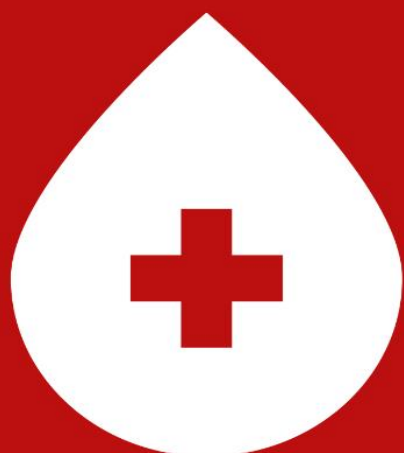


**NK STAT
CONSULTING
2025**



**Analyse et Visualisation
des campagnes de dons
de sang au Cameroun**

Plan du Rapport

05

Introduction

1. Contexte
2. Objectifs
3. Plan du rapport

09

Présentation des données et Méthodologie d'analyse

1. Vue d'ensemble
2. Structure et contenu des données
3. Méthodologie

22

Présentation des principaux résultats

1. Cartographie de la Répartition des Donneurs
2. Conditions de Santé & Éligibilité
3. Profilage des Donneurs Idéaux
4. Efficacité des Campagnes
5. Caractéristiques des donneurs effectifs de sang

Prédiction d'éligibilité et AI assistant

Limites et recommandations

Conclusion

Bibliographie

44

49

51

I. CONTEXTE

Au Cameroun, le don de sang reste un enjeu critique de santé publique. Avec un système de santé en développement constant et des besoins transfusionnels en hausse, notamment en raison des urgences obstétricales, des accidents de la route, des maladies comme le paludisme et la drépanocytose, ainsi que des interventions chirurgicales, le pays fait face à un défi permanent d'approvisionnement en produits sanguins. Selon les estimations du Centre National de Transfusion Sanguine (CNTS) du Cameroun, le pays n'atteint qu'environ 50% de ses besoins annuels en sang, ce qui entraîne des conséquences dramatiques pour de nombreux patients.



Les campagnes de don de sang organisées au Cameroun se heurtent à plusieurs obstacles, notamment

des croyances culturelles défavorables, une méconnaissance des procédures, et une répartition géographique inégale des centres de collecte entre zones urbaines et rurales. De plus, le faible taux de donneurs réguliers (moins de 20% selon les dernières statistiques nationales) complique davantage la situation.

Ce projet s'inscrit dans une volonté d'optimiser les futures campagnes de don de sang au Cameroun en exploitant les données collectées lors des précédentes initiatives, afin d'améliorer l'efficacité des collectes et de garantir un approvisionnement plus stable en produits sanguins à l'échelle nationale.

2. OBJECTIFS

Le tableau de bord développé dans le cadre de ce challenge vise à optimiser les campagnes de don de sang au Cameroun pour augmenter l'approvisionnement en produits sanguins.

Les trois objectifs spécifiques regroupent de manière logique les différentes analyses proposées :

- L'analyse géographique et démographique (qui et où)
- L'analyse des facteurs d'éligibilité et de rétention (pourquoi)
- Les recommandations stratégiques basées sur les analyses temporelles et qualitatives (comment)

3. STRUCTURE DU RAPPORT

Le présent rapport s'articule autour de 4 principaux chapitres. Dans un premier temps, une présentation des données et de la méthodologie sont faites. Ensuite, les principaux résultats sont présentés et le modèle de prédiction d'éligibilité de don de sang avec l'assistant AI. Enfin, les conclusions et recommandations sont présentées.



Chapitre 1

PRESENTATION DES DONNEES ET METHODOLOGIE D'ANALYSE

1. Vue d'ensemble

La base de données contient des informations détaillées sur les donneurs de sang potentiels au Cameroun, avec un total de **1915 enregistrements** (individus) et **39 colonnes** (variables). Cette collection représente un échantillon significatif pour l'analyse des caractéristiques des donneurs et des facteurs influençant l'éligibilité au don de sang dans le contexte camerounais.




2. Structure et contenu des données

Notre jeu de données sur le don de sang au Cameroun se compose de

1915 enregistrements individuels décrits par 39 variables différentes. La grande majorité des données sont de nature catégorielle (36 colonnes de type 'object'), complétées par deux variables numériques (de type 'float64') correspondant aux mesures physiques, et une variable temporelle (au format 'datetime64[ns]') pour les dates. Ce riche ensemble de données couvre plusieurs dimensions essentielles à la compréhension des profils des donneurs et des facteurs d'éligibilité au don de sang dans le contexte camerounais.

Les informations démographiques constituent une part importante des variables collectées, incluant les données personnelles (date de naissance, genre, taille, poids et situation matrimoniale) qui permettent d'établir le profil



fondamental des donneurs potentiels. Le niveau socio-économique est représenté par le niveau d'étude et la profession, tandis que la dimension géographique est capturée par l'arrondissement et le quartier de résidence, offrant une granularité spatiale précieuse pour l'analyse territoriale. Les caractéristiques culturelles, notamment la nationalité et la religion, complètent ce portrait démographique en apportant une dimension socioculturelle essentielle dans le contexte camerounais.

Les informations médicales et d'éligibilité forment le cœur analytique de la base de données, avec des variables documentant l'historique de don (expérience antérieure et date du dernier don), les paramètres médicaux critiques

comme le taux d'hémoglobine, et surtout, la variable cible d'éligibilité au don qui détermine la capacité du candidat à effectuer un don de sang. Cette variable centrale est complétée par un système détaillé de documentation des motifs d'inéligibilité, structuré en trois catégories principales : les indisponibilités temporaires (comme l'antibiothérapie en cours, un taux d'hémoglobine insuffisant, un don trop récent ou une infection sexuellement transmissible récente), les indisponibilités spécifiques aux femmes (liées au cycle menstruel, à l'allaitement, à un accouchement récent, à une interruption de grossesse ou à une grossesse en cours), et les causes d'inéligibilité permanente (antécédents de transfusion, statut sérologique positif pour le VIH, l'hépatite B ou C, interventions chirurgicales,

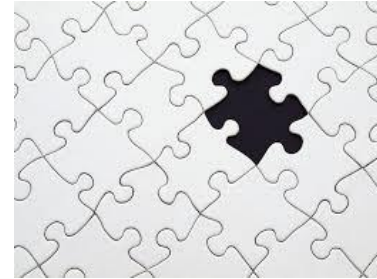
drépanocytose, diabète,
hypertension, asthme, problèmes
cardiaques, présence de tatouages
ou de scarifications).



3. Méthodologie

Cette section présente la méthodologie adoptée pour le traitement et l'analyse des données dans le cadre de cette étude. L'objectif principal est de détailler les différentes étapes de prétraitement, de nettoyage et d'analyse des variables clés. L'étude repose sur l'utilisation de Python et de bibliothèques telles que Pandas, fuzzywuzzy, scikit-learn et geopy pour assurer un traitement rigoureux des données.

3.1. Gestion des valeurs manquantes



Les valeurs manquantes ont été identifiées et traitées selon leur nature.

- Pour les variables numériques, l'imputation a été réalisée à l'aide de l'algorithme `IterativeImputer` de `scikit-learn`, en prédisant les valeurs manquantes à partir des autres variables numériques du jeu de données.
- Pour les variables catégoriques, l'imputation s'est faite en remplaçant les valeurs manquantes par la modalité la plus fréquente (mode).

