Entité de formation : Institut de Formation et de Recherche en Informatique (IFRI)

Grade / Année académique : Licence 2 IA / 2023-2024

Titre de la thématique : Projet de l'UE « *Concepts de base de l'apprentissage automatique* »

Enseignant : Dr Ing. (MA) Ratheil V. HOUNDJI

Sujet : Méthodes d'apprentissage supervisée

1. Justification :

Développer des modèles d'apprentissage automatique supervisé avec la librairie scikit-learn

2. Objectifs

2.1. Objectif général

Le but de ce TP est de construire des modèles d'apprentissage automatique en utilisant principalement la librairie *Scikit-learn*, **tout en suivant une méthodologie rigoureuse**.

2.2. Objectifs spécifiques

A la fin de ce projet, l'apprenant sera capable de :

- Réaliser convenablement un pré-traitement (*preprocessing)* avant la modélisation,
- Mettre en place des modèles d'apprentissage automatique avec la librairie Scikit-learn,
- Evaluer des modèles d'apprentissage automatique,
- Identifier le surapprentissage dans un modèle d'apprentissage automatique,
- Valider le modèle construit,
- Prédire sur de nouvelles observations avec les modèles,

3. Prérequis

Dans le cadre de la réalisation de ce projet, on partira du principe que les apprenants ont des notions théoriques et pratiques sur les méthodes d'apprentissage automatique en général et celles d'apprentissage supervisé en particulier.
De même, les apprenants devront-être en mesure de :

- Importer un dataset avec la librairie pandas ;
- Réaliser des opérations classiques de nettoyage sur des données structurées ;
- Effectuer une analyse statistique descriptive d'un ensemble de données (univariée et multivariée);
- Communiquer les résultats d'analyse via figures et visuels avec les librairies matplotlib et seaborn ;
- Entrainer des modèles d'apprentissage automatique supervisés ;
- Evaluer des modèles d'apprentissage automatique supervisés.

4. Matériel et outils nécessaires

Ce projet sera réalisé sous Python avec les librairies Pandas, Matplotlib, Seaborn et Scikit-Learn essentiellement

5. Description du projet

Les données (disponibles sous Kaggle) utilisées dans le cadre de ce TP sont des données réelles provenant d'une compagnie d'assurance américaine X et sont utiles pour prédire les primes d'assurance maladie d'un citoyen américain. En effet, de nombreux facteurs influencent le montant des primes d'assurance maladie fixées par les compagnies d'assurance et qui sont indépendants de la volonté des assurés. Entre autres facteurs influençant le coût des primes d'assurance maladie, nous avons :

- L'âge du principal bénéficiaire (age)
- Le sexe de l'assureur (sex)
- L'indice de masse corporelle, permettant de comprendre le corps, les poids relativement élevés ou faibles par rapport à la taille, indice objectif de poids corporel (kg/m^2) utilisant le rapport taille/poids, idéalement 18,5 à 24,9 (bmi)
- Le nombre d'enfants couverts par l'assurance maladie ou le nombre de personnes à charge (children)
- Le statut de l'assuré par rapport au tabagisme, fumeur ou non (smoker)
- La zone résidentielle du bénéficiaire aux États-Unis, nord-est, sud-est, sud-ouest, nord-ouest (region)
- Les frais médicaux individuels (primes) facturés par l'assurance maladie (charges)

En tant que responsable de la cellule Informatique Décisionnelle et Gestion de Portefeuille, vous disposez de données mises à disposition par le département IT pour proposer un système intelligent de prédiction des primes d'assurance de clients désirant souscrire à un produit d'assurance maladie auprès de la compagnie X.
Pour ce faire, dans le cadre de ce projet, votre mission est de :

Partie 1 : Analyse + modèle KNN de base

a. Faire une première description pour présenter les caractéristiques de la dataset
b. Faire l'analyse descriptive des données :
    - Faire une analyse descriptive univariée des variables quantitatives ;
    - Faire une analyse descriptive univariée des variables qualitatives ;
    - A travers des boîtes à moustache, faites une analyse descriptive bivariée entre la variable cible et les variables qualitatives ;
    - A travers des nuages de points, faites une analyse descriptive bivariée entre la variable cible et les variables quantitatives (ne pas hésiter à intégrer des variables qualitatives pour affiner l'analyse) ;
    - Concevoir une matrice de corrélation entre les variables quantitatives de la base de données.
    - Interpréter l'ensemble des résultats issus de l'analyse descriptive.
c. Faire le pré-traitement des données
    - Convertir les variables qualitatives au format approprié pour la modélisation

- Recoder les modalités des variables qualitatives au format approprié pour la modélisation
- Diviser la base de données en échantillon d'apprentissage (80%) et en échantillon test (20%)
- Séparer la variable cible des variables explicatives.

d. Construire un premier modèle basé sur l'algorithme des *k plus proches voisins*
   - Construire un premier modèle M0 à partir des paramètres par défaut de l'algorithme
   - Modifier le paramètre K pour construire trois autres modèles M1, M2 et M3, respectivement pour K=3, K=7 et K=10
   - Evaluer les performances de chacun de ces modèles
   - Interpréter vos résultats

e. Optimiser le modèle pour K=10
   - Optimiser les hyperparamètres ***weights*** *(méthode de pondération des voisins : uniform ou distance)****, p*** *(paramètre de choix de la distance : Manhattan ou Euclidienne) et* ***algorithm*** *(algorithme utilisé pour calculer les plus proches voisins : ball tree, kd_tree, brute, auto).*
   - Evaluer les performances du modèle utilisé
   - Comparer vos résultats au modèle par défaut
   - Faites le choix du modèle optimal

f. Analyser le meilleur modèle
   - Représenter un nuage de points qui nous permet de visualiser la relation entre les valeurs réelles et les valeurs prédites
   - Faire une analyse des résidus issus du meilleur modèle: nuage de points entre les valeurs prédites et les résidus issus du modèle
   - Interpréter les résultats

g. Faire une synthèse générale de l'étude
   - Commenter l'ensemble des résultats
   - Tirer une conclusion
   - Proposer des solutions

h. Utiliser le modèle optimal pour prédire sur la nouvelle vague de clients voulant souscrit à une assurance maladie dans la compagnie X (feuille PREDICT).

<u>Partie 2 : Autres modèles d'apprentissage supervisé</u>

Tester plusieurs autres modèles suivant un protocole expérimentale rigoureux (choix des modèles, optimisation des hyperparamètres, caractéristiques importantes, évaluations, interprétations des résultats).

(Bonus) Implémenter de petites interfaces graphiques qui vont prendre de nouvelles entrées dans un formulaire et prédire la sortie en utilisant le meilleur modèle.