2020 - 2021

**Rapport projet annuel**

**Nom - Prenom** : BENDAVID Natane, WADE Cheikh Abdourahm

**Promotion** : 4 IABD 2

**Matière** : Projet Annuel

**Intervenant**: WAJNBERG Charles David

Table des matières

[Introduction 3](#_Toc77081524)

[Contexte du projet 3](#_Toc77081525)

[Projetsimilaire 3](#_Toc77081526)

[Objectifs 4](#_Toc77081527)

[Description fonctionnelle 5](#_Toc77081528)

[Pipeline de déploiement des données et du modèle de ML 5](#_Toc77081529)

[Pipeline de déploiement pour la génération de la data visualisation 9](#_Toc77081530)

[Pipeline pour l’api de prédiction immobilière 9](#_Toc77081531)

[Aperçu des contenus 9](#_Toc77081532)

[Architecture technique 10](#_Toc77081533)

[Données mises en œuvre 10](#_Toc77081534)

[Algorithmes mis en œuvre 10](#_Toc77081535)

[Gestion de projet 10](#_Toc77081536)

[Résultats obtenus 10](#_Toc77081537)

[Adresse des dépôts de code (github, etc.) 10](#_Toc77081538)

# Introduction

Dans ce rapport de projet annuel, nous allons mettre en évidence toute l’élaboration et le déroulement de la réalisation de ce projet. Nous expliquerons la partie technique et fonctionnelle mise en place pour ce projet. Nous mettrons en évidence les choix réaliser pour aboutir à notre résultat final

## Contexte du projet

Nous commençons par un rappel de l’objectif du projet Annuel. Ce projet a pour but de mettre en application les savoirs issus des cours suivies à l’ESGI.

Le but du projet annuel est de réaliser un projet qui respecte les différents critères :

* L’utilisation d’une plateforme et d’une architecture dans le Cloud
* L’automatisation de tous les traitements réalisés
* Le déploiement de notre application
* La réalisation de l’objectif via le savoir issus des cours de l’ESGI
* La mise en place d’une partie Intelligence artificielle

Le sujet du projet que nous avions décidé de travaille concerne l’immobilier.

Lors de nos premières réunions en groupe, nous avons évoqué nos différents sujets qui pourrait nous intéresser. C’est alors que le sujet de l’immobilier nous intéresser car dans notre vie privée nous étions nous deux dans une période de recherche d’appartement. Le sujet de l’immobilier à été tous de suite une idée validé par l’ensemble de l’équipe.

## Projetsimilaire

Une fois le sujet validé par l’ensemble de l’équipe, nous avons décider de s’informer sur le monde de l’immobilier ainsi que les sites existants sur le marché.

Le projet de référence que nous avions vu et utilisé durant nos recherches d’appartement dans le cadre de notre vie privée est le site internet [meilleurs agent](https://www.meilleursagents.com/prix-immobilier/paris-75000/rue-du-faubourg-saint-antoine-2672/242-b/).

Une image contenant texte

Description générée automatiquement

Sur ce site web nous pouvons renseigner une adresse postale du bien à estimer. Le site meilleurs agents nous renvoie alors une estimation moyenne du prix au mettre carré de l’adresse donnée, ainsi qu’une fourchette de prix.

Nous nous sommes alors donnée objectif de réalisé une application du même type.

## Objectifs

L’objectif de notre projet est de réaliser une Api et une application avec de l’intelligence artificielle capable d’estimer un bien immobilier pour une adresse et surface donnée. Le but est que notre application et Api nous donne une estimation la plus précise du bien donné.

Comme dit au-dessus, ce projet a deux objectifs :

Le premier est de créer une Api permettant de renvoyer plusieurs informations sur un bien immobilier (prix estimer, prix du voisinage et distance entre les biens estimé). Cette api permettra au développeur voulant intégrer sur le site internet ou application immobilière un estimateur d’un bien immobilier.