3.2. Traitement des variables clés

3.2.1. *Traitement de la variable profession*

Le problème principal rencontré avec cette variable était la multiplication des formes d'écriture d'une même profession. Pour homogénéiser les données, les étapes suivantes ont été suivies :

- Création d'une variable normalisée '**Profession_Commune**' ;
- Utilisation d'un dictionnaire associant les différentes variations d'une profession à une seule catégorie commune ;
- Transformation de la variable en variables indicatrices (**dummy variables**) pour les besoins d'analyse quantitative ;

3.2.2. *Traitement des variable Nationalité et Région*

Une fonction (**transform_nation**) est définie pour nettoyer et standardiser les valeurs de la colonne 'Nationalité' (suppression des espaces et points, conversion en minuscules, remplacements spécifiques). Un algorithme est ainsi

construit pour cette tâche et pour la variable Religion.

- Les valeurs des variables sont nettoyées et standardisées :
 - Suppression des espaces et caractères spéciaux.
 - Conversion en minuscules.
 - Utilisation de dictionnaires de remplacement pour regrouper des termes similaires.

Cela diminue en effet le nombre de modalités de cette variable afin de permettre une bonne structuration de l'analyse et de la modélisation.

3.2.3. *Traitement des Données Temporelles*

Notre approche de traitement des dates repose sur plusieurs principes méthodologiques :

- **Détection des Incohérences** : Nous avons développé une méthode systématique pour identifier les dates invalides en utilisant la conversion au format datetime et en repérant les valeurs NaT (Not a Time).

- **Correction des Années** : Une fonction de transformation spécifique a été conçue pour corriger les années incorrectes. Par exemple, pour la '*Date de remplissage de la fiche*', les années invalides sont systématiquement remplacées par '2019'.
- **Normalisation** : En cas de dates totalement invalides, nous utilisons la date la plus fréquente (mode) parmi les dates valides comme valeur de remplacement.

Pour les dates de naissance, nous avons mis au point une stratégie de correction basée sur :

- L'analyse des deux derniers chiffres de l'année
- L'attribution d'un siècle (2000 ou 1900) en fonction de ces chiffres
- La gestion des erreurs de conversion avec l'option `errors='coerce'`

3.2.4 Quartier de résidence et arrondissement

Tout d'abord la colonne est convertie en minuscules et supprimé des caractères indésirables. La librairie

`fuzzywuzzy` est utilisée pour regrouper les noms de quartiers similaires en fonction d'un seuil de similarité de 90% en raison qu'il peut y avoir les noms des quartiers qui sont très similaires mais désignant bien deux zones totalement différentes. Ensuite un dictionnaire (`grouped_quartiers`) est généré pour représenter ces groupes. Une fonction (`get_grouped_quartier`) est ensuite utilisée pour créer une nouvelle colonne ('`vrai_quartier`') contenant les noms de quartiers standardisés. Pour contrôler une fois de plus la variable Arrondissement un fichier de données téléchargé sur Wikipédia a été utilisé pour vérifier la cohérence des quartiers à leur arrondissement pour ville de Douala.

3.2.3. Normalisation et géolocalisation de la variable Quartier

La récupération des coordonnées géographiques est ensuite réalisée via le service Nominatim de geopy, avec une recherche spécifique ciblant les quartiers de Douala, au Cameroun. En définitive, cette méthodologie a permis de

géolocaliser précisément 211 quartiers distincts, transformant des données textuelles brutes en informations géographiques exploitables, avec une latitude et une longitude pour chaque quartier. L'utilisation d'un fichier CSV préexistant ('Vrai_Quartier.csv') comme référentiel géographique et la suppression des quartiers non localisables garantissent la qualité et la précision du jeu de données géolocalisées.

En somme, les variables Profession et Quartier de résidence ont présenté des problèmes d'orthographe et d'incohérence dans la saisie par exemple, on pouvait trouver 'Commerçant (e)', 'COMMERCEANT', 'COMMERÇANTE'. Pour les quartiers le dictionnaire a permis d'avoir un dictionnaire de ce genre : 'logbaba': ['logbaba ', 'log-baba', 'log baba', 'logbaba']

Les étapes suivantes ont été mises en place pour résoudre ce problème :

- Conversion de toutes les valeurs en minuscules.

- Regroupement des noms similaires en utilisant la librairie fuzzywuzzy avec un seuil de similarité de 90 %.
- Géolocalisation des quartiers à l'aide de geopy et du service Nominatim d'OpenStreetMap.
- Exportation et importation des coordonnées géographiques pour réduction du temps de calcul.
- Suppression des quartiers non géolocalisés.
- Calcul et agrégation des indicateurs par quartier.

3.3. Modélisation et analyse

3.3.1. Analyse textuelle et sentimentale

Une analyse de sentiment a été réalisée sur la variable **Autre** à **préciser** en appliquant les méthodes suivantes :

- Suppression des **stop Words**.
- Construction d'un **nuage de mots** pour visualiser les termes les plus fréquents.

4. Analyses statistiques et modélisation du phénomène du statuts d'éligibilité

Analyses statistiques:

- **Statistiques descriptives** sont calculées pour comprendre les distributions des variables.
- **Tests du Chi-2** sont réalisés pour évaluer l'indépendance entre les variables catégorielles et leur relation avec l'éligibilité au don.
- **Analyses de la variance (ANOVA)** sont effectuées pour comparer les moyennes des variables numériques entre différents groupes (par exemple, l'âge en fonction de l'éligibilité ou du groupe sanguin).
- **Calcul des taux de croissance** annuels et mensuels du nombre de dons
- **Test de Kolmogorov-Smirnov** pour comparer les distributions d'âge entre les genres et le statut d'éligibilité.
- **Clustering** : L'algorithme K-Means est appliqué pour identifier des groupes (clusters) de donateurs\candidats en fonction de leurs caractéristiques démographiques et de leur historique de dons. Le nombre

optimal de clusters est déterminé à l'aide de la méthode du coude et du score de silhouette. Les résultats du clustering sont visualisés et profilés.

- **Analyse des Correspondances Multiples (ACM)** : L'ACM est utilisée pour analyser les relations entre plusieurs variables catégorielles simultanément et pour visualiser les proximités entre les modalités de ces variables⁷.... Des techniques de clustering (mélange gaussien et DBSCAN) sont appliquées aux coordonnées de l'ACM.

Analyse de l'efficacité des campagnes : L'historique des dons (date du dernier don) est analysé en relation avec les facteurs démographiques pour potentiellement évaluer l'impact des campagnes passées.

3.3.2 Modélisation du statut d'éligibilité

Le prétraitement a été mené de manière différenciée pour les variables numériques et catégorielles, en utilisant des techniques sophistiquées de la librairie scikit-learn.

Pour les variables catégorielles ('Niveau d\'etude', 'Genre', 'Situation Matrimoniale (SM)', 'Profession_Commune', 'Religion', 'A-t-il (elle) déjà donné le sang') ont suivi un traitement spécifique : imputation des valeurs manquantes par la modalité la plus fréquente, puis encodage via TargetEncoder. Cette méthode permet de capturer la relation entre les variables catégorielles et la variable cible. La variable d'intérêt étant le statut d'éligibilité au don est transformée en deux modalités : éligible et non éligible (définitivement ou temporairement).

L'imputation des valeurs manquantes des variables numériques a été réalisée selon les étapes suivantes :

- Sélection des variables '**Taux d'hémoglobine**', '**Poids**' et '**Taille**'.
- Application de **IterativeImputer** avec 10 itérations pour estimer les valeurs manquantes.

La division des données a été réalisée selon un ratio 80/20 pour les ensembles d'entraînement et de test,

avec un `random_state` de 42 garantissant la reproductibilité.

Nous avons opté ensuite pour un classificateur

`RandomForestClassifier` du fait de sa puissance à prédire pour les classes déséquilibrées, intégré dans un pipeline complet combinant prétraitement et classification.

