POLYTECHNIQUE MONTRÉAL

UNIVERSITÉ D'INGÉNIERIE

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

MTH6312 - MÉTHODES STATISTIQUES D'APPRENTISSAGE

Rapport du Projet de Session - Automne 2024

Thématique: Détection précoce du risque d'AVC

Automne 2024

Apprenants :

Gervais Presley Koyaweda – 2305686

Jesuton Turibe Zinsou-ply - 2188765

Enseignant :

Luc Adjengue

1. Introduction

1.2. Contexte du Projet

Dans le cadre du cours "Méthodes Statistiques d'Apprentissage", ce projet vise à appliquer les techniques enseignées à un problème concret. Nous avons choisi un dataset disponible sur Coursera, spécifiquement dans le projet intitulé "**Showcase: Build and Deploy a Stroke Prediction Model**". Ce dataset se prête parfaitement à une analyse prédictive visant à identifier les patients à risque d'accident vasculaire cérébral (AVC), une problématique majeure en santé publique. La mise en œuvre de modèles d'apprentissage automatique permet de démontrer l'utilité des approches statistiques dans un contexte pratique.

Lien du projet : https://www.coursera.org/projects/showcase-build-and-deploy-a-stroke-prediction-model-with-r#details

1.3 Problématique

Comment développer un modèle prédictif robuste capable d'identifier les patients à haut risque d'AVC en exploitant des données cliniques et démographiques, afin de faciliter des interventions préventives ciblées et améliorer les résultats pour les patients ?

Objectifs Spécifiques

De manière spécifique, les objectifs sont :

- **Analyser et caractériser les données disponibles** pour identifier les variables explicatives critiques associées aux AVC.
- **Développer un modèle prédictif performant** intégrant des méthodologies avancées de machine learning.
- **Valider les performances du modèle** à l'aide de techniques rigoureuses d'évaluation statistique et de validation croisée.

Méthodologie

Afin de répondre à cette problématique, une approche structurée a été adoptée :

1. **Analyse exploratoire des données (AED)** :

- Analyse structurelle des données pour examiner leur composition, identifier les variables pertinentes et évaluer leur qualité.
- Exploration approfondie pour détecter des tendances et des relations entre les variables, en particulier celles associées au risque d'AVC.

2. **Prétraitement des données** :
   - Imputation des valeurs manquantes (KNN).
   - Encodage des variables qualitatives en représentations binaires (OneHot Encoding).
   - Application de SMOTE pour le rééquilibrage des classes.
3. **Développement et optimisation des modèles prédictifs** :
   - Entraînement de plusieurs modèles (régression logistique, KNN, arbres de décision, forêts aléatoires et gradient boosting).
   - Optimisation des hyperparamètres du meilleur modèle avec GridSearchCV.
   - Évaluation rigoureuse basée sur des métriques telles que l'AUC, le rappel et le F1-score.

Analyse Exploratoire des Données

Analyse de la Structure des Données

Le jeu de données contient **5110 observations** réparties sur **12 variables**, comprenant des variables entières, continues et qualitatives. La variable cible stroke, binaire, est fortement déséquilibrée : seulement **4.87%** des cas sont positifs. Une exploration initiale a permis d'identifier **3.93% de valeurs manquantes** dans bmi, nécessitant une imputation.

