Projet 3 :
W-LAB-Analytics

WILD CODE SCHOOL

Février 2024

L'ÉQUIPE

Fanny GRANCHER

Marie MAMDY

Nassim IRID

Thibault QUAGHEBEUR

01

Le Projet

Contexte

Les maladies chroniques, qu'est-ce que c'est ?

- **de longue durée** dont les **facteurs de risque** sont :
  - **L'obésité,**
  - **le tabagisme,**
  - **la consommation d'alcool,**
  - **la mauvaise alimentation,**
  - **l'inactivité physique.**

En France, **15 millions**
de personnes sont atteintes de maladies chroniques

80 % **des décès prématurés**
dus aux maladies non transmissibles

Objectifs du projet

- **Acquérir des connaissances** de base en recherche scientifique
- **Développer un modèle** prédictif
- **Concevoir une application** permettant de prédire le risque de développer l'une des maladies suivantes :
  - Diabète
  - Cancer du sein
  - Maladie rénale chronique
  - Maladie chronique cardiaque
  - Maladie du foie

Réflexion sur l'Éthique et la Confidentialité

Les **données de santé** sont considérées comme **sensibles** en vertu du RGPD.
Le règlement européen vise à renforcer la protection des données de santé à caractère personnel.
Le **RGPD** interdit en principe leur traitement et impose des **règles strictes** concernant leur transfert en dehors de l'Union Européenne, il prévoit des **formalités spécifiques** pour le transfert **en dehors de l'UE**.

Les principes éthiques dans le traitement des données, en particulier l'**éthique clinique** est ancrée sur **4 principes** :

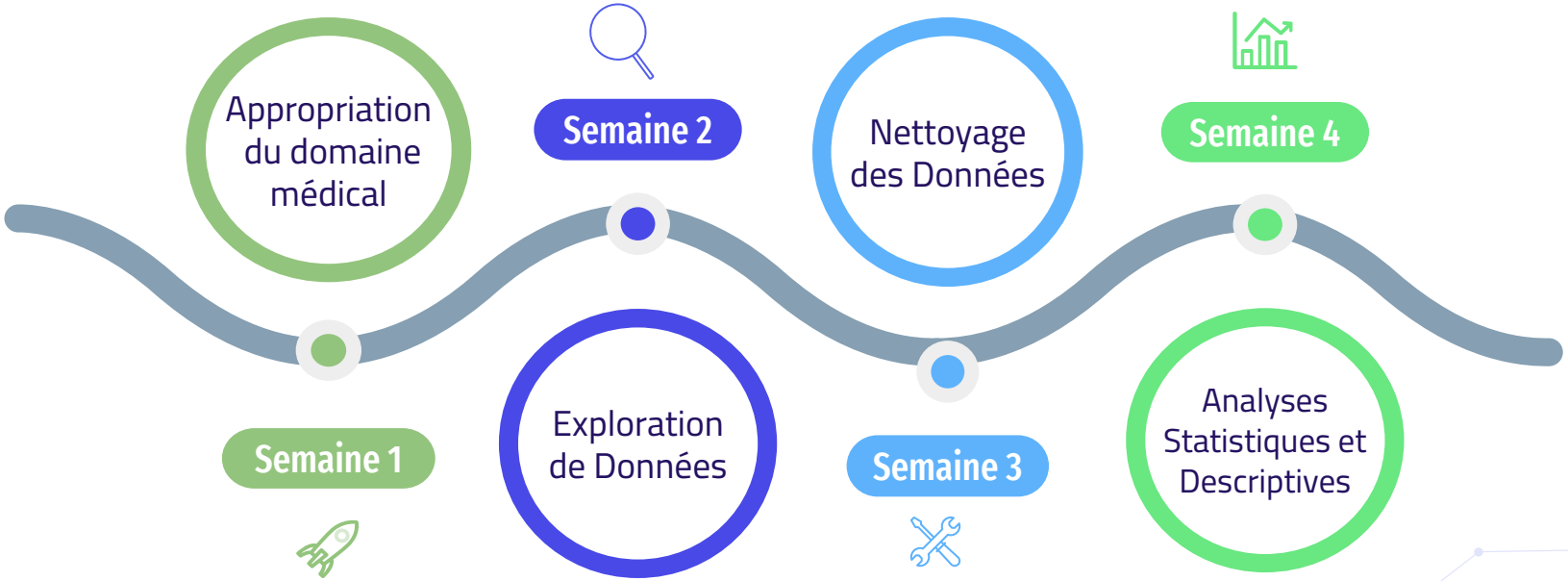
- La bienfaisance
- La non-malfaisance
- Le respect de l'autonomie du patient
- La justice

