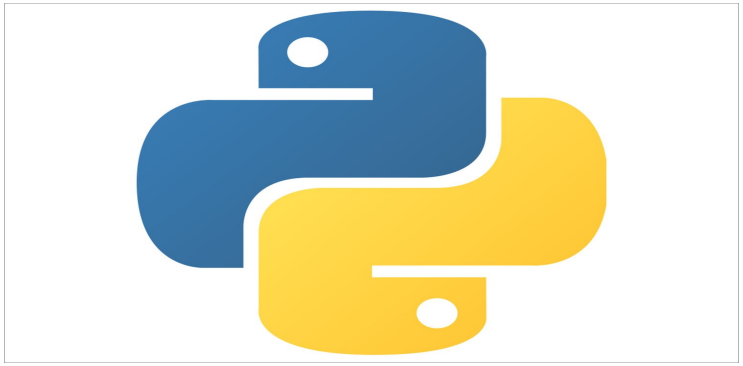
Formation sur la création d'une application simple de Machine Learning



Dr Yaya TRAORE
Maître de conférences en Informatique UJKZ
Email : yaytra@gmail.com

Bibliothèque

- **Python** : 1 langage de programmation en data sciences

- **Scikit-learn** est une bibliothèque Python libre et Open Source destinée à l'apprentissage automatique

- **Pandas** est une bibliothèque écrite pour le langage de programmation Python permettant la manipulation et l'analyse des données

- **Streamlit** est un framework Python open source qui est en mesure de transformer des scripts de données en applications web

Liens pour apprendre

- Pour ceux d'entre vous intéressés par la DATA SCIENCE, avec python : lien utile pour démarrer :
- https://youtu.be/xE97torN8zM
- https://www.youtube.com/channel/UCE-613S-bsuLukwHDhnRxIA/?sub_confirmation=1
- https://www.youtube.com/watch?v=u0Syto1oAGA
- https://www.youtube.com/watch?v=VsXWs4AvxqM

Outils pour utiliser l'IA sans savoir coder

- Teachable Machine
- What-If Tool
- Google AI Platform
- Data Robot
- RapidMiner Studio
- Microsoft Azure Automated Machine Learning
- BigML

Agenda

- Démarche de travail
- Compréhension des données
- Prétraitement – Pre-processing
- Construction d’un modèle
- Evaluation du modèle
- Sauvegarde et utilisation du modèle

Démarche de travail

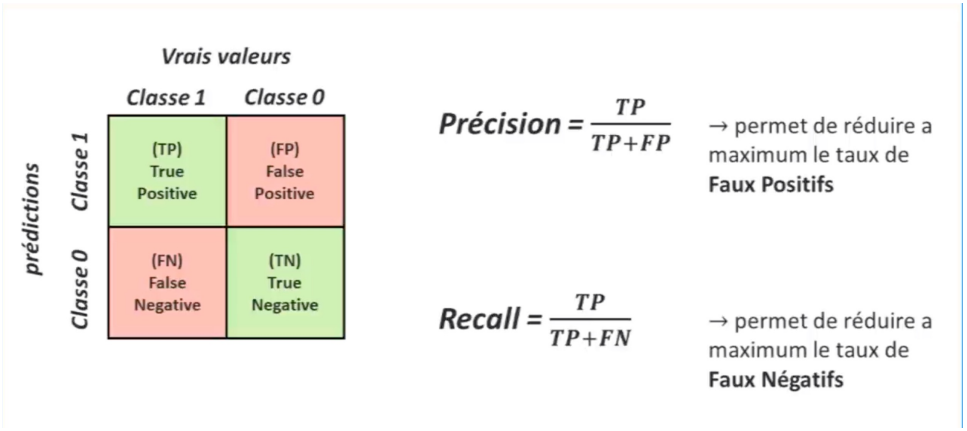
1. Définir un objectif mesurable
2. Analyse et exploration des données
3. Prétraitement – Pre-processing
4. Construction d’un modèle
5. Evaluation du modèle
6. Sauvegarde et utilisation du modèle

Démarche de travail

1. Définir un objectif mesurable

- Objectif : Prédire si une personne est infectée en fonction des données cliniques disponible.
- **Métrique :** Accuracy → 90%

- **Métrique :** Précision→60%, Rappel (sensibilité) → 70%, F1→50%



Démarche de travail

2. Analyse et Exploration de nos Données (EDA = Exploratory Data Analysis)

- Objectif : Comprendre au maximum les données dont on dispose pour définir une stratégie de modélisation.
  - **Analyse de la forme :** Identification de la target, Nombre de lignes et de colonnes, Types de variables, Identification des valeurs manquantes,……..
  - **Analyse du fond :** visualisation de la target (histogramme/ Boxplot), compréhension des différentes variables, visualisation des relations features-target (histogramme/ Boxplot), identification des outliers

Démarche de travail

3. Pétraitement (Pre-processing)

- Objectif : Transformer le data pour le mettre dans un format propice au Machine Learning
  - Checklist de base
    - Définir une fonction d'évaluation
    - Entrainement de différents modèles
    - Optimisation avec GridSearchCV
    - Analyse des erreurs et retour au prétraitement / EDA
    - Learning Curve et prise de décision

Démarche de travail

4. Modélisation

- Objectif : Développer un modèle de machine learning qui réponde à l'objectif final.
  - Checklist de base
    - Création du Train set / Test Set
    - Elimination des NaN : dropna(), imputation, colonnes vides, Encodage, Suppression des outliers néfastes au modèle, Feature Selection, Feature Engineering, Feature Scaling

Démarche de travail

4. Modélisation

- Objectif : Développer un modèle de machine learning qui réponde à l'objectif final.
  - Checklist de base
    - Création du Train set / Test Set
    - Elimination des NaN : dropna(), imputation, colonnes vides, Encodage, Suppression des outliers néfastes au modèle, Feature Selection, Feature Engineering, Feature Scaling