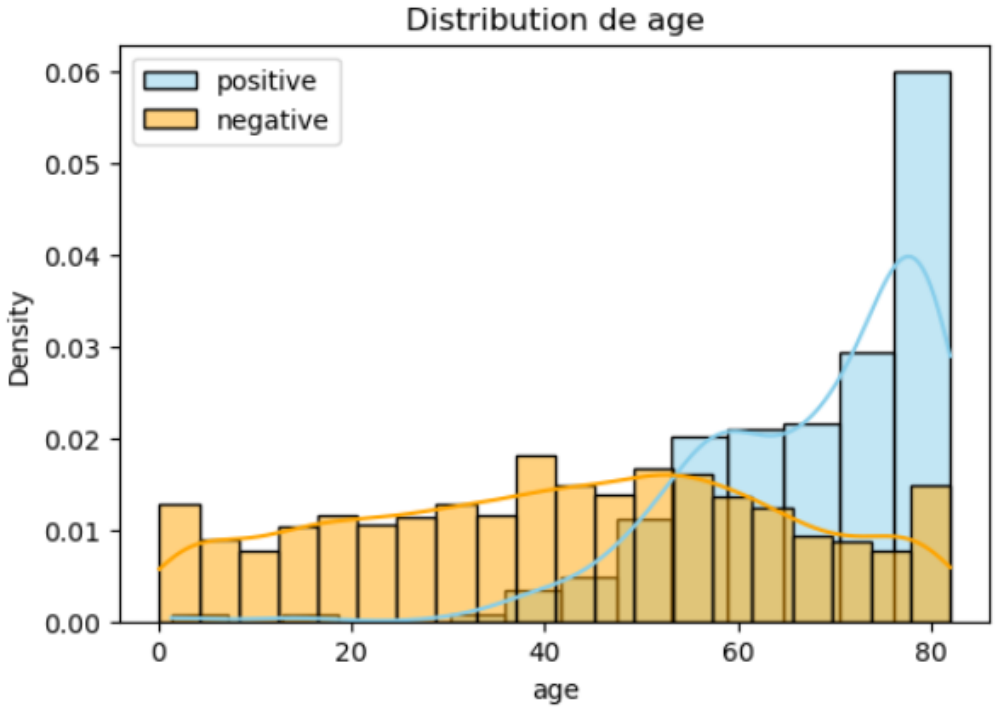
```
Repartition de la variable

stroke
0    95.127202
1     4.872798
```

```
Proportion des valeurs manquantes

gender               0.000000
age                  0.000000
hypertension         0.000000
heart_disease        0.000000
ever_married         0.000000
work_type            0.000000
Residence_type       0.000000
avg_glucose_level    0.000000
smoking_status       0.000000
stroke               0.000000
bmi                  3.933464
dtype: float64
```

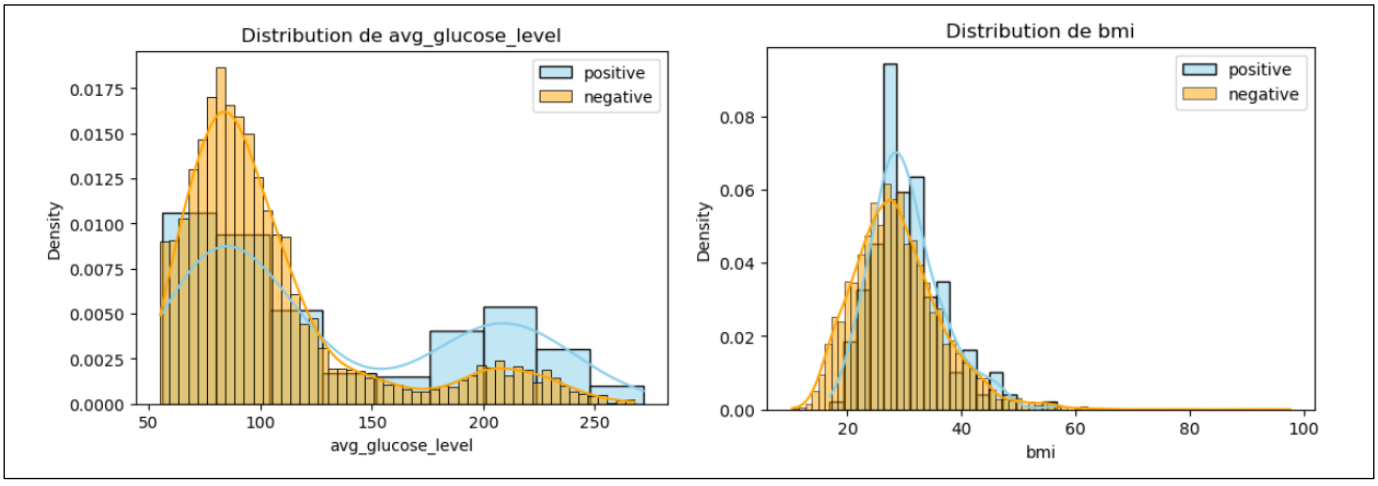
Analyse de Fond

L'analyse approfondie a mis en évidence des corrélations cliniquement significatives :

- Les individus âgés de plus de **60 ans** présentent un risque accru d'AVC.



- Des niveaux élevés de glucose sanguin et un indice de masse corporelle (IMC) supérieur à la moyenne sont fréquemment associés à des AVC.