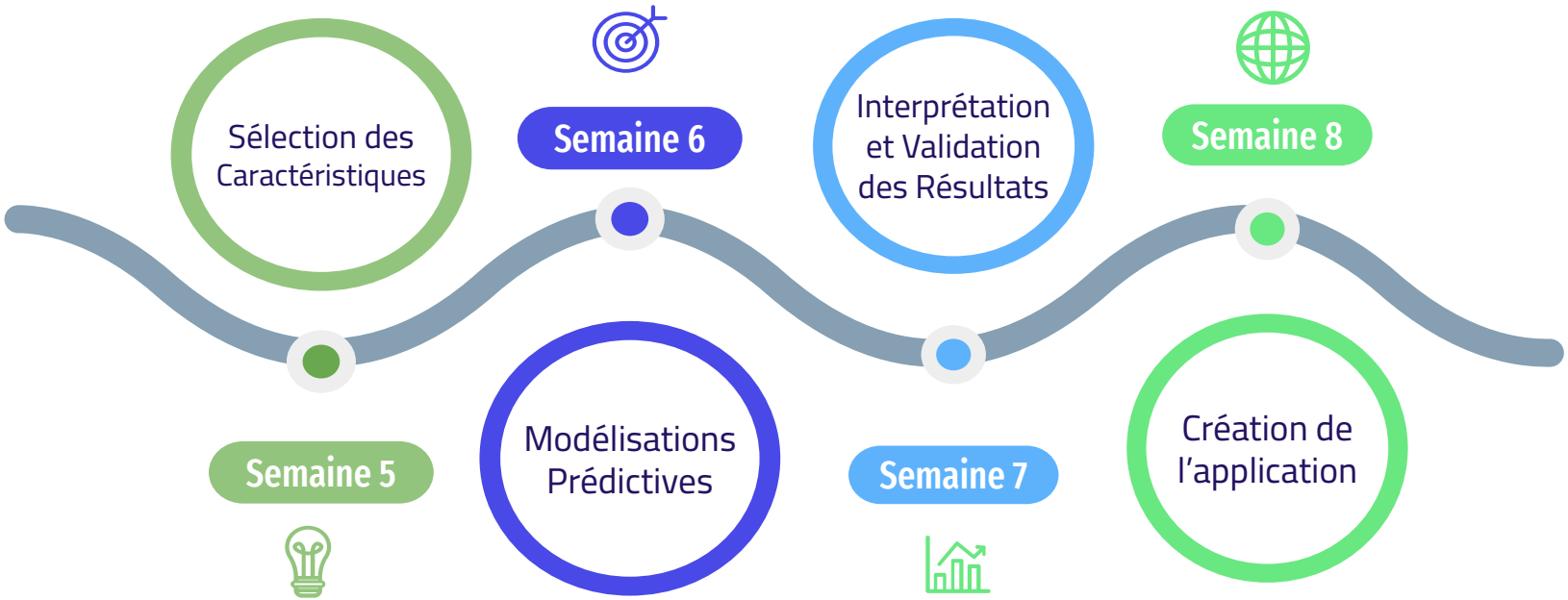
Visent à assurer le **respect des patients** et la **protection** de leurs données, tout en **garantissant des soins** de qualité et équitables.

L'utilisation des **données massive** en santé, par exemple, peut soulever des **enjeux éthiques** liés au **secret médical**, à la **responsabilité de la décision médicale** et au **respect de l'autonomie** des patients.