Évaluation avec des métriques appropriées : Nous avons mis en évidence l'utilisation du **score F1** en plus de la précision et du rapport de classification. Le score F1 est une moyenne harmonique de la précision et du rappel, ce qui le rend plus sensible aux performances sur la classe minoritaire que la précision seule. L'utilisation de cette métrique suggère une prise en compte du déséquilibre potentiel des classes dans la variable cible '**ÉLIGIBILITÉ AU DON**'.

Validation croisée stratifiée : L'implémentation de **StratifiedKFold** assure que chaque pli de validation croisée contient des proportions approximativement égales de chaque classe. Ceci est crucial lors du travail avec des données déséquilibrées, car une validation croisée non stratifiée

pourrait conduire à des plis avec une représentation très faible, voire nulle, de la classe minoritaire, entraînant une évaluation biaisée du modèle.

Optimisation du seuil de probabilité : Cette technique est couramment utilisée pour ajuster le compromis entre la précision et le rappel, ce qui est particulièrement important dans les scénarios de classes déséquilibrées où l'on peut vouloir favoriser la détection de la classe minoritaire (rappel plus élevé) au détriment d'une augmentation potentielle des faux positifs (précision plus faible), ou vice versa.

Bien que ces éléments indiquent une certaine sensibilité au problème des classes déséquilibrées, Nous pourrions aussi mentionner l'utilisation de techniques explicites d'équilibrage des classes telles que :

- **Suréchantillonnage (Oversampling)** de la classe minoritaire.
- **Sous-échantillonnage (Undersampling)** de la classe majoritaire.
- **Génération de données synthétiques** pour la classe

minoritaire (par exemple, SMOTE).

Chapitre 2

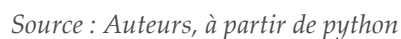
PRESENTATION DES RESULTATS





1.1. Distribution des candidats au don de sang à Douala

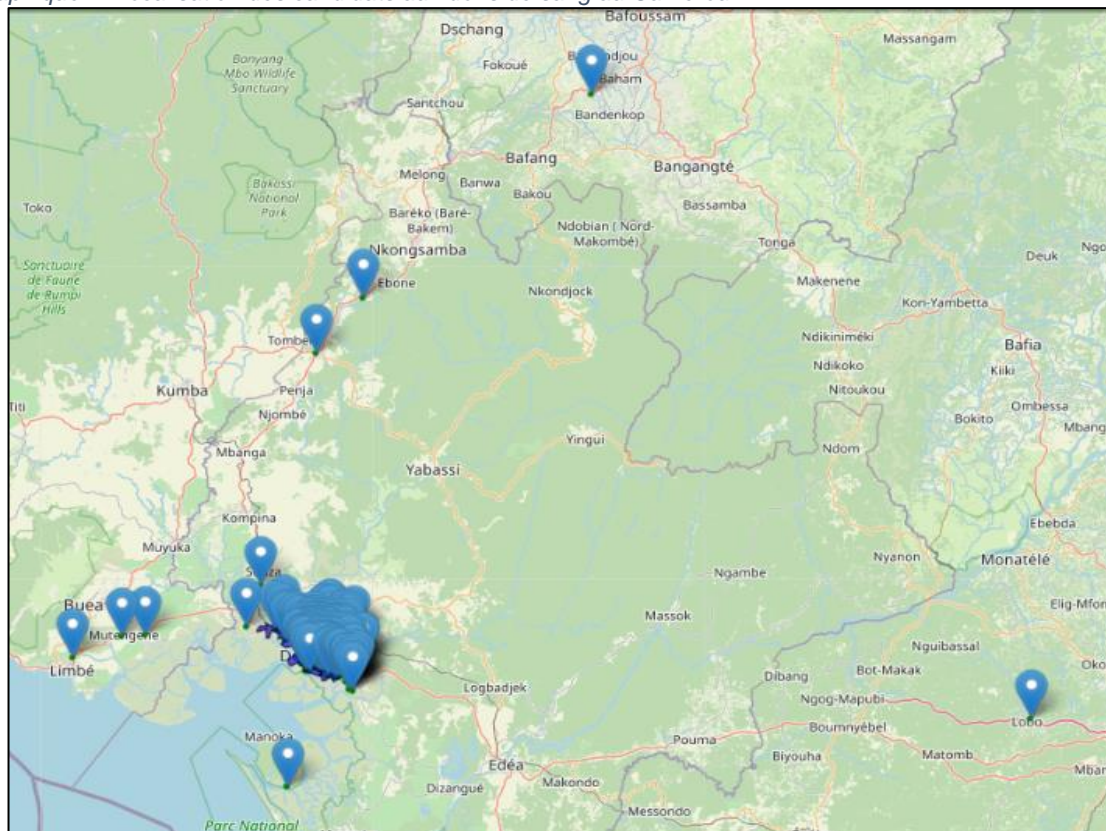
Graphique 1: Cartographie de la distribution des candidats au don de sang par arrondissement (Douala)



1.2. Localisation des candidats au don de sang sur le territoire national

Cette carte illustre la localisation géographique des centres de collecte de sang ou des donneurs, révélant une concentration très marquée dans la région urbaine de Douala (centre-sud de la carte) où un amas dense de marqueurs bleus est visible, tandis que d'autres points sont dispersés de façon beaucoup plus éparse sur le territoire : quelques-uns au nord vers Bafoussam et Bangangté, plusieurs le long de la côte ouest près de Limbé et Buea, des marqueurs alignés semblant suivre un axe routier principal, et des points isolés comme celui proche de Lobo à l'est. Cette répartition déséquilibrée suggère une forte centralisation des activités liées au don de sang dans les principales zones urbaines, reflétant probablement des disparités en termes de densité de population, d'infrastructures sanitaires disponibles et d'accessibilité, ce qui soulève des questions sur l'équité d'accès aux services de don de sang pour les populations des zones rurales ou éloignées du pays.

