ISEN ALL IS DIGITAL! OUEST | yncréa

Projet Année 3
Big Data/IA/Web

Big Data

Benoit Lardeux

Bilel Benziane

Contexte du projet

Sujet

Objectif

Concevoir et développer une application d'étude des accidents de la route

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 accidents corporels de la circulation routière en France.

Objectifs de la partie Big Data :

- Nettoyer et analyser un jeu de données
- Comprendre les interactions entre les variables
- Construire les visualisations appropriées
- Concevoir des modèles explicatifs des données

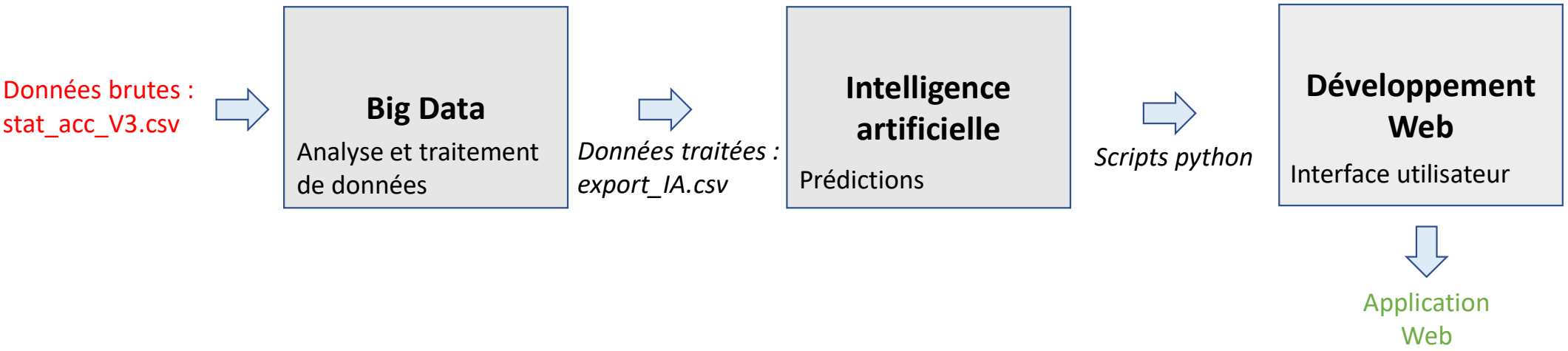
BL0

Diapositive 3

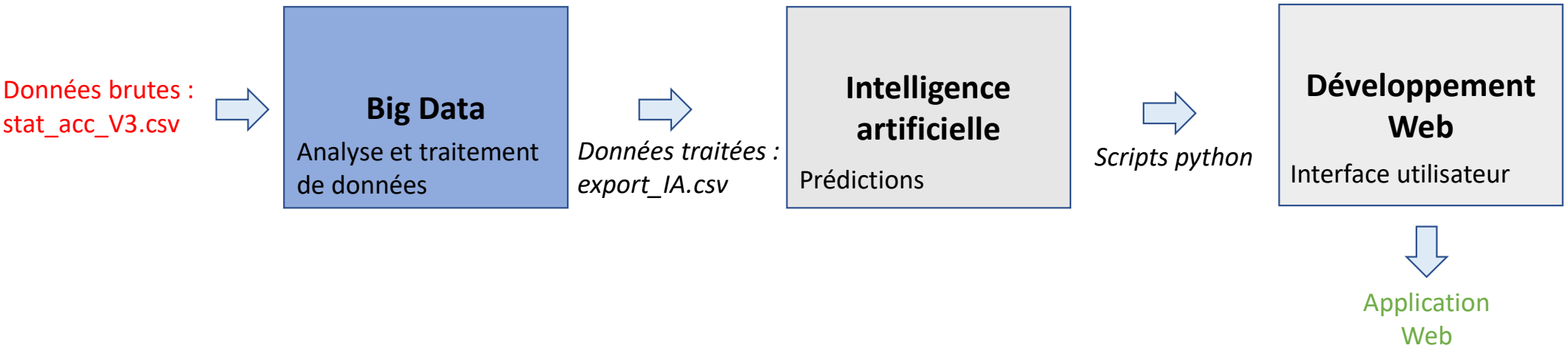
BL0 A revoir

Benoit LARDEUX; 2023-06-09T12:58:37.274

Déroulement du projet



Déroulement du projet



Descriptifs des données

- Données disponibles sur data.gouv
  - Lien vers données brutes et descriptif
- Une partie des pré-traitements déjà réalisés
- Un fichier csv avec les données est disponible sur l'ENT

Descriptifs des données

- Quelles sont les informations présentes ?
  - 4 types d'informations globales :
    - usagers,
    - véhicules,
    - lieux,
    - Caractéristiques

- Exemple : sur un accident corporel de la circulation routière, on a les données suivantes : type véhicule, gravité de l'accident, lieu de l'accident (latitude, longitude), heure, luminosité, type de voie, le port de la ceinture de sécurité, etc…

