TELECOM SudParis

IP PARIS

Projet Spark

Hoe Ziet WONG

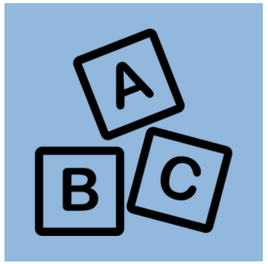
Sommaire

- Sélection du jeu de données et objectif
- Chaîne de traitement
- Utilisation de Spark
- Résultats
- Démonstration

Jeu de données

- Données sur la santé générale issu d'un sondage CDC au Etats-Unis
- Exemples de colonnes : Diabètes, Hypertension, BMI, âge, cholestérol, etc.
- Source de jeu de données : Kaggle
- Objectif : **Multi-Class Classification** pour prédire l'état diabétique à partir des autres données santé

Colonnes du dataset



Catégorique

HighBP
HighChol
CholCheck
Smoker
Stroke
HeartDiseaseorAttack
PhysActivity
Fruits
Veggies
AnyHealthcare
NoDocbcCost
GenHlth
DiffWalk
Sex
Age
Education
HvyAlcoholConsum
Income

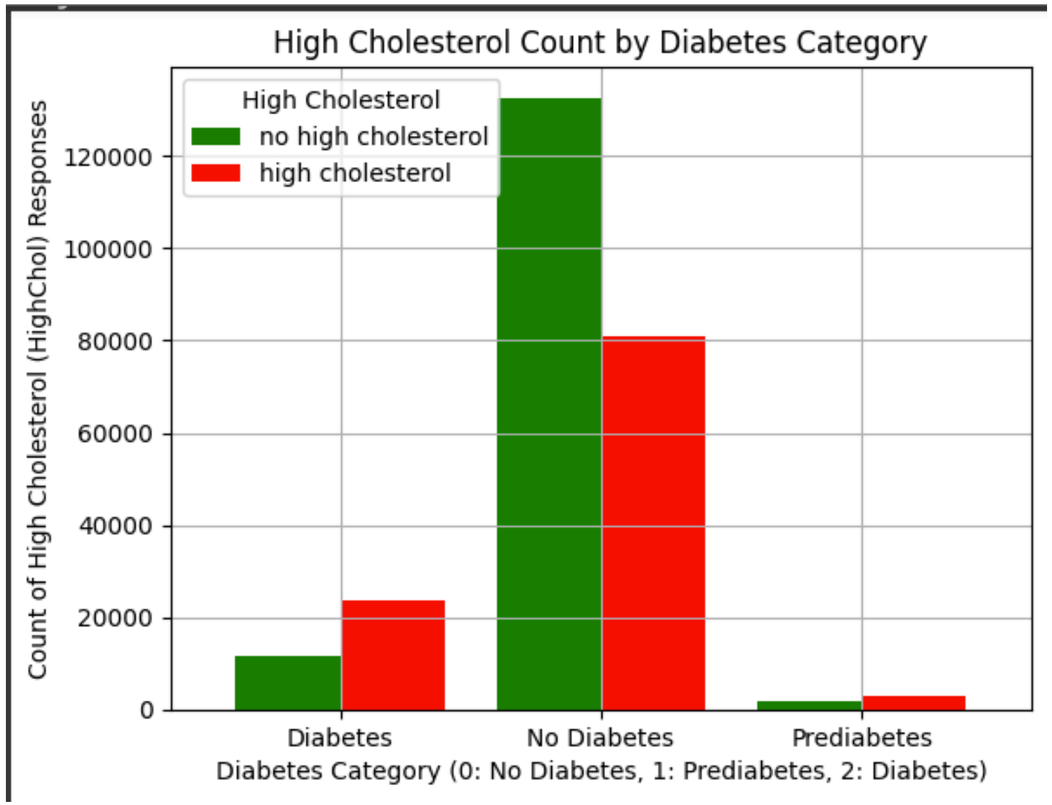
Numérique

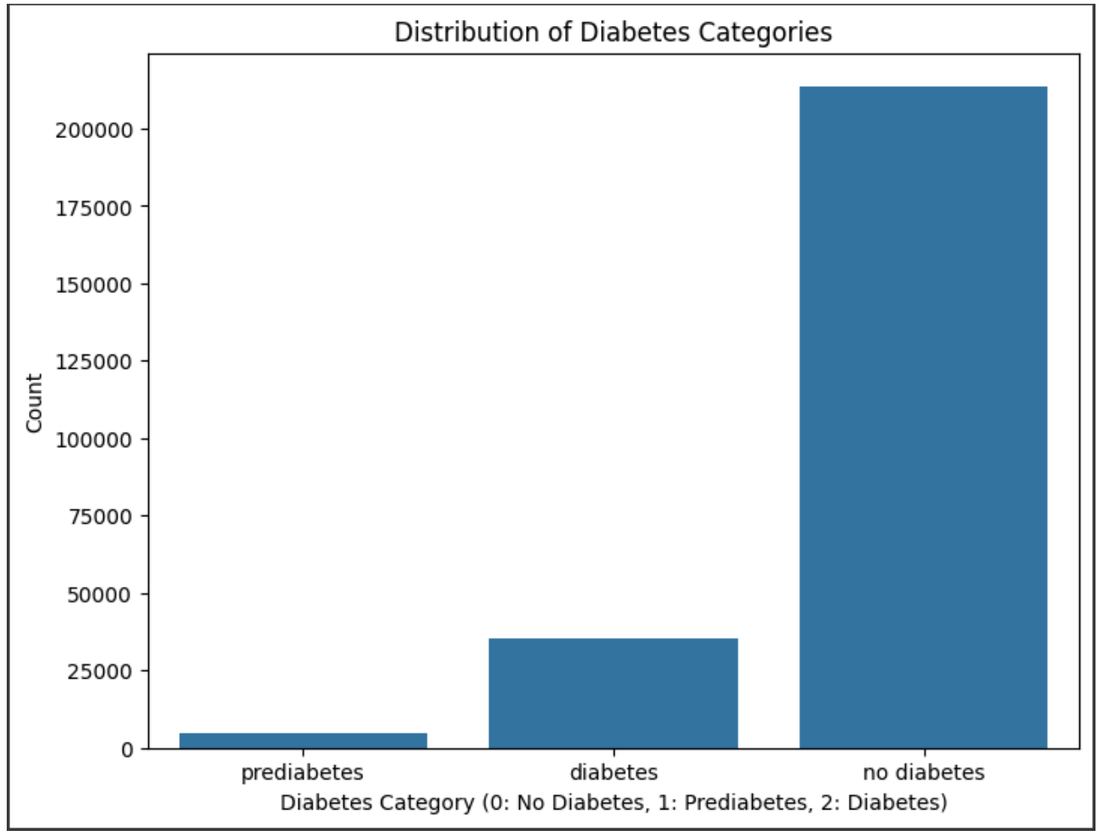
MentHlth
PhysHlth
BMI

Chaîne de traitement

Data analysis	Preprocessing	Machine Learning	Tuning des params
• Lire les données • Analyse sur la correlation des features avec le label • Illustrations via graphes • Suppression de certains features	• Forcer le type des colonnes numériques • Encoder les colonnes catégoriques • Scaling des colonnes numériques	• Logistic Regression • Decision Tree • Random Forest • Evaluation des modèles	• GridSearch

Exemple du data analysis





AnyHealthcare	Diabetes_012	no	yes
0	diabetes	1422	33924
1	no diabetes	10741	202962
2	prediabetes	254	4377

