Rapport d'Analyse Exploratoire

Étude de cas : *Modélisation des risques AVC*

Étudiants : *Adama Sall*

Les différentes étapes

___ Chargement et inspection des données

___Correction des types et nettoyage

___Imputation des valeurs manquantes

___Traitement des valeurs extrêmes

___Analyse univariée et bivariée

___Visualisation des distributions suivie des test d'hypothèse.

1. Description des données

- **Source :** (non renseigné pour le moment)
- **Nombre d'observation:** 5 110
- **Variables :** 11 (3 numériques et 8 catégorielles)

Variable	Description	Type
Genre	Sexe du patient (Male, Female, Other)	Catégorique
Age	Âge du patient	Numérique
Hypertension	1 si le patient souffre d'hypertension, 0 sinon	Catégorique
Maladie_Cardiaque	1 si le patient a une maladie cardiaque, 0 sinon	Catégorique
Situation_Matrimoniale	Statut matrimonial (Yes/No)	Catégorique
Type_travail	Type d'emploi (Private, Self-employed, etc.)	Catégorique
Residence	Lieu de résidence (Urban/Rural)	Catégorique

Taux_glucose_moyen	Moyenne du taux de glucose	Numérique
IMC	Indice de masse corporelle	Numérique
Statut_Fumer	Statut tabagique (never smoked, formerly smoked, smokes, unknown)	Catégorique
AVC	1 si AVC, 0 sinon	Catégorique

3. Inspection initiale

- **Types de variables :** 3 numériques, 8 catégorielles. Certaines de ces variables n'étaientt pas de type nomal.
- **Valeurs manquantes :**
- 201 valeurs manquantes sur la variable IMC.
- Statut_Fumer : 1 544 valeurs « Unknown » nous l'avons consideré comme valeur manquante d'après la description reçue (remplacées par NaN)
- **Doublons :** Aucun doublon détecté
- Seul la variable **IMC** suit la distribution normale par la confirmation d'un test de shapiro
- Detection des valeurs aberrantes sur la variable « Taux moyen de glucose» et « IMC»

4. Nettoyage et traitement des données

Correction des types

- Transformation des variables « Hypertension », « Maladie_Cardiaque » et « AVC » en catégories.

Traitement des valeurs manquantes

- **IMC :** Nous avons fais une imputation par la moyenne en se justifiant que la distribution de **IMC** est symetrique (Moyenne=mdedian=mode) avec un écart-type faible.
- **Statut_Fumer :** Remplacement des « Unknown » par NaN, puis imputation par le mode (never smoked).

Traitement des valeurs aberrantes

- Méthode des bornes (IQR) appliquée à « Age », « Taux_glucose_moyen », « IMC ».Toute valeur inferieur à Q1 sera remplacer par Q1 et toute valeur superieurà Q3 sera remplacer par Q3

5. Analyse univariée

Variables numériques apres traitement

Variable	Moyenne	Écart-type	Min	25%	Médiane	75%	Max	Skewness	Kurtosis
Age	43,2	22,6	0,08	25	45	61	82	-0.14	-0.99
Taux_glucose_moyen	101,0	33,2	55,12	77,2	91,9	114,1	169,4	0.94	-0.17
IMC	28,7	7,1	10,3	23,8	28,4	32,8	46,3	0.44	-0.08

Normalité (test de Shapiro-Wilk) :

- **Age** et **Taux_glucose_moyen** : distributions non normales (p-value < 0.05)
- **IMC** : distribution non normale (p-value < 0.05)

Visualisation

- Apres traitement les distributions sont globalement symétriques pour l'IMC, légèrement asymétriques et presentattion de pic pour le glucose et IMC.

Variables catégorielles

Variable	Modalités principales	Effectif	Pourcentage
Genre	Female	2 994 sur 5110	58,6 %
Hypertension	0	4 612 sur 5110	90,3 %
Maladie_Cardiaque	0	4 834 sur 5110	94,6 %
Situation_Matrimoniale	Yes	3 353 sur 5110	65,6 %

Type_travail	Private	2 925 sur 5110	57,2 %
Residence	Urban	2 596 sur 5110	50,8 %
Statut_Fumer	never smoked	3 436 sur 5110	67,3 %
AVC	0 (pas d'AVC)	4 861 sur 5110	95,1 %

Visualisation

- Les barplots montrent un déséquilibre important sur la variable cible (peu d'AVC).
- La majorité des patients sont des femmes.
- La majorité des patients ne souffrent pas d'hypertension, ni de maladie cardiaque.
- La majorité des patients ne fument pas ou n'ont jamais fumé
- La majorité des patients sont dans le privé.
- Il y'a un equilibre sur les patients selon leur résidence

6. Analyse bivariée

Linéarité entre variables numériques

Aucune relation linéaire detectée entre les variables.

Corrélations entre variables numériques

- Corrélation moderée entre Age et IMC.
- Corrélation très faible entre Age et Taux_glucose_moyen.
- Corrélation faible entre IMC et Taux_glucose_moyen .

Ces correlations ont été tester (spearman) et toutes les p-values sont < 0.05, ce qui indique que toutes les corrélations sont statistiquement significatives.

Croisement avec la variable cible (AVC)

- L'AVC est légèrement plus fréquent chez les femmes que les hommes.
- La majorité des cas AVC ont un statut matrimonial
- La répartition des patients ayant AVC ou non est équitable selon leur résidence.
- Chez les patients AVC, On remarque que y'a plus d'ancien fumeur.

Relation des variables catégorielle avec la variable cible (Test de chi 2)

Toutes les variables ont une relation statistiquement significative avec la variable cible (AVC). Aucune relation significative detectée entre le genre et l'AVC , et entre la résidence et l'AVC, selon les résultats du test du chi carré.

- Les patients ayant eu un AVC présentent plus souvent de l'hypertension ou une maladie cardiaque.
- L'AVC est plus fréquent chez les personnes mariées.
- Les patients ayant eu un AVC sont souvent des fumeurs et des anciens fumeurs.
- Les AVC sont sont plus fréquents chez les personnes travaillant dans le secteur privé, mais aussi chez les travailleurs indépendants et les employés du secteur public. Les enfants et ceux qui n'ont jamais travaillés sont moins exposés.

Relation des variables quantitative avec la variable cible (AVC)

- Les patients ayant eu un AVC sont en moyenne plus âgés et presente un taux de glucose plus importante.
- Les patients ayant eu un AVC ont en moyenne un IMC légèrement supérieur à ce qui n'en ont pas.

8. Conclusion

Variable ciblée : âge, hypertension, maladie cardiaque, tabagisme, taux de glucose

- **Déséquilibre de la cible** à prendre en compte lors de la modélisation (techniques de rééchantillonnage recommandées).
- **Prochaines étapes :** modélisation prédictive (régression logistique, GEV , GPD, POT, random forest, gradient boosting ,SVM réseau de neurone.).