Graphique 2: Localisation des candidats aux dons de sang au Cameroun



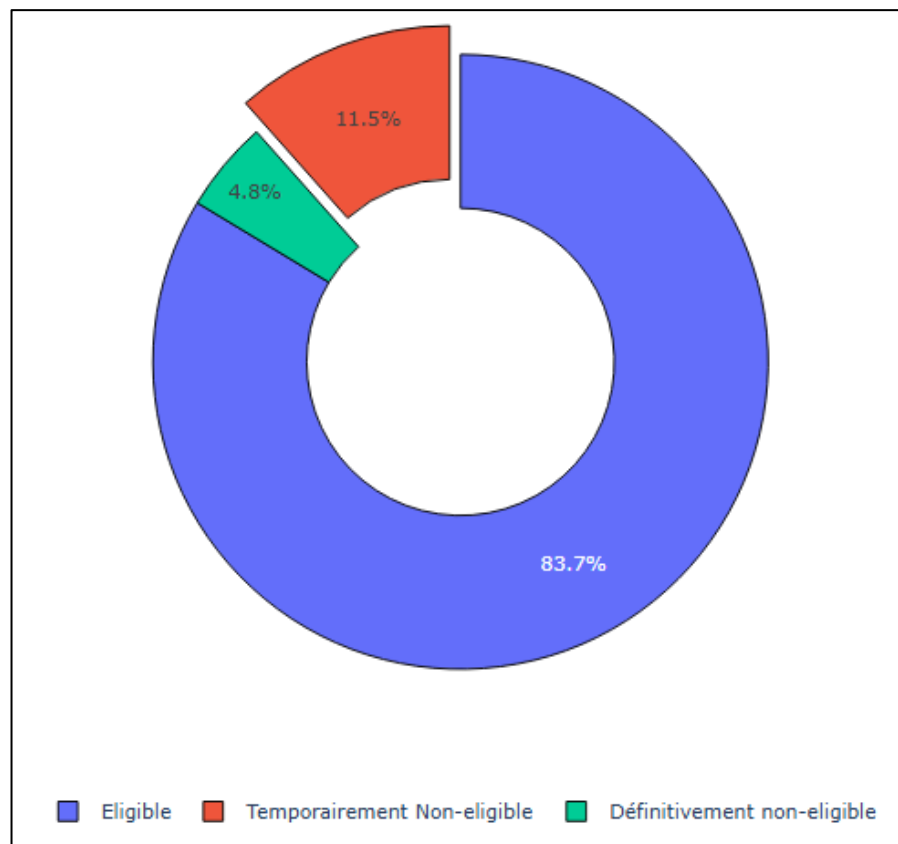
Source : Auteurs, à partir de python

2. Conditions de Santé & Éligibilité

2.1. Distribution des 1915 répondants selon leur éligibilité au don de sang

L'analyse de ce graphique révèle la distribution des 1915 répondants selon leur éligibilité au don de sang, montrant que la grande majorité (1602 personnes, soit 83,7%) est éligible au don, tandis qu'une proportion modérée (221 personnes, soit 11,5%) fait face à une inéligibilité temporaire qui pourra être levée ultérieurement, et qu'une minorité (92 personnes, soit 4,8%) est confrontée à une inéligibilité définitive qui les empêchera permanemment de donner leur sang, ce qui indique globalement un fort potentiel de donneurs dans cette échantillon étudié.

Graphique 3: Distribution des 1915 répondants selon leur éligibilité au don de sang

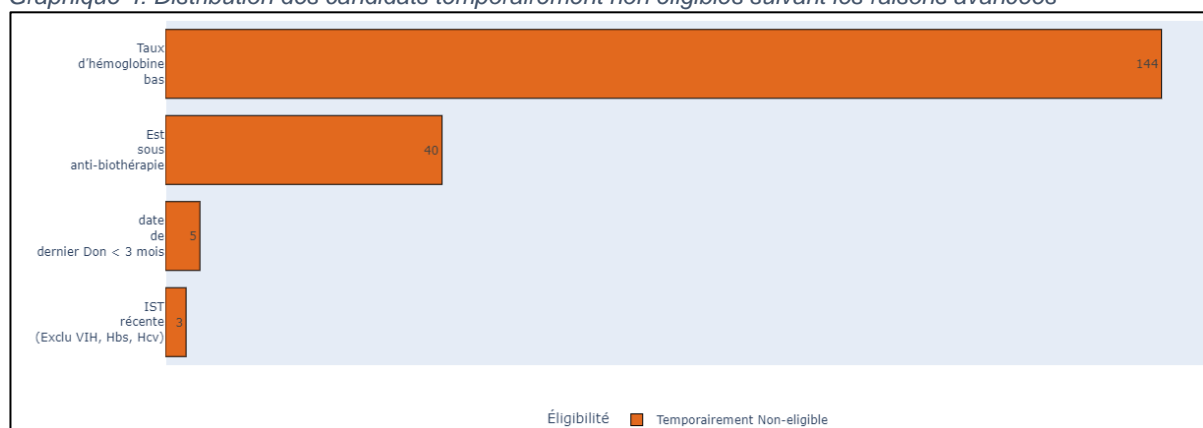


Source : Auteurs, à partir de python

2.2. Distribution des candidats temporairement non éligibles suivant les raisons avancées

Ce Graphique présente les raisons pour lesquelles des personnes sont temporairement non éligibles au don de sang au Cameroun, ainsi que le nombre de cas pour chaque raison. Parmi les motifs d'exclusion temporaire figurent une IST récente (3 cas, hors VIH, hépatite B et C), un dernier don de sang datant de moins de 3 mois (5 cas), un traitement antibiotique en cours (40 cas) et un taux d'hémoglobine bas (144 cas). Les données suggèrent que l'anémie (taux d'hémoglobine bas) est la principale cause d'exclusion temporaire, suivie des traitements antibiotiques, des dons récents et des IST récentes.

Graphique 4: Distribution des candidats temporairement non éligibles suivant les raisons avancées



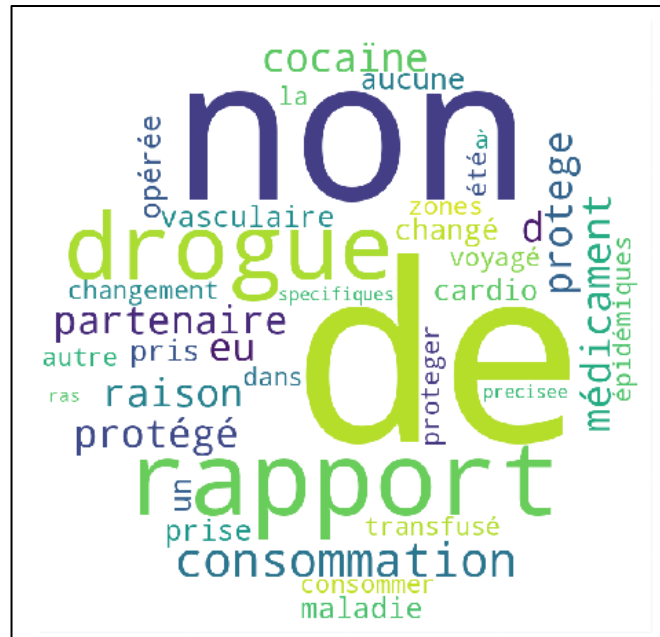
Source : Auteurs, à partir de python

1.3. Autres raisons de non éligibilité temporaire

Ce nuage de mots illustre les autres principales causes d'inéligibilité temporaire au don de sang, mettant en évidence des facteurs tels que la consommation de drogues (notamment la cocaïne), les rapports sexuels non protégés ou avec de nouveaux partenaires, ainsi que diverses considérations médicales comme les opérations récentes, les problèmes cardio-vasculaires, la prise de médicaments spécifiques, ou des antécédents de transfusion. Les termes "non", "de", "rapport" et "drogue" apparaissant de façon proéminente soulignent l'importance particulière des comportements à risque dans les critères d'exclusion temporaire, ces restrictions visant à garantir la sécurité tant des donneurs que des receveurs dans le processus de don sanguin.