Le second est la création d’une application web permettant aux utilisateurs lambda, agent immobilier ou même particulier ou professionnels voulant estimer, vendre sont bien immobilier et connaitre une estimation du prix de son bien

Comme décrit un peu plus haut notre objectif pour ce projet est de réalise cet estimateur de bien immobilier pour :

* des particuliers ou professionnels voulant estimer leurs biens immobiliers.
* des particuliers ou professionnels voulant vendre leurs biens immobiliers
* des particuliers ou professionnels voulant acheter un bien immobilier
* des particuliers ou professionnels voulant investir dans un bien immobilier
* des développeurs voulant mettre en place un système d’estimation de bien immobilier sur leurs sites web

# Description fonctionnelle

Pour la partie fonctionnelle de ce projet d’estimateur immobilier nous avons 3 grandes parties fonctionnelles :

1. Une première partie récupération, nettoyage, traitement et préparation des données avec la création de modelé de machine Learning de régression et de clustering ainsi que le stockage de l’ensemble dans le cloud. Tous cette partie est automatisé
2. Une seconde partie qui permet de générer toute la partie data visualisation de l’ensemble des données mise à jour. Ce traitement est aussi automatisé.
3. Une dernière partie qui est l’API qui permet de prédire le prix d’un bien immobilier par rapport à son adresse postale et ça surface.

Dans chaque partie nous détaillerons précisément toute les démarche réaliser pour aboutir à chaque fonctionnement et nous schématiserons le résultat.

## Pipeline de déploiement des données et du modèle de ML

Dans cette pipeline automatiser via une tache CRON, nous avons réalisé différentes étapes.

La première partie de cette pipeline est la récupération des données. Pour établir notre modèle de Machine learning plus tard nous avons besoin d’une grande quantité de donnée pour avoir un modèle performant.

Nous avons alors utilisé les données proposées par Etalab sur data-gouv.fr : **Demandes de valeurs foncières géolocalisées** (<https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres-geolocalisees/>).Sur ce dataset, **Etalab** nous fournie un historique de données sur les 5 dernières années.

Une image contenant texte

Description générée automatiquement

Comme sur l’image si dessus, nous avons à disposition un file système avec un dossier par années (sur les 5 dernières années courante) qui contient un fichier compressé csv « **full.csv.gz »** qui comprend toutes les données pour chaque département. Pour notre pipeline, nous téléchargeons les 5 fichiers csv des 5 dernières années courantes. Nous décompressons toutes les données et l’utilisons dans un seul et même dataset.

Une fois les données mise à disposition, nous téléchargeons un second dataset de data-gouv qui comprend le numéro de région correspondant au numéro de département (<https://www.data.gouv.fr/fr/datasets/regions-departements-villes-et-villages-de-france-et-doutre-mer/>). Ce dataset est stocké puis récupère depuis le bucket S3 créer pour le projet immobilier. Une fois le dataset charger, on le merge avec le dataset immobilier par rapport au code départements données pour obtenir le code régions correspondant

Une fois toutes les données correctement chargées nous nettoyons les données mise a disposition pour ne pas avoir de mauvaises données, des données incomplètes ou des données qui pourrait détériorer la création du modèle ML. Sur les 40 variables du dataset nous gardons uniquement celle-ci :

*‘id\_mutation', 'date\_mutation', 'nature\_mutation', 'valeur\_fonciere', 'code\_postal', 'code\_commune', 'code\_region', 'code\_departement', 'code\_type\_local', 'type\_local', 'surface\_reelle\_bati', 'nombre\_pieces\_principales', 'latitude', 'longitude'*

Maintenons-nous filtrons les données :

* Les données correspondantes à une vente ou vente en l’état futur d’achèvement
* Les valeurs foncières non vide et différentes de 0
* Les surface du biens non vide et différent de 0
* La longitude et la latitude non vide
* Un nombre de lots égales a 0
* Un code départements non vide

Une fois les données nettoyer et filtrer, nous procédons a quelque traitement sur certaine variables. Nous créons aussi la variable *prix\_au\_mettre\_carre* qui nous servira plus tard dans notre modèle. Pour terminer sur le traitement des données, nous regroupons toutes nos données des différentes années dans un même dataframe pour l’utiliser ensuite dans nos modèles.