Essentiel : Veiller à ce que le traitement des données de Santé soit conforme à ces principes éthiques fondamentaux

Méthodologie




Sélection des Caractéristiques
Semaine 5
Semaine 6
Modélisations Prédictives
Interprétation et Validation des Résultats
Semaine 7
Semaine 8
Création de l'application

Outils et langages utilisés

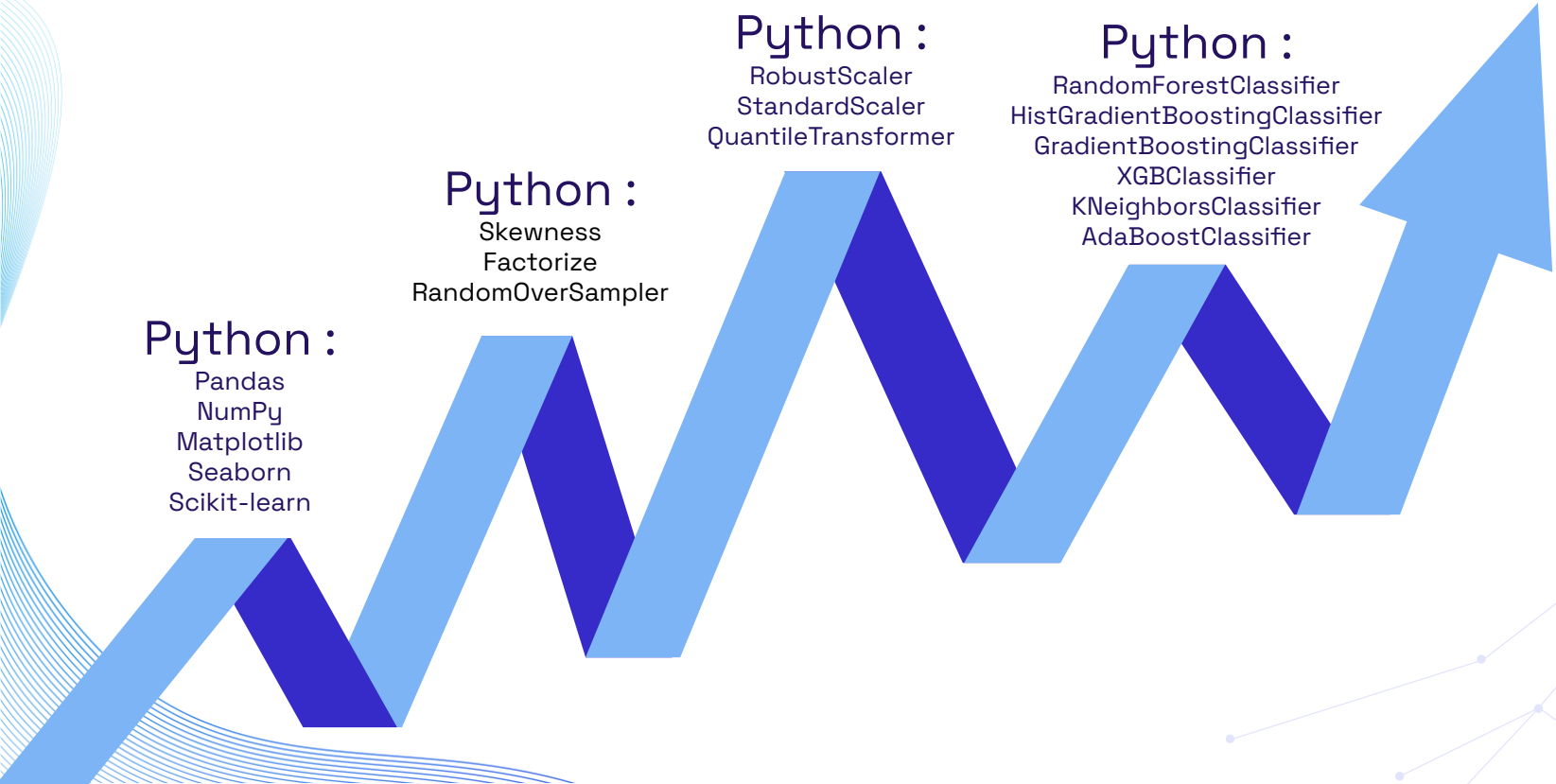
Streamlit

Python :
Pandas
NumPy
Matplotlib
Seaborn
Scikit-learn

Python :
Skewness
Factorize
RandomOverSampler

Python :
RobustScaler
StandardScaler
QuantileTransformer

Python :
RandomForestClassifier
HistGradientBoostingClassifier
GradientBoostingClassifier
XGBClassifier
KNeighborsClassifier
AdaBoostClassifier