Graphique 5: Autres raisons de non éligibilité

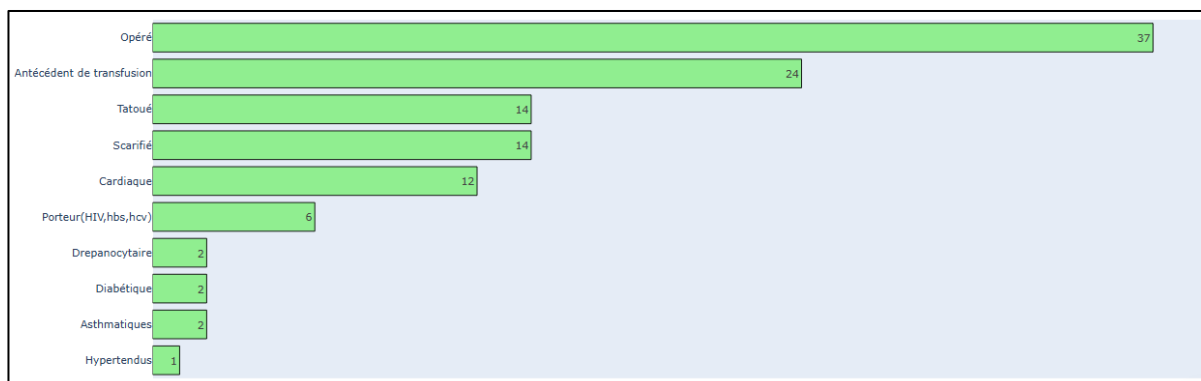


Source : Auteurs, à partir de python

2.3. Répartition des répondants selon les différentes raisons d'inéligibilité définitive au don de sang

Ce graphique présente les différentes raisons d'inéligibilité définitive au don de sang, classées par nombre de cas. On observe que parmi les personnes définitivement non éligibles, la raison la plus fréquente est liée aux antécédents d'opération chirurgicale (37 cas), suivie par les antécédents de transfusion sanguine (24 cas). Les personnes tatouées et scarifiées représentent chacune 14 cas. Les problèmes cardiaques concernent 12 personnes. Les porteurs de virus (HIV, HBs, HCV) comptent 6 cas. Les conditions médicales moins représentées incluent les drépanocytaires, les diabétiques et les asthmatiques (2 cas chacun), ainsi que les hypertendus (1 cas). Toutes ces conditions entraînent une exclusion permanente du don de sang selon les critères établis pour cette population.

Graphique 6: Répartition des répondants selon les raisons de non éligibilité permanente



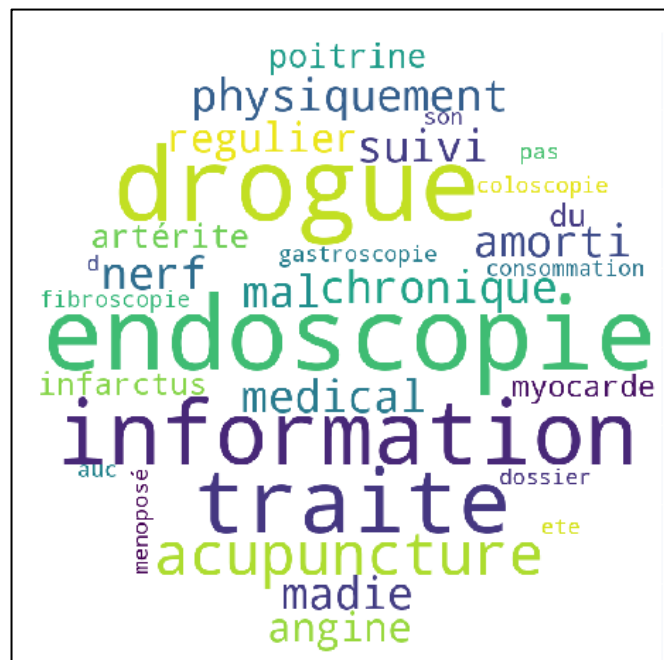
Source : Auteurs, à partir de python

1.4. Autres raisons de non éligibilité permanente

Ce nuage de mots illustre les raisons additionnelles d'inéligibilité permanente au don de sang, où les termes "information", "traité", "endoscopie" et "drogue" dominent, révélant l'importance des antécédents médicaux dans la disqualification des donneurs potentiels; on y trouve diverses catégories de causes incluant les procédures invasives (endoscopie, fibroscopie, coloscopie, gastroscopie), les conditions cardiaques (infarctus, myocarde, angine, artérite), les traitements médicaux spécifiques (acupuncture), les problèmes neurologiques (nerf, malchronique) et l'usage de substances (drogue, consommation), indiquant que ces facteurs médicaux constituent des critères rigoureux d'exclusion définitive pour préserver la sécurité du système transfusionnel.



Graphique 7: raisons additionnelles d'inéligibilité permanente au don de sang



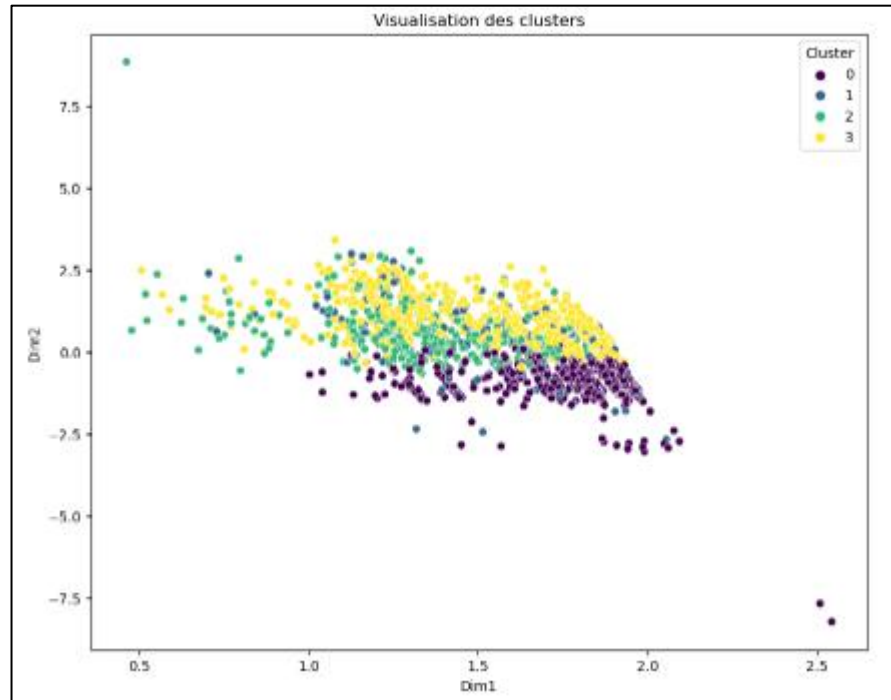
Source : Auteurs, à partir de python

3. Profilage des Donneurs Idéaux

Pour établir le profil des donneurs de sang idéaux, nous avons mis en œuvre une approche par clustering basée sur l'algorithme K-means. Notre analyse a intégré à la fois des variables numériques (âge, profession et quartier de résidence) et catégorielles (genre, situation matrimoniale, religion, niveau d'études et expérience antérieure de don). Un prétraitement rigoureux a été appliqué via un pipeline combinant une standardisation des variables numériques et un encodage one-hot des variables catégorielles. Pour déterminer le nombre optimal de segments, nous avons employé conjointement la méthode du coude et le score de silhouette, analysant les configurations de 2 à 8 clusters. Cette double validation nous a permis d'identifier 4 clusters comme configuration optimale. La visualisation des résultats a été facilitée par une réduction dimensionnelle utilisant TruncatedSVD, permettant de projeter les données multidimensionnelles sur un plan bidimensionnel. Cette segmentation nous a permis d'identifier distinctement les caractéristiques

dominantes des donneurs les plus susceptibles de contribuer régulièrement aux campagnes de don de sang.