Calendrier prévisionnel

Déroulement

Lundi	Mardi	Mercredi	Jeudi	Vendredi
• Présentation du projet • Suivi du MOOC sur Git • Organisation du travail	• Visualisations des données	• Analyse des données	• Analyse de données • Préparation de la soutenance et de l'export pour l'IA	• Dépôt des livrables sur l'ENT • Soutenances
• Préparation des données	• Visualisations des données	• Analyse des données		• Soutenances

Organisation

- Travail par groupe de 3 (+ 2 groupes de 4)
  - Il est attendu que chaque étudiant connaisse l'ensemble du projet (GIT)
  - Nécessité de se répartir le travail convenablement

- Ressources externes:
  - Tous les documents sont autorisés
  - Mais les documents extérieurs doivent être utilisés avec une grande précaution!

Cahier des charges

MOOC Git et gestion de projet

- Répartition des taches dans le groupe, diagramme de Gantt

- Faire une description des données, des différentes variables, etc…

Préparation des données

- Quels traitements appliquer lorsqu'il manque des informations ou que ces informations ne sont pas exploitables ?
  - Présenter quelques exemples et expliquer les traitements effectués

- Recoder des variables multimodales en chiffres (0,1,2,3, etc..) : catégories des véhicules, niveau de gravité de l'accident

- Mettre les variables numériques sous format numériques, date sous format date, etc…

Préparation des données (2)

- Construire des séries chronologiques sur l'évolution du nombre d'accidents par mois et semaines sur l'ensemble de la période
  - A quel niveau d'agrégation (mois ou semaine) les données collectées permettraient-elles de faire une prévision de bonne qualité avec une régression linéaire ?

- Construire un jeu de données avec le nombre d'accidents selon la gravité pour 100.000 habitants par région (qui servirait à l'Analyse en Composantes Principales (ACP) discutée dans la section 3 de l'analyse de données)

Visualisations des données

- Créer des représentations graphiques pour:
  - Nombre d'accidents en fonction des conditions atmosphériques
  - Nombre d'accidents en fonction de la description de la surface
  - Nombre d'accidents selon la gravité
  - Nombre d'accidents par tranches d'heure
  - Nombre d'accidents par ville

- Créer des histogrammes :
  - Quantité d'accidents en fonction des tranches d'âges
  - Moyenne mensuelle des accidents

Visualisations des données (2)

- Proposer une représentation sous formes de carte de la quantité d'accidents enregistrés par région puis par départements
- Même chose avec les taux d'accidents graves
- Exporter et sauvegarder vos figures en png

Analyse des données

- Etude des relations entre variables qualitatives
  - Faire des tableaux croisées et des tests d'indépendance du chi2 sur les tableaux entre les différentes variables
  - Représenter graphiquement ces tableaux (mosaicplot) et les analyser

- Calculer les régressions linéaires sur l'évolution du nombre d'accidents par mois, puis par semaine.
  - Comparer les résultats obtenus par les deux régressions mentionnées ci-dessus
    - o Analyser les performances de la régression (proportion de la variabilité due aux résidus et aux variables explicatives)
    - o Analyser des erreurs types associés aux estimateurs
    - o Calculer les intervalles de confiance à 95% pour ces estimateurs
    - o Calculer les R2 et R2 ajusté pour les deux modèles. Qu'en déduire ?
    - o Quelle est la qualité des prédictions basées sur ces modèles de régression (à la semaine et au mois)?

Analyse des données (2)

- Bonus: ACP-AFC-ACM sur le jeu de données avec le nombre d'accidents selon la gravité pour 100.000 habitants par région
  - Eboulis des valeurs propres pour déterminer le nombre de variables principales à utiliser
  - A quelle part d'inertie, ça correspond
  - Contributions des variables (gravité) et des individus (régions) sur les axes principaux
  - Représentation des variables sur le cercle des corrélations
  - Interprétation des axes dans le plan principal

Export pour l’IA

- Exporter le fichier nettoyé en format csv

Evaluations

Modalités de l'évaluation

1- Soutenance de 15 mns, plus 5 mn de questions (30%)

2- Rapport de 10/15 pages (30%)

3- Dépôt des programmes sur l'ENT (30%)

4- Export des data pour l'IA (10%)

Livrable

- Le rendu final doit contenir
  - La présentation au format pdf
  - Le rapport
  - L'intégralité des codes sources commentés
  - Fichiers d'input des programmes

- Livrable à déposer sur l'ENT. Les retards seront pénalisés
  - Format de l'archive: *ZIP*, *TGZ* ou *7ZIP*, pas de *RAR*: **`projetbd_groupeX.zip`**

- Remarques
  - Malus possible pour 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)

ISEN
ALL IS DIGITAL!

yncréa

MERCI

Des questions ?