- L’hypertension et les maladies cardiaques augmentent de manière significative le risque d’AVC.

```
Tableau croisé pour hypertension (en pourcentage par ligne):

hypertension      0      1
stroke
0             91.11   8.89
1             73.49  26.51

Tableau croisé pour heart_disease (en pourcentage par ligne):

heart_disease      0      1
stroke
0              95.29   4.71
1              81.12  18.88
```

Tests Statistiques

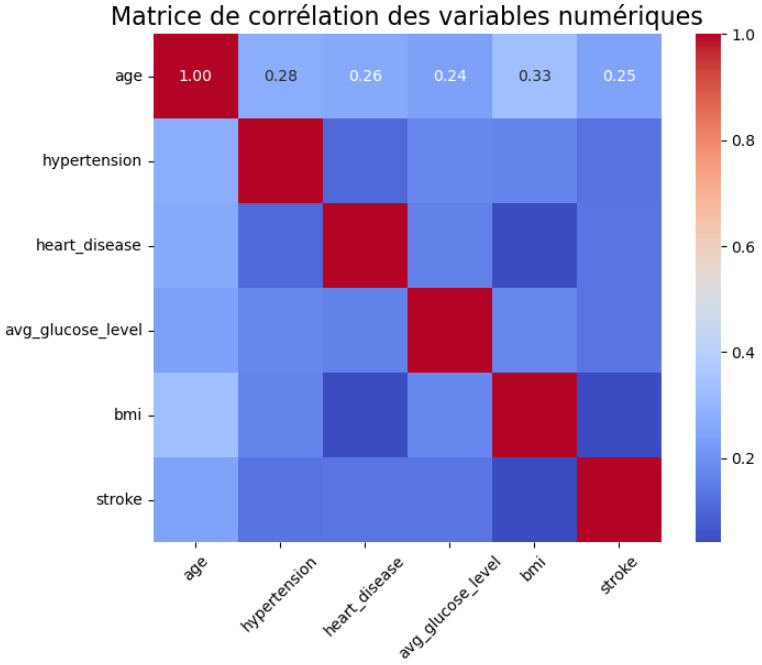
- **Tests de normalité** : Les variables continues ne suivent pas une distribution normale.
- **Analyse de variance (ANOVA)** : Les différences entre les groupes stroke sont statistiquement significatives pour age, avg_glucose_level et bmi.

```
Target et la variable age
 Test de normalite de Shapiro-Wilk: Statistique = 0.9672, p-value=0.0000
 Test de Levene d'égalite de la variance : Statistique =129.7366, p-value=0.0000
  ANOVA (comparaison des moyennes): Statistique = 326.9166, p-value = 0.0000

Target et la variable avg_glucose_level
 Test de normalite de Shapiro-Wilk: Statistique = 0.8059, p-value=0.0000
 Test de Levene d'égalite de la variance : Statistique =94.1085, p-value=0.0000
  ANOVA (comparaison des moyennes): Statistique = 90.5039, p-value = 0.0000

Target et la variable bmi
 Test de normalite de Shapiro-Wilk: Statistique = 0.9535, p-value=0.0000
 Test de Levene d'égalite de la variance : Statistique =10.4249, p-value=0.0013
  ANOVA (comparaison des moyennes): Statistique = 8.8265, p-value = 0.0030
```

- **Tests de chi-carré** : Les variables qualitatives telles qu'hypertension et heart_disease montrent une association forte avec stroke.

```
Target et la variable hypertension
  Test de Chi-carré: Statistique = 81.6054, p-value = 0.0000

Target et la variable heart_disease
  Test de Chi-carré: Statistique = 90.2596, p-value = 0.0000

Target et la variable ever_married
  Test de Chi-carré: Statistique = 58.9239, p-value = 0.0000

Target et la variable work_type
  Test de Chi-carré: Statistique = 49.1635, p-value = 0.0000

Target et la variable Residence_type
  Test de Chi-carré: Statistique = 1.0816, p-value = 0.2983

Target et la variable smoking_status
  Test de Chi-carré: Statistique = 29.1473, p-value = 0.0000

Target et la variable gender
  Test de Chi-carré: Statistique = 0.4726, p-value = 0.7895
```

Pas de problème de multicolliearité.