02

Exploration et Nettoyage des Données

Import, exploration, traitement

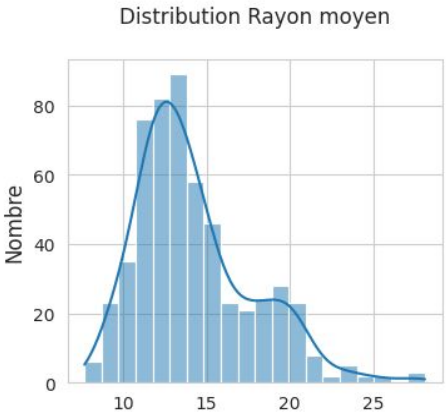
Chaque Dataset a ses spécificités, cependant, certains traitement sont communs :

- Import, info(), describe()
- Création de dictionnaires (variables numériques / catégorielles
- Recherche de valeurs manquantes, de doublons, de zéros
- Recherche d'irrégularités (virgule, caractères spéciaux, espaces)
- Modifications de certaines valeurs avec l'accord du client

Analyse Statistique et Descriptive

Visualisations - Analyses univariées :

- Distribution
- Commentaires / Constats
- Corrélations



Médiane: 13.455
Moyenne: 14.24
Max: 28.11
Min: 7.691
Nombre de lignes avec 0 : 0

Observation et calcul de l'asymétrie des distributions (skewness)

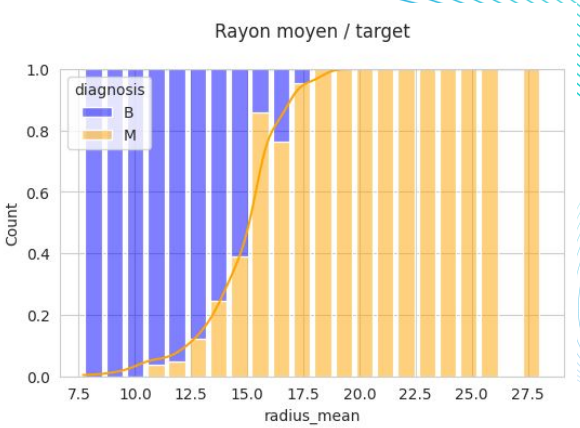
- Déterminer par quoi remplacer les valeurs manquantes

Calcul de la moyenne et de la médiane par classe (malade ou non)

- Remplacement des valeurs manquantes

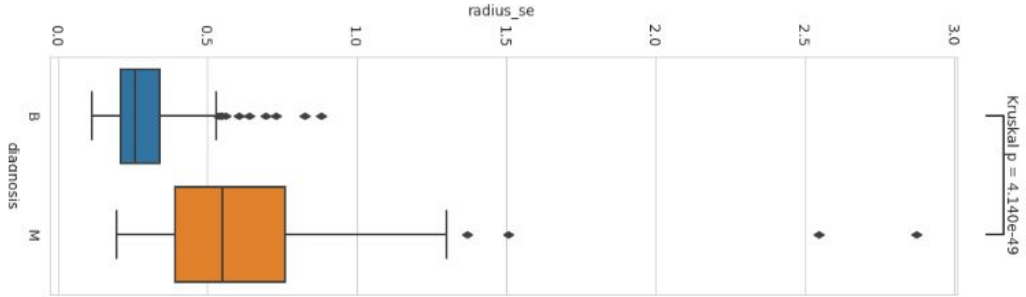
Visualisations - Analyse bivariée :

- Répartition des données par variable
  - Impact de la classe “malade” sur les variables
- Commentaires / Constats



Visualisations - Indication de la significativité des variables selon la classe (calcul de la p-value) :