Exemple du preprocessing

Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits
no diabetes	high BP	high cholesterol	yes cholesterol c...	40.0	yes	no	no	no	no
no diabetes	no high BP	no high cholesterol	no cholesterol ch...	25.0	yes	no	no	yes	no
no diabetes	high BP	high cholesterol	yes cholesterol c...	28.0	no	no	no	no	yes
no diabetes	high BP	no high cholesterol	yes cholesterol c...	27.0	no	no	no	yes	yes
no diabetes	high BP	high cholesterol	yes cholesterol c...	24.0	no	no	no	yes	yes
no diabetes	high BP	high cholesterol	yes cholesterol c...	25.0	yes	no	no	yes	yes
no diabetes	high BP	no high cholesterol	yes cholesterol c...	30.0	yes	no	no	no	no
no diabetes	high BP	high cholesterol	yes cholesterol c...	25.0	yes	no	no	yes	no
diabetes	high BP	high cholesterol	yes cholesterol c...	30.0	yes	no	yes	no	yes
no diabetes	no high BP	no high cholesterol	yes cholesterol c...	24.0	no	no	no	no	no
diabetes	no high BP	no high cholesterol	yes cholesterol c...	25.0	yes	no	no	yes	yes
no diabetes	high BP	high cholesterol	yes cholesterol c...	34.0	yes	no	no	no	yes
no diabetes	no high BP	no high cholesterol	yes cholesterol c...	26.0	yes	no	no	no	no
diabetes	high BP	high cholesterol	yes cholesterol c...	28.0	no	no	no	no	no
no diabetes	no high BP	high cholesterol	yes cholesterol c...	33.0	yes	yes	no	yes	no
no diabetes	high BP	no high cholesterol	yes cholesterol c...	33.0	no	no	no	yes	no
no diabetes	high BP	high cholesterol	yes cholesterol c...	21.0	no	no	no	yes	yes
diabetes	no high BP	no high cholesterol	yes cholesterol c...	23.0	yes	no	no	yes	no
no diabetes	no high BP	no high cholesterol	no cholesterol ch...	23.0	no	no	no	no	no
no diabetes	no high BP	high cholesterol	yes cholesterol c...	28.0	no	no	no	no	no

label	features
0.0	(27,[3,4,6,11,12,...
0.0	(27,[0,1,3,4,8,11...
0.0	(27,[2,3,4,5,11,1...
0.0	(27,[1,2,3,4,5,6,...
0.0	(27,[2,3,4,5,6,7,...
0.0	(27,[3,4,5,6,7,13...
0.0	(27,[1,3,4,8,11,1...
0.0	(27,[3,4,6,8,11,1...
1.0	(27,[3,5,6,11,12,...
0.0	(27,[0,1,2,3,4,6,...
1.0	(27,[0,1,3,4,5,6,...
0.0	(27,[3,4,5,6,8,11...
0.0	(27,[0,1,3,4,6,8,...
1.0	(27,[2,3,4,6,10,1...
0.0	(27,[0,4,6,10,11,...
0.0	(27,[1,2,3,4,7,11...
0.0	(27,[2,3,4,5,6,8,...
1.0	(27,[0,1,3,4,7,15...
0.0	(27,[0,1,2,3,4,6,...
0.0	(27,[0,2,3,4,7,21...

Utilisation de Spark

PySpark.SQL

- DataFrame
- SparkSession
- RandomSplit()

PySpark.ML

- Pipeline
- StringIndexer, OneHotEncoder
- VectorAssembler, StandardScaler
- LogisticRegression, DecisionTreeClassifier, RandomForestClassifier
- MulticlassClassificationEvaluator
- CrossValidator, ParamGridBuilder

Résultats des modèles ML

Sans Tuning

Accuracy of LogisticRegression model:
0.8451483223416275

Accuracy of DecisionTree model:
0.8448987098977903

Accuracy of RandomForest model:
0.840891773299351

Après Tuning

Accuracy of LogisticRegression model:
0.8453059723061562

Accuracy of DecisionTree model:
0.8448987098977903

Accuracy of RandomForest model:
0.8424551354475945

Meilleur modèle

LogisticRegression

Best Params :

Best regParam: 0.001
Best elasticNetParam: 0.0
Best maxIter: 10

DEMO TIME