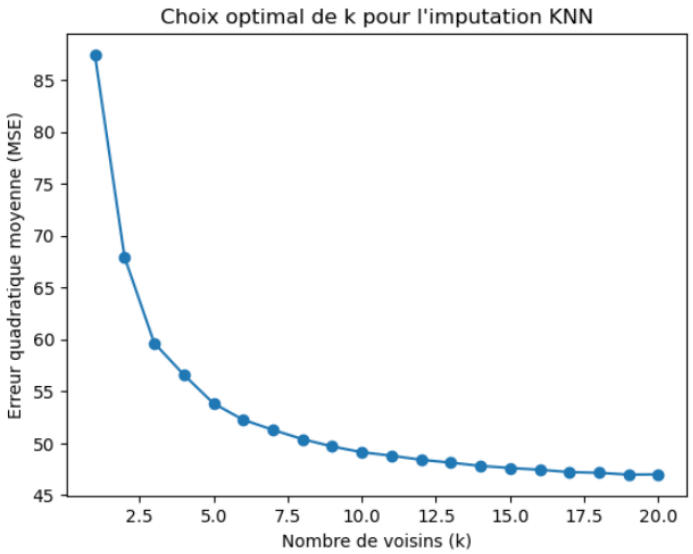


Prétraitement des Données

Imputation des Valeurs Manquantes

La méthode KNN Imputer a été utilisée pour imputer les valeurs manquantes de bmi. Un paramètre optimal de **19 voisins** a été déterminé par validation croisée (K-Fold), minimisant l'erreur quadratique moyenne (MSE).

```
Meilleur k : 19
Erreurs pour chaque k : [87.41954879575478, 67.92875080189592, 59.656495473281794, 56.59699437336377, 53.838410222164086, 52.26
573497130014, 51.294547940695736, 50.39715596696837, 49.69437840556656, 49.13992672251392, 48.80469657499664, 48.4122602949765
9, 48.13439835177339, 47.824553149550084, 47.620677077442664, 47.45031730914521, 47.22408890618882, 47.15457043333406, 46.98241
88887484, 47.005448934033815]
```



Encodage des Variables Catégorielles

Les variables qualitatives ont été transformées en représentations binaires à l'aide de One-Hot Encoding, garantissant leur compatibilité avec les modèles prédictifs.

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke	ever_married_Yes	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_t
0	67	0	1	228	36	1	1	0	1	0	
1	61	0	0	202	34	1	1	0	0	1	
2	80	0	1	105	32	1	1	0	1	0	
3	49	0	0	171	34	1	1	0	1	0	
4	79	1	0	174	24	1	1	0	0	1	
...	...	...	...	...	...	...	...	...	...	...	
5105	80	1	0	83	26	0	1	0	1	0	
5106	81	0	0	125	40	0	1	0	0	1	
5107	35	0	0	82	30	0	1	0	0	1	
5108	51	0	0	166	25	0	1	0	1	0	
5109	44	0	0	85	26	0	1	0	0	0	

5110 rows × 14 columns

Rééquilibrage des Classes

La méthode SMOTE a été appliquée pour créer des exemples synthétiques dans la classe minoritaire, équilibrant ainsi efficacement les proportions de la variable cible.

Entraînement des Modèles et Résultats

Entraînement Initial

Cinq modèles ont été évalués : régression logistique, KNN, arbres de décision, forêts aléatoires et gradient boosting. La méthode SMOTE a été utilisée pour équilibrer les classes avant l'entraînement. Le modèle KNN a émergé comme le plus performant, avec un rappel de **81%** pour la classe minoritaire et une AUC de **0.7809**.

```
=== Logistic Regression ===
=== Évaluation du modèle : LogisticRegression (Seuil=0.3) ===
Matrice de confusion :
[[3638 1223]
 [  93  156]]

Rapport de classification :
              precision    recall  f1-score   support

           0       0.98      0.75      0.85      4861
           1       0.11      0.63      0.19       249

    accuracy                           0.74      5110
   macro avg       0.54      0.69      0.52      5110
weighted avg       0.93      0.74      0.81      5110
```

```
=== KNN ===
=== Évaluation du modèle : KNeighborsClassifier (Seuil=0.3) ===
Matrice de confusion :
[[3183 1678]
 [  48  201]]

Rapport de classification :
              precision    recall  f1-score   support

           0       0.99      0.65      0.79      4861
           1       0.11      0.81      0.19       249

    accuracy                           0.66      5110
   macro avg       0.55      0.73      0.49      5110
weighted avg       0.94      0.66      0.76      5110
```

```
=== Decision Tree ===
=== Évaluation du modèle : DecisionTreeClassifier (Seuil=0.3)
Matrice de confusion :
[[4385  476]
 [ 185   64]]

Rapport de classification :
              precision    recall  f1-score   support

           0       0.96      0.90      0.93      4861
           1       0.12      0.26      0.16       249

    accuracy                           0.87      5110
   macro avg       0.54      0.58      0.55      5110
weighted avg       0.92      0.87      0.89      5110
```