- Première observations des variables significatives pour le modèle de prédiction



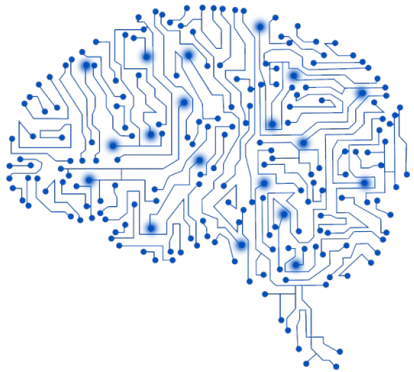
03

Machine Learning

Pré-traitement

- **Rééquilibrage des classes (malade/non-malade)** dans les datasets du diabète et de la maladie du foie (Augmente artificiellement la Classe par duplication)

- **Transformation, Structuration** des données en un format numérique uniforme (même échelle) pour être traitées par un algorithme de façon optimale

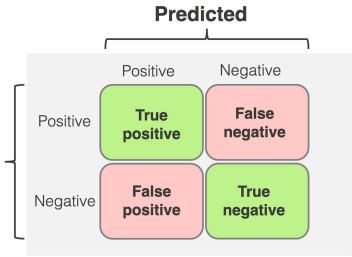


Identification des modèles

- Utilisation d'une fonction pour **croiser plusieurs scaler (8) avec plusieurs algorithmes (9)** :
  - Renvoi de toutes les combinaisons avec le meilleur score et hyperparamètres
  - Choix du meilleur Scaler et des 4 ou 5 meilleurs modèles (donnant le moins de faux négatifs)

- Utilisation de la **stratégie de Vote** (voting) avec le scaler et les modèles choisis :
  - Ici, la classe qui reçoit le plus de vote est choisie comme prédiction finale
  - Choix de cette solution pour avoir des résultats plus robustes et stables

- **Calcul des scores** de la stratégie de vote :
  - Précision des prédictions sur l'ensemble des données (Accuracy)
  - Vrais positifs / Ensemble des positifs (Recall)
  - Matrice de Confusion



L'objectif :
Réduire le nombre de faux négatifs

Interprétations

Vérification de la significativité des variables :

Lasso : méthode statistique qui identifie et conserve les variables les plus significatives pour le modèle, alors que les variables non significatives ou redondantes sont écartées

A permis d'**éliminer les variables redondantes** qui pourraient :

- **biaiser les résultats** des modèles,
- éviter le **surapprentissage**.

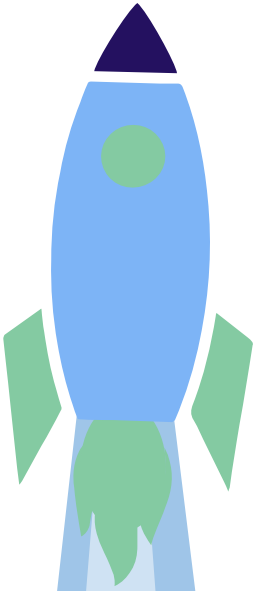
Précision des Prédictions entre 0.83 et 0.98

04 Application

L’application doit inclure :

- Le **chargement des modèles** pré-entraînés pour chaque maladie
- Une interface pour la **saisie** des données utilisateurs selon les caractéristiques requises par chaque modèle.
- Les résultats de prédiction aux utilisateurs avec un **avertissement clair** que ces prédictions sont informatives et ne remplacent pas un diagnostic médical professionnel.

Lien vers l'application Streamlit



Limites et Axes d’amélioration

01 Approfondir les connaissances médicales pour améliorer nos analyses (éviter les biais, redondances des variables)

02 Améliorer des modèles par la collaboration avec des Data Scientists

03 Automatiser de la méthodologie (Classes, Fonctions)

04 Axer l’application sur la prévention des maladies

05 Améliorer la carte des spécialistes

Merci pour votre attention !

Avez-vous des questions ?

Annexes

Liens vers les notebooks

- Diabète
- Cancer du Sein
- Maladie Rénale Chronique
- Maladie Chronique Cardiaque
- Maladie Chronique du Foie