Mini-projet IA (classification) : Risques Cardiovasculaires

1. Objectif

L'objectif général de ce mini-projet est la création et comparaison des performances de plusieurs modèles de classification, en appliquant les principes d'apprentissage automatique (machine learning) sur un ensemble de données (Data Set).

L'objectif des modèles de classification à développer est de prédire les risques de développer une maladie cardiovasculaire.

Généralement, la création d'un modèle de classification est un processus impliquant un ensemble d'étapes, à savoir : compréhension des données, préparation des données, modélisation et évaluation.

En réalisant ce mini-projet, vous allez ainsi mettre en pratique l'ensemble des principes théoriques de ce processus en utilisant le langage **Python** et son écosystème (ensemble de packages) d'apprentissage automatique.

2. Organisation et déroulement

- Le mini-projet doit être réalisé sous forme d'un **Notebook** sous **Jupyter Notebook** ou tout autre plateforme équivalente.
- Les présentations des projets auront lieu le mercredi **22 mai 2024**.
- Le **Notebook** (**.ipynb**) est le seul fichier à remettre. Cela veut dire que vous devez ajouter des cellules de type **Markdown** pour commenter et expliquer les différentes étapes de réalisation de votre projet.
- Veillez à **bien présenter** votre **notebook** en ajoutant des titres, des descriptions, des explications, des couleurs, des figures, etc.

3. Ensemble des données (Data Set)

Les données du dataset de ce mini-projet, proviennent d'une enquête réalisée par le Système de Surveillance des Facteurs de Risque Comportementaux (The Behavioral Risk Factor Surveillance System), un outil utilisé par de nombreux pays pour collecter, suivre et analyser des données sur les comportements liés à la santé de leur population.

Le dataset est constitué généralement de caractéristiques liées au mode de vie des personnes (citoyens américains) qui ont participé à cette enquête.

Les caractéristiques de cet ensemble de données concernent des informations d'ordre démographiques telles que l'âge, le sexe, les antécédents médicaux ainsi que des attributs liés à l'alimentation, l'activité physique, les habitudes de tabagisme, la consommation d'alcool, les antécédents médicaux familiaux, et plus encore.

Essentiellement, il s'agit d'une collection de données visant à capturer différents aspects de la vie et de la santé d'un individu qui pourraient potentiellement contribuer au risque de développer des maladies cardiovasculaires.

L'ensemble de données est constitué de 308 854 lignes et 19 colonnes.

Voici le **dictionnaire de données** :

Variable	Définition	Valeurs
General_Health	Diriez-vous qu'en général, votre santé est...	Very Good Good Excellent
Checkup	Depuis combien de temps environ n'avez-vous pas consulté un médecin pour un examen de routine ?	Within the past year Within the past 2 years Within the past 5 years 5 or more years ago Never

Exercise	Au cours du dernier mois, avez-vous participé à des activités physiques ou à des exercices, comme la course ?	Yes-No
Heart_Disease	Répondants ayant déclaré souffrir d'une maladie cardiovasculaire ?	Yes-No
Skin_Cancer	Répondants ayant déclaré avoir un cancer de la peau ?	Yes-No
Other_Cancer	Répondants ayant déclaré avoir d'autres types de cancer ?	Yes-No
Depression	Répondants ayant déclaré souffrir d'un trouble dépressif ?	Yes-No
Diabetes	Répondants ayant déclaré souffrir de diabète. Si oui, de quel type de diabète s'agit-il ?	No Yes No, pre-diabetes or borderline diabetes Yes, but female told only during pregnancy
Arthritis	Répondants ayant déclaré souffrir d'arthrite ?	Yes-No
Sex		Female-Male
Age_Category		50-54 55-59 60-64 65-69 70-74
Height_(cm)		
Weight_(kg)		
BMI	Indice de masse corporelle	
Smoking_History		Yes-No
Alcohol_Consumption	combien de fois buvez-vous de l'alcool par mois ?	
Fruit_Consumption	combien de fois mangez-vous des fruits par mois ?	
Green_Vegetables	combien de fois mangez-vous des légumes par mois ?	
FriedPotato_Consumption	combien de fois mangez-vous des frites par mois ?	

L'objectif est de développer un modèle de classification permettant de prédire le risque de développer une maladie cardiovasculaire (***Heart_Disease***) afin d'aider les médecins et les personnes à risque à prévenir cette maladie mortelle.

4. Spécifications techniques

Le mini-projet doit être réalisé en respectant les spécifications techniques suivantes :

- Langage de programmation : **Python**
- Environnement de développement **: Jupyter Notebook** ou autre plateforme équivalente
- Packages du calcul scientifique et manipulation des données **: numpy et pandas**
- Package de visualisation des données : **matplotlib, seaborn** ou autre.
- Package d'apprentissage automatique (machine learning) : **scikit-learn**

5. Spécifications fonctionnelles

Au cours de ce mini-projet vous allez aborder l'ensemble des étapes du processus de Machine Learning.

Alors, voici l'ensemble des exigences que vous devez satisfaire (et que vous devez intégrer dans le notebook) en réalisant ce mini-projet.

5.1. Data understanding (compréhension des données) :

- **Chargement** de l'ensemble des données (Data set)
- **Affichage des données** : afficher un aperçu des 10 premières instances de l'ensemble des données.
- **Description** et **analyse** des données de l'ensemble des données : Afficher le **volume** (nombre total d'instances) et la **dimension** des données (nombre total des

attributs), le **type** et le codage des attributs et quelques **statistiques descriptives** (moyenne, écart-type, quartiles, valeur minimale, valeur maximale, etc.).

Analyser et **interpréter** les différentes valeurs.

- **Visualisation des données** : afin d'approfondir votre compréhension des données et chercher d'éventuelles **corrélations** entre la variable cible et les attributs prédictifs de l'ensemble des données, vous devez réaliser plusieurs types de **graphiques**, notamment des histogrammes, des nuages de points et des boîtes à moustaches.

  Analyser et **interpréter** les différents graphiques.

5.2. Nettoyage des données

Rechercher et éliminer les **doublons** dans les **enregistrements**. Afficher dans un tableau, le nombre de doublons de l'ensemble des données. **Supprimer** les doublons détectés, puis afficher de nouveau, le nombre de doublons de l'ensemble des données afin de prouver leur disparition.

Détection et traitement des **valeurs manquantes** de l'ensemble des données. Afficher dans un tableau, le nombre de valeurs manquantes pour chaque attribut de l'ensemble des données. Sur la base de votre compréhension des données, proposer puis appliquer une technique de traitement des valeurs manquantes. Afficher de nouveau, le nombre de valeurs manquantes pour chaque attribut de l'ensemble des données afin de prouver leur disparition.

5.3. Transformation des données

- En fonction de votre compréhension des données, et **si nécessaire**, proposez (avec justification) la **suppression** des **attributs prédictifs** que vous estimez **non discriminants** (non pertinents) par rapport à la création des modèles de classification. Ensuite, afficher un aperçu des données après l'application de cette transformation, le cas échéant.
- En fonction de votre compréhension des données, et **si nécessaire**, proposer (avec justification) la **création** de **nouveaux attributs prédictifs** à partir des autres attributs. Ensuite, afficher un aperçu des données après l'application de cette transformation, le cas échéant.
- **Normaliser**, **si nécessaire** (avec justification), les valeurs des **attributs** prédictifs **numériques**. Ensuite, afficher un aperçu des données après l'application de la normalisation.
- Appliquez (avec justification) les **transformations** nécessaires pour convertir les colonnes de type **catégoriel** en type **numérique**. Ensuite, afficher un aperçu des données après l'application de ces transformations.

5.4. Création et validation des modèles

Après les étapes de compréhension et de préparation des données, il est temps de créer des **modèles de classification supervisée**.

L'objectif est de créer **quatre modèles** de classification, à savoir, les **k-NN (KNeighborsClassifier)**, les **Arbres de Décision (DecisionTreeClassifier)**, la **régression logistique** (**LogisticRegression)** et les **Support Vector Machines (SVM)**. Ensuite, vous devez comparer leurs performances et choisir bien évidemment celui qui donne les meilleurs résultats de classification.

Dans ce mini-projet nous considérons : **Exactitude** (**Accuracy**), **précision**, **rappel** (recall) ainsi que **F1-Score** comme étant les **mesures** d'évaluation des **performances** pour tous les modèles.

Pour atteindre cet objectif, vous devez :

- Diviser **l'ensemble des données** en deux sous-ensembles :
  - **Ensemble d'apprentissage** : constitué de **75%** des lignes du dataset initial.
  - **Ensemble de test** : constitué de **25%** des lignes du dataset initial.
- En utilisant la technique de **validation croisée, ajuster** (optimiser) les **hyperparamètres** des quatre modèles de classification.

  Voici les hyperparamètres que vous devez ajuster :
  - **k-NN** : ***n_neighbors*** (nombre des plus proches voisins : tester différentes valeurs) et ***metric*** (*mesure de distance* : 'euclidien', 'Manhattan') ;
  - **Arbres de décision** : ***criterion*** *(fonction qui mesure la qualité de découpage :* 'entropy', 'gini') ;
  - **Régression logistique** : ***max_iter*** (nombre maximum d'itérations) et ***solver*** (algorithme d'optimisation : 'lbfgs', 'liblinear', 'newton-cg', 'sag', 'saga')

- **Support Vector Machines** : ***kernel*** *('linear', 'poly', 'rbf', 'sigmoid')* et ***C*** (*100, 10, 1.0, 0.1, 0.001)*

- Après ajustement, afficher la **matrice de confusion** (sous forme d'une heatmap) ainsi que les différentes mesures de performances (Accuracy, Précision, Rappel et F1-score) du modèle le plus performant pour chaque type d'algorithme.
- Après ajustement, afficher les **courbes** de **Précision-Rappel** et de **ROC-AUC** du modèle le plus performant pour chaque type d'algorithme.
- Après ajustement, **afficher** les valeurs des **meilleurs paramètres** pour chaque type algorithme.
- **Comparer** les performances des quatre modèles et choisir le meilleur.
- Utiliser le sous-ensemble d'apprentissage pour construire le modèle de classification le plus performant en employant l'algorithme et les hyperparamètres qui ont présenté les meilleurs résultats de performances durant la phase de validation.

5.5. Test du modèle

Selon les résultats obtenus dans la phase précédente, vous devez appliquer le modèle de classification le plus performant sur l'ensemble de **test**, afficher la matrice de confusion (sous forme d'une heatmap) ainsi que les mesures de performances suivantes :

- Accuracy
- Précision
- Rappel
- F1-score