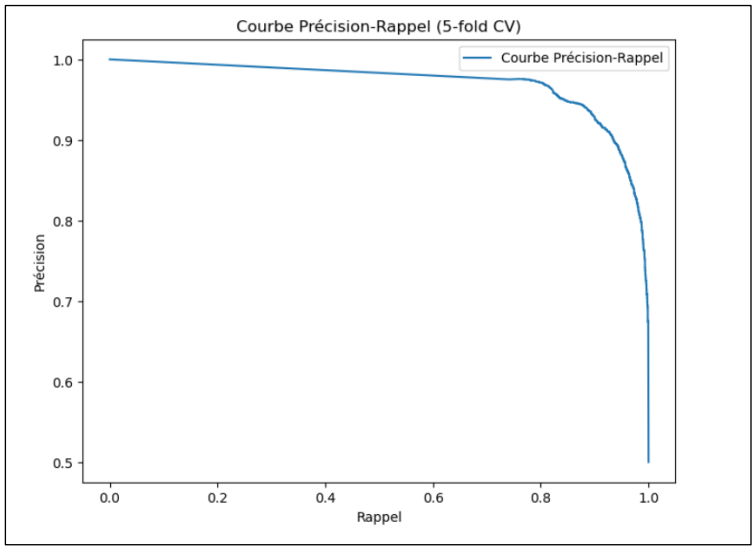
```
=== Random Forest ===
=== Évaluation du modèle : RandomForestClassifier (Seuil=0.3)
Matrice de confusion :
[[4175  686]
 [ 160   89]]

Rapport de classification :
              precision    recall  f1-score   support

           0       0.96      0.86      0.91      4861
           1       0.11      0.36      0.17       249

    accuracy                           0.83      5110
   macro avg       0.54      0.61      0.54      5110
weighted avg       0.92      0.83      0.87      5110
```

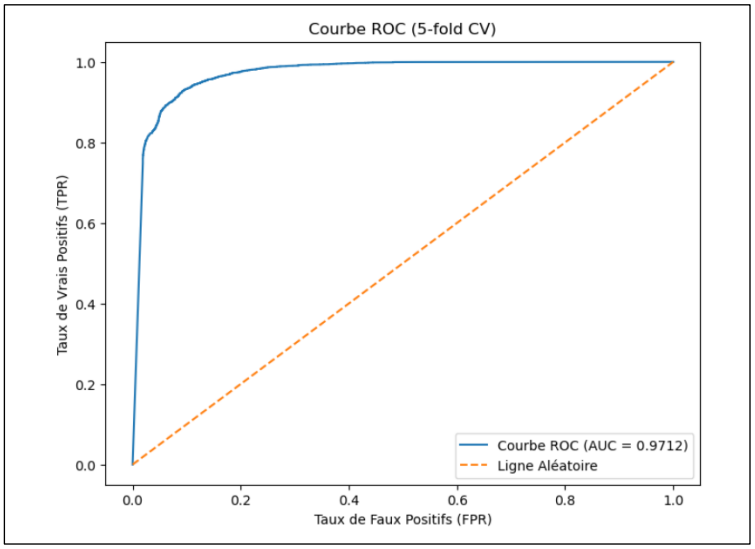
```
=== Gradient Boosting ===
=== Évaluation du modèle : GradientBoostingClassifier (Seuil=0.3)
Matrice de confusion :
[[3804 1057]
 [ 111  138]]

Rapport de classification :
              precision    recall  f1-score   support

           0       0.97      0.78      0.87      4861
           1       0.12      0.55      0.19       249

    accuracy                           0.77      5110
   macro avg       0.54      0.67      0.53      5110
weighted avg       0.93      0.77      0.83      5110
```

Optimisation des Hyperparamètres

Le modèle KNN a été optimisé à l'aide de GridSearchCV, en explorant différentes configurations de métriques et de poids. La configuration optimale (« Manhattan » et poids proportionnels à la distance) a permis d'atteindre une AUC de **0.9713**.

```
Rapport de classification (5-fold CV) :
              precision    recall  f1-score   support

           0       0.99      0.64      0.78      4861
           1       0.74      0.99      0.85      4861

    accuracy                           0.82      9722
   macro avg       0.86      0.82      0.81      9722
weighted avg       0.86      0.82      0.81      9722

AUC Score (5-fold CV) : 0.9712

Matrice de confusion :
[[3131 1730]
 [  29 4832]]
```





Évaluation et Interprétation des Résultats

- **Précision** : 99% pour la classe majoritaire.
- **Rappel** : 99% pour la classe minoritaire, réduisant significativement les faux négatifs.
- **F1-Score** : 85%, reflétant un bon équilibre entre précision et rappel.
- **AUC** : 0.9712, confirmant une capacité discriminante excellente.

Conclusion

Le modèle KNN optimisé, combiné avec SMOTE pour le rééquilibrage des classes, s'est révélé performant pour la prédiction des AVC. Ce projet démontre l'application concrète des méthodes statistiques d'apprentissage dans un contexte pédagogique, en intégrant une démarche complète d'analyse, de prétraitement, et de modélisation. Les résultats obtenus montrent que ces techniques peuvent être déployées pour résoudre des problématiques réelles.