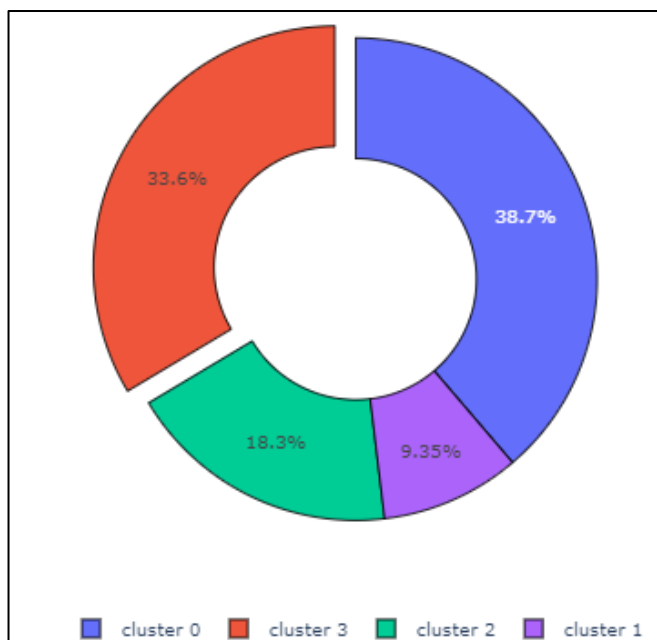
Graphique 8: Nuage des individus suivant les clusters créés



Source : Auteurs, à partir de python

D'après les statistiques descriptives, notre segmentation par clustering a généré quatre profils distincts de donneurs potentiels, avec une répartition inégale entre les segments. Le cluster 0 apparaît comme le groupe majoritaire (avec 38,7% de l'échantillon), suivi de près par le cluster 3 qui comprend 33,6% de personnes. Les clusters 2 et 1 sont moins représentés avec respectivement 18,3% et 9,35%.

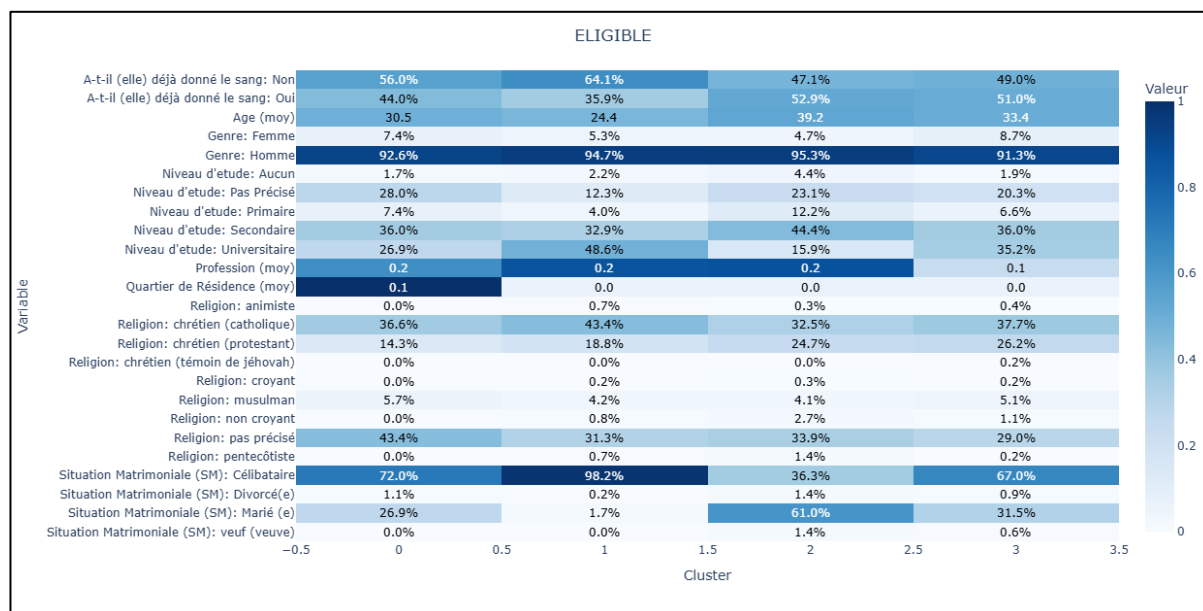
Graphique 9: Répartition des répondants suivants les clusters



Source : Auteurs, à partir de python

A présent, nous passons à la caractérisation des différents clusters.

Graphique 10: Caractérisation des clusters



Source : Auteurs, à partir de python



La visualisation ci-dessus permet de caractériser en détail les quatre clusters de donneurs potentiels selon diverses variables démographiques et comportementales. En analysant systématiquement le graphique :

Cluster 0 (741 individus - 38,7%) :

- Forte présence d'individus n'ayant jamais donné de sang (56,0%)
- Surreprésentation de personnes de niveau d'études secondaire (36,0%)
- Prévalence élevée des religions chrétiennes (catholiques 36,6% et protestantes 14,3%)
- Majoritairement des hommes (92,6%), célibataires (26,5%)
- Caractéristiques moyennes pour les autres variables

Cluster 1 (179 individus - 09,35%) :

- taux relativement faible d'individus ayant déjà donné du sang (35,9%)
- Plus forte proportion de niveau d'études supérieur (48,6%)
- Taux élevé de personnes célibataires (98,2%)
- Présence notable de catholiques (43,4%) mais plus diversifié religieusement

Cluster 2 (350 individus - 18,3%) :

- Taux relativement élevé de personnes ayant déjà donné le sang (52,9%)
- Forte concentration de niveau d'études secondaire (44,4%)
- Prévalence élevée des religions chrétiennes (catholiques 32,5% et protestantes 24,7%)
- Proportion significative de personnes mariées (61,0%) et célibataires (36,3%)
- Majoritairement des hommes

Cluster 3 (645 individus - 33,6%) :

- Taux élevé de personnes ayant déjà donné (51,0%)
- Niveau d'études équilibré entre secondaire (36,0%) et supérieur (35,2%)
- Forte proportion de catholiques (37,7%)
- Taux important de mariés (31,5%) et de célibataires (67%)



Recommandations Stratégiques

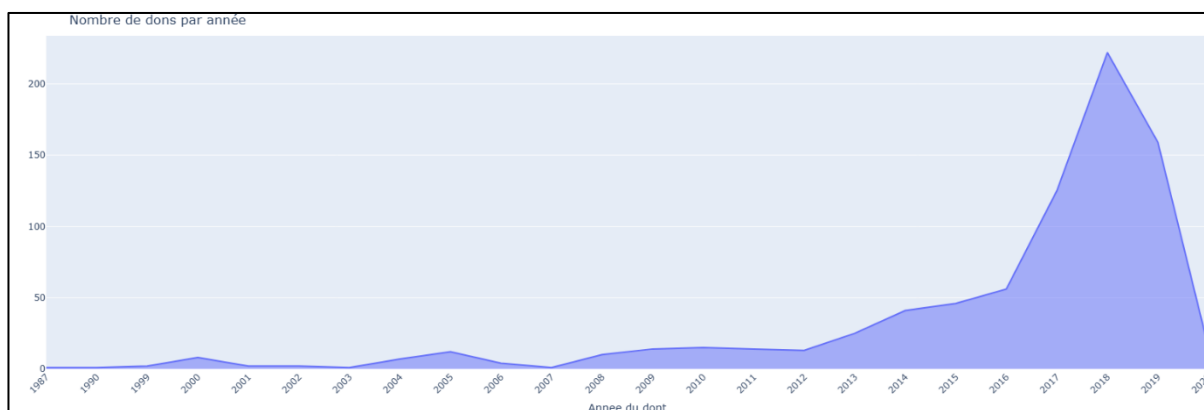
- **Fidéliser** les Clusters 2 & 3 via des programmes de reconnaissance (cartes de donneur, rappels SMS).
- **Convertir** le Cluster 1 par des partenariats avec écoles/entreprises.
- **Mobiliser** le Cluster 0 via des campagnes de proximité (leader religieux, affichage local).

4. Analyse de l'Efficacité des Campagnes

4.1. Evolution du nombre de dons de sang par année

Ce graphique de l'évolution du nombre de dons de sang par année révèle une tendance relativement stable et modeste entre 1997 et 2012, avec de légers pics occasionnels mais sans variation majeure. À partir de 2013, on observe le début d'une croissance progressive qui s'accélère nettement à partir de 2016, culminant en 2018 avec plus de 200 dons annuels, soit une augmentation spectaculaire par rapport aux années précédentes. L'année 2019 montre une légère diminution par rapport au pic de 2018, tout en maintenant un nombre de dons significativement plus élevé qu'avant 2016, tandis que 2020 marque une chute brutale, probablement attribuable à la pandémie de COVID-19 qui a perturbé les systèmes de santé et les collectes de sang dans le monde entier, ramenant le nombre de dons à un niveau comparable à celui observé avant 2013.

Graphique 11: l'évolution du nombre de dons de sang par année

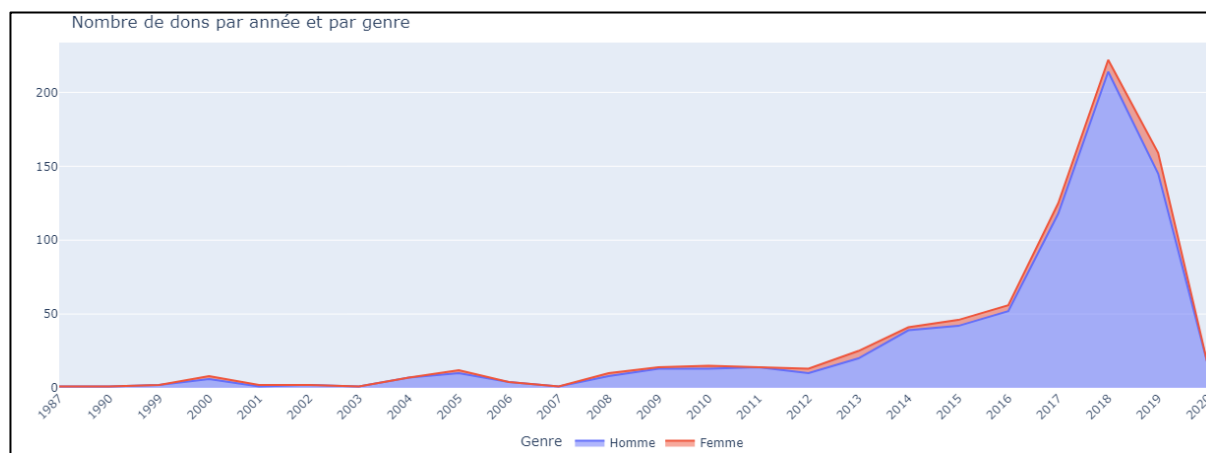


Source : Auteurs, à partir de python

4.2. Evolution du nombre de dons de sang par année et par genre

Ce graphique présente l'évolution du nombre de dons de sang par année et par genre entre 1987 et 2020. On observe une tendance similaire aux graphiques précédents : des dons relativement stables et faibles jusqu'en 2013, une légère augmentation jusqu'en 2016, puis une hausse spectaculaire culminant en 2018 (environ 210 dons) avant une chute brutale en 2019-2020. La zone bleue dominante représente les dons des hommes, tandis que la fine bande rouge au-dessus montre la contribution marginale des femmes. Cette disproportion marquée entre les genres confirme les observations du diagramme précédent, indiquant une participation au don de sang très majoritairement masculine au Cameroun sur toute la période étudiée.

Graphique 12: l'évolution du nombre de dons de sang par année et par genre



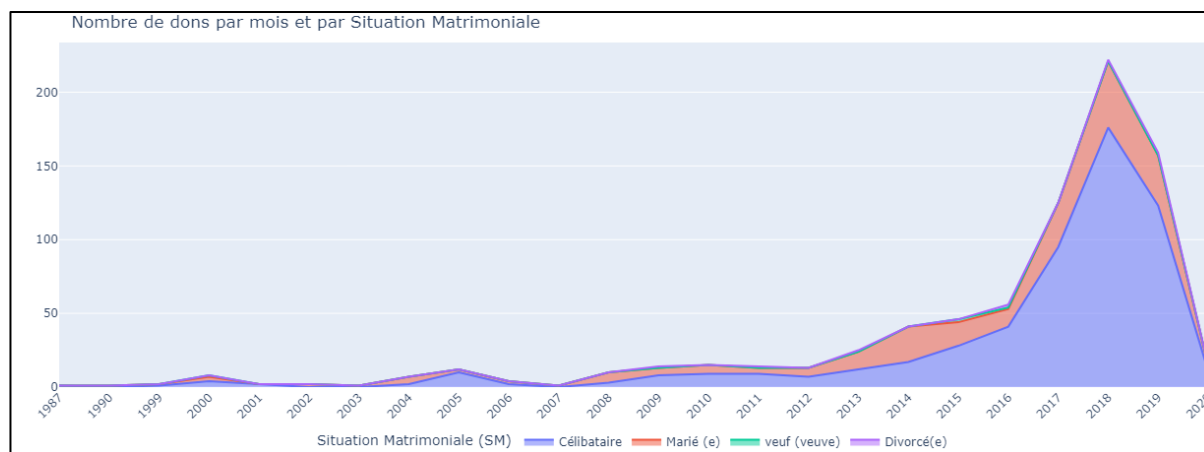
Source : Auteurs, à partir de python

4.3. Nombre de dons par mois et par Situation Matrimoniale

L'analyse du graphique "Nombre de dons par mois et par Situation Matrimoniale" révèle une évolution frappante des dons de sang entre 1998 et 2020, caractérisée par une période de relative stabilité à des niveaux bas jusqu'en 2015, suivie d'une légère augmentation progressive puis d'une hausse spectaculaire entre 2017 et 2018, culminant à environ 200 dons mensuels en 2018 avant de connaître une chute drastique en 2019-2020. Les donateurs célibataires (représentés par la courbe bleue) constituent nettement la majorité des contributeurs, particulièrement lors du pic de 2018, suivis des donateurs mariés (zone rouge), tandis que les donateurs veufs et divorcés (courbes verte et violette respectivement) représentent une proportion

infime des dons, suggérant soit des campagnes de sensibilisation particulièrement efficaces ciblant les célibataires en 2017-2018, soit des changements importants dans les politiques de collecte de sang ayant influencé cette dynamique remarquable.

Graphique 13: Nombre de dons par mois et par Situation Matrimoniale

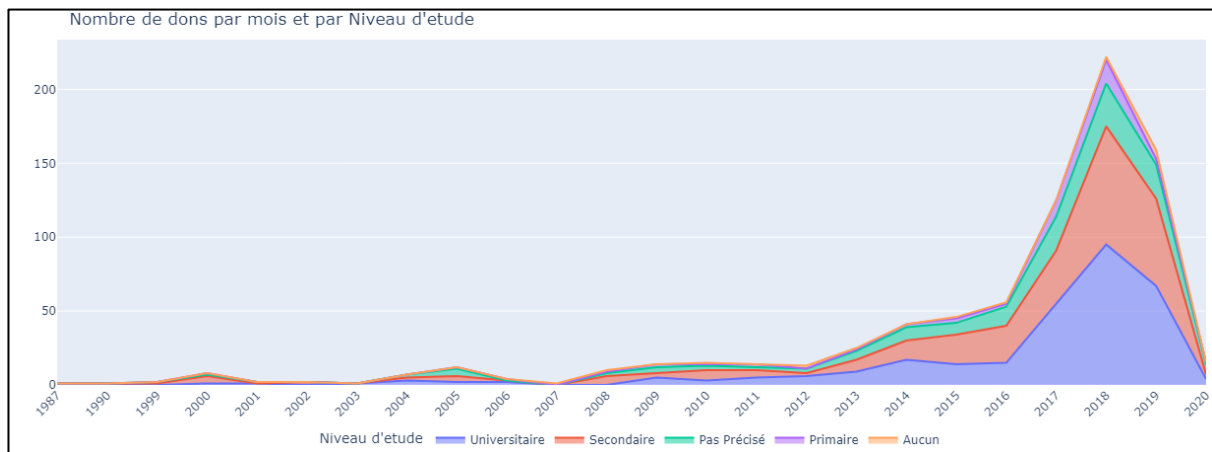


Source : Auteurs, à partir de python

4.4. Evolution mensuelle des dons de sang par niveau d'étude entre 1998 et 2020

Ce graphique montre l'évolution mensuelle des dons de sang par niveau d'étude entre 1987 et 2020. On observe une tendance similaire au graphique précédent: des dons relativement stables et faibles jusqu'en 2013, suivis d'une légère augmentation jusqu'en 2016, puis une hausse spectaculaire culminant en 2018 (environ 220 dons mensuels) avant une chute brutale en 2019-2020. Les donateurs universitaires (zone bleue) constituent la plus grande proportion, suivis des donateurs de niveau secondaire (zone rouge) et ceux dont le niveau n'est pas précisé (zone verte). Les donateurs de niveau primaire ou sans éducation formelle sont quasi inexistantes. Cette distribution suggère que les campagnes de don de sang ont été particulièrement efficaces auprès des populations éduquées, notamment universitaires, pendant la période de pic.

Graphique 14: l'évolution mensuelle des dons de sang par niveau d'étude entre 1987 et 2020

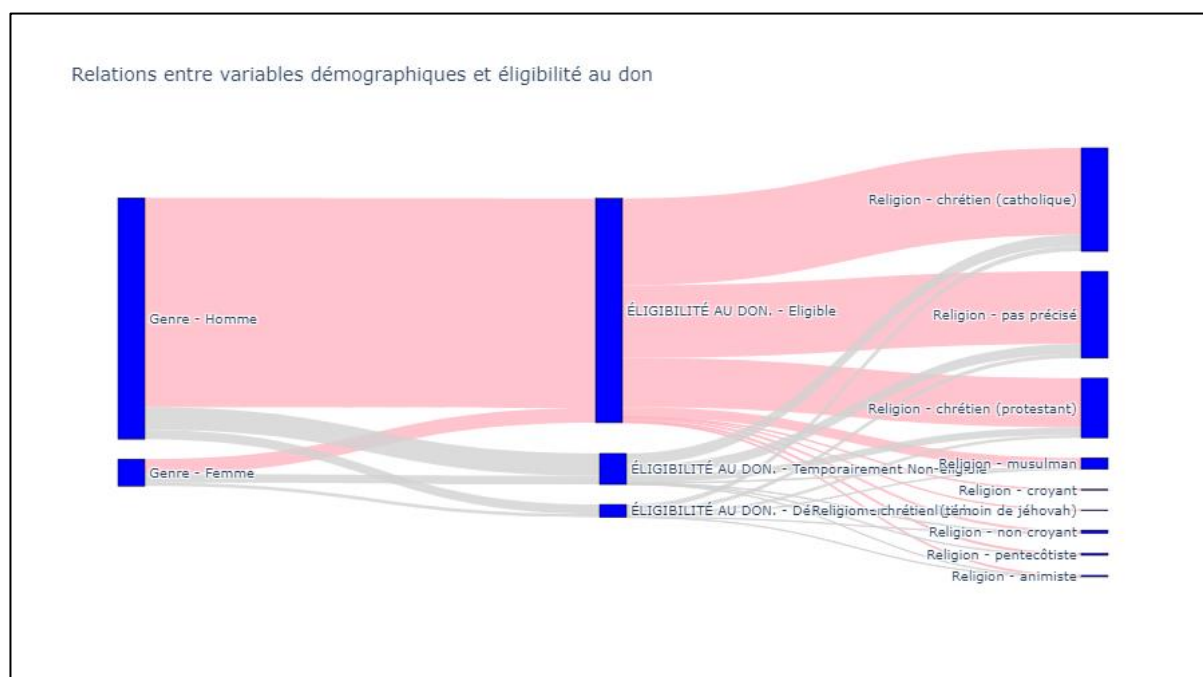


Source : Auteurs, à partir de python

4.5. Relation entre les variables démographiques et l'éligibilité au don de sang

Le diagramme de flux illustre la relation entre les variables démographiques et l'éligibilité au don de sang, révélant une nette prédominance des hommes parmi les donateurs éligibles, avec une forte représentation des chrétiens catholiques comme groupe religieux majoritaire, suivis par les personnes dont la religion n'est pas précisée et les protestants. À l'inverse, les femmes sont surreprésentées dans les catégories "Temporairement Non-Éligible" et "Déféré", suggérant des obstacles spécifiques à leur participation, tandis que certaines affiliations religieuses (musulmans, témoins de Jéhovah, non-croyants, pentecôtistes et animistes) apparaissent en proportions marginales parmi les donateurs, ce qui souligne l'influence déterminante du genre et de l'appartenance religieuse sur les patterns d'éligibilité et de participation au don de sang.

Graphique 15: relation entre les variables démographiques et l'éligibilité au don de sang



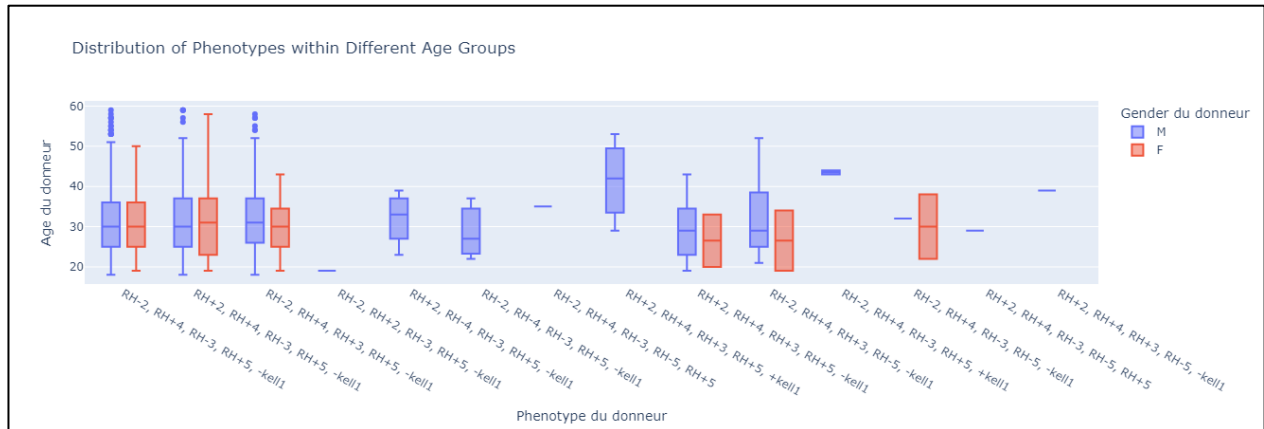
Source : Auteurs, à partir de python

5. Analyse des caractéristiques des donneurs effectifs de sang

5.1. Distribution des phénotypes sanguins selon l'âge et le genre

La campagne de don de sang au Cameroun en 2019 révèle une distribution variée des phénotypes sanguins selon l'âge et le genre. Les donneurs ont majoritairement entre 25 et 40 ans, avec une représentation relativement équilibrée entre hommes (en bleu) et femmes (en rose). Les phénotypes Rh+/- sont les plus fréquents, tandis que certains phénotypes plus rares montrent une distribution d'âge plus restreinte ou une prédominance de genre. Quelques valeurs aberrantes sont observées dans les groupes d'âge supérieurs pour certains phénotypes, suggérant une participation occasