➢ Prédiction de la variable VesselType en fonction des variables pertinentes.

➢ Autre exemple : sélection de quelques bateaux, calculer leur vitesse et mesurer quantitativement l'erreur commise par votre méthode.

Fonctionnalité

Export pour l’IA

- Exporter le fichier nettoyé en format csv.

Cahier des charges

Technologies à utiliser

Déroulement

Organisation

Travail en trinôme :

- Attention à bien se répartir le travail en prévoyant les tâches de chacun avec un **diagramme de Gantt**

Ressources externes :

- Tous les documents sont autorisés
- Attention à utiliser avec une grande précaution tout document extérieur : site de vulgarisation, forum, code d'autrui

Documentation du projet :

- Au fur et à mesure
- Standardisée
- Livraison de code ou de documents :
  - Ne pas attendre la dernière minute pour poster un livrable
  - Préparer des livrables intermédiaires (surtout pour les sources)
  - Sauvegarder régulièrement vos données

Livrables et évaluations

Livrable

Format de l'archive :

Archive *ZIP*, *TGZ*, *7ZIP*, pas de *RAR :* `projetbigdata_groupeX.zip` (remplacer X par votre numéro de trinôme)

Le rendu final doit contenir :

- Un rapport (10 pages) qui décrit vos résultats (représentations graphiques, cartes, …)
- L'intégralité de vos codes sources commenté avec vos ressources **ainsi que les données de l'export à l'IA** (scripts *R*)
- Votre diagramme de Gantt en *PDF*

Remarques :

- Malus possible sur l'un des membres du groupe si l'investissement est jugé trop faible
- Possibilité d'être interrogé durant le projet de façon individuelle
- Plagiat sévèrement sanctionné pour TOUS les membres du/des groupe(s)

Attention

Les livrables seront à poster sur l'intranet. Tout retard sera sanctionné (l'heure du réseau faisant foi).
Les fichiers au mauvais format ou avec un mauvais nommage seront pénalisés.

Évaluation

Présentation orale :

- Soutenance de 10 minutes (strict) + 5 minutes de questions
- Présentation en trinôme (pensez à vous répartir la parole)
- Présentez l'essentiel de votre projet

Rapport :

- Description de vos résultats dont les représentations graphiques

Evaluations des compétences :

- Des questions seront posées tout au long du projet pour vérifier les acquis

Code :

- Rendu de l'intégralité de vos codes sources avec les ressources associées

Barème indicatif : Soutenance 40% – Compétences 40% – Évaluation du code/Rapport 20%

ISEN
ALL IS DIGITAL!

yncréa

MERCI

Des questions ?