Une fois toute la partie traitement terminer, nous commençons la partie ML. Mais avant cela, pour reproduire l’estimation d’un agent immobilier nous avons besoin d’une autre donnée qui n’est pas présente dans le dataset. Pour estimer correctement un bien, nous ne pouvons pas uniquement nous fixer sur le prix au mètre carré d’un bien, mais nous avons besoin du prix au mètre carré des autres biens aux l’alentours du bien estimé. Pour cela nous allons alors récupère les 10 biens les plus proches et récupérer la moyenne du prix au mettre carré de l’ensemble de ces biens. Pour déterminer cette variable et récupère les bien les plus proche, nous devons déterminer la distance entre le bien estimé et les autre bien de la région et récupère les 10 biens dont la distance est la plus petite. C’est à partir des données latitute/longitude que nous pouvons calculer cette distance. Nous mettons alors en place un modèle de partitionnement spatial appelé le model *BallTree* ou modèle de *voisinage* disponible dans la librairie sklearn.

Malheureusement, il est impossible d’utilise ce modèle directement sur l’ensemble des données car le traitement serrait trop long. C’est pour cela que nous découpons notre dataset par régions et créons alors un modèle de partitionnement par régions.

Ce modèle nous permettra ensuite d’estimer pour chaque bien le prix au mètre carré du voisinage.

Pour réaliser ce modèle de partitionnement :

* On sépare les données par numéros de région
* On crée ensuite le modèle avec les données latitude/longitude
* On enregistre et charges les modèles de chaque région dans le bucket S3
* On utilise ensuite le modèle pour récupérer la distance moyenne entre les 10 bien voisin et les indices des 10 biens voisins pour chaque bien. La distance obtenue est en haversine. Pour la convertir en Km on doit multiplier par la surface de la terre soir 6371 km)
* Avec ces données on récupère pour chaque bien c’est 10 voisins les plus proche et on calcule la moyenne du prix au mettre carré de ces 10 biens. On appelle cette variable *prix moyen du cartier*
* On sauvegarde chaque dataset par régions et les charges sur le bucket S3

Une fois toutes ces étapes réalisées, on obtient bien alors la variable qui nous intéresser qui est le prix moyen du cartier que nous utiliseront comme variable pour notre modèle de ML

Pour terminer cette pipeline, nous devons maintenant entrainer notre modèle de Machine Learning. Pour cela nous avons réalisé différents tests avec différents modèles :

* Un MLP avec la lib Keras de Tenserflow
* Un GridSearch avec la lib XGboost
* Un RandomForestRegressor avec la lib scikit-learn

Après ces différents tests réalisés avec ces modèles, nous avons obtenues de meilleur résultat avec un RandomForestregressor.

Nous avons obtenu avec ce modèle 7% d’erreur. Les inputs (variables) que nous utilisons pour entrainer se modèles sont :

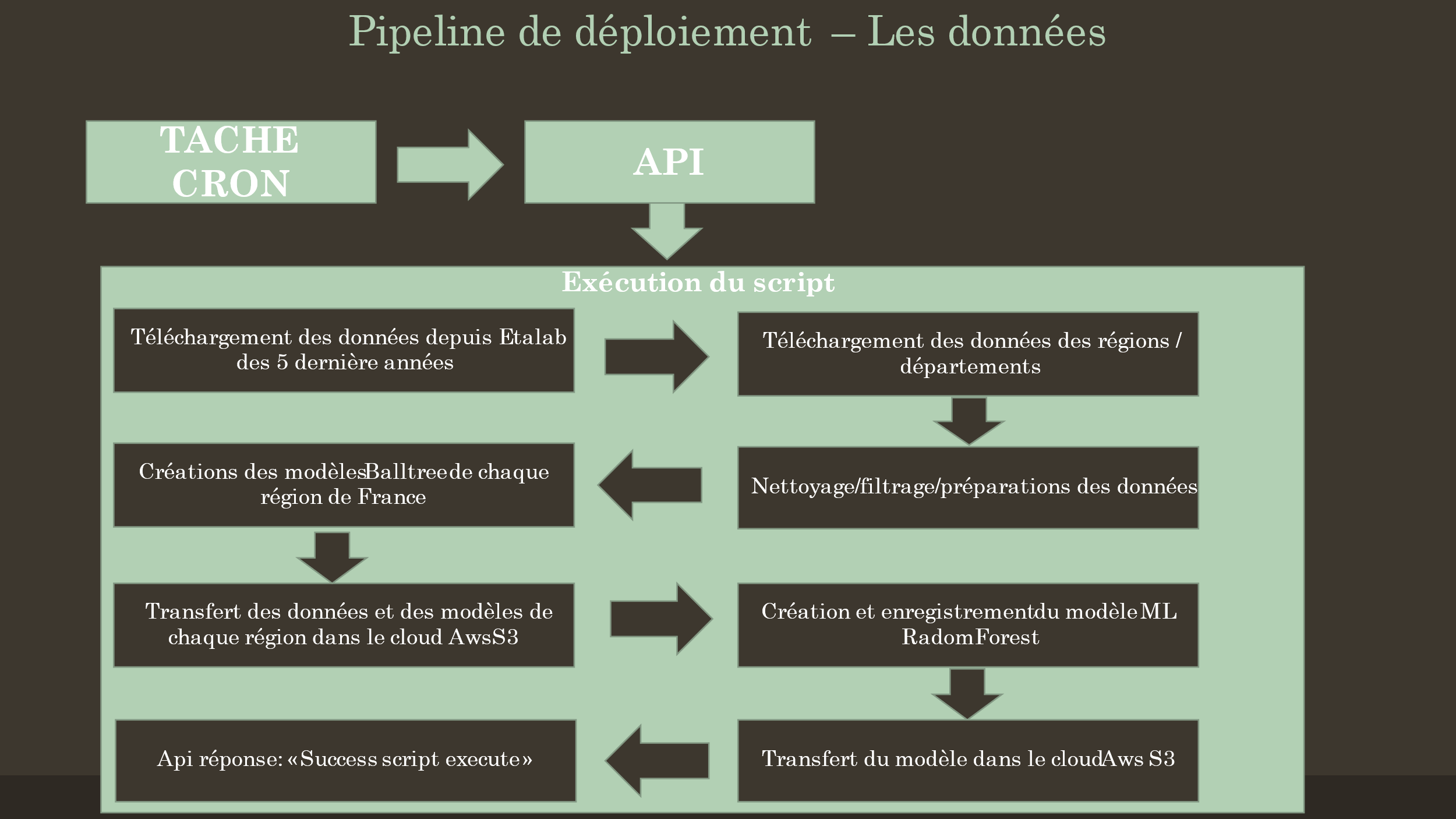
*'latitude', 'longitude', 'surface\_reelle\_bati', 'prix\_metre\_carre', 'code\_region', 'prix\_moyen\_cartier'*

Une fois le modèle entrainé nous le sauvegardons via la lib pickel (disponible directement sur sklearn) et le chargeons sur le bucket S3

Tous ces traitements sont exécutés via l’appelle d’une API. Pour cela nous utilisons la librairie python Flask pour mettre en place cette API.

Pour automatiser alors tous les traitements, nous lançons une tache CRON tous les 6 mois pour récupère les nouvelles données mise à jour par Etalab, effectuer les traitements sur les données, les mette à jour sur le bucket S3 et réentraîné le modèle de ML.

Pour schématiser tous les traitements réaliser et expliquer si dessus, voici un schéma qui résume les grandes étapes de réalisation de ce workflow :



## Pipeline de déploiement pour la génération de la data visualisation

## Pipeline pour l’api de prédiction immobilière

## Application web pour estimer un bien immobilier

# Architecture technique

# Données mises en œuvre

* + Formats
  + Flux
  + Scrapping
  + Saisie et feedback utilisateur
  + etc.

# Algorithmes et outils utilisés

* + Librairies existantes
  + Conception
  + Paramétrage
  + Performances
  + etc.

# Gestion de projet

* + Planning de réalisation
  + Répartition du travail
  + Problèmes rencontrés

# Résultats obtenus

* + Bilan fonctionnel
  + Améliorations possibles

# Adresse des dépôts de code (github, etc.)