Rapport de Traitement et Analyse des Données d'Assurance Maladie

Introduction

Ce rapport présente les étapes de traitement, d'analyse et de modélisation des données d'assurance maladie contenant plus de 4 millions de lignes. Les données incluent des informations sur les effectifs de patients pris en charge par divers régimes d'assurance maladie, classées par pathologie, traitement chronique, épisode de soins, sexe, classe d'âge, région et département.

Objectifs

1. Nettoyer et préparer les données pour l'analyse.
2. Imputer les valeurs manquantes pour améliorer la qualité des données.
3. Construire et évaluer des modèles de machine learning pour prédire la prévalence des pathologies et les coûts des soins.

Description des Données

Les données comportent les colonnes suivantes :

- **annee** : Année de l'enregistrement
- **patho_niv1, patho_niv2, patho_niv3** : Catégorisation des pathologies
- **top** : Indicateur binaire de la pathologie principale
- **cla_age_5** : Classe d'âge en tranches de 5 ans
- **sexe** : Sexe du patient (1 pour masculin, 2 pour féminin)
- **region** : Code de la région
- **dept** : Code du département
- **Ntop** : Nombre de patients avec la pathologie principale
- **Npop** : Population totale
- **prev** : Prévalence de la pathologie
- **Niveau prioritaire** : Niveau de priorité de la pathologie
- **libelle_classe_age** : Libellé de la classe d'âge
- **libelle_sexe** : Libellé du sexe

Étapes de Traitement des Données

1. **Chargement des Données** :
    - Les données ont été chargées dans un DataFrame Spark pour une manipulation efficace.
2. **Identification des Valeurs Manquantes** :
    - Un comptage des valeurs manquantes a été effectué pour chaque colonne.
3. **Imputation des Valeurs Manquantes** :
    - **Colonnes Numériques** : Les colonnes numériques (`Ntop`, `prev`) ont été imputées en utilisant la moyenne des valeurs non manquantes.
    - **Colonnes Catégorielles** : La colonne `patho_niv2, patho_niv3, Niveau prioritaire` ont été imputée avec la valeur la plus fréquente (mode) ou une valeur par défaut si aucune valeur n'était disponible.

Modélisation

1. **Modèle de Régression pour la Prévalence** :
    - **Algorithme Utilisé** : Régression par Forêt Aléatoire.
    - **Entraînement et Test** : Les données ont été divisées en ensembles d'entraînement (80%) et de test (20%).
    - **Évaluation** : Le modèle a été évalué à l'aide de la mesure RMSE (Root Mean Square Error).
    - **Résultat** : Le modèle a obtenu un RMSE de 3.018, indiquant que les prédictions du modèle sont en moyenne à 3 unités de la valeur réelle de prévalence.
2. **Modèle de Randomforest pour la classification des Patients** :
    - **Algorithme Utilisé** : RadomForestClassifier
    - **Entraînement et Test** : Les mêmes étapes d'entraînement et de test que pour
    - **Évaluation** : Le modèle a été évalué pour prédire si un patient a une certaine pathologies

Analyse et Conclusion

- **Performance du Modèle** : Le RMSE de 3.018 pour le modèle de prévalence est un indicateur raisonnable, mais pourrait être amélioré avec des techniques de feature engineering, de sélection de caractéristiques, et de tuning des hyperparamètres.
- **Qualité des Données** : L'imputation des valeurs manquantes a amélioré la qualité des données, permettant des analyses plus fiables.
- **Prochaines Étapes** : Pour améliorer les performances, il serait bénéfique d'explorer des modèles plus complexes comme les Gradient Boosting Machines ou les réseaux neuronaux, et de continuer à affiner les techniques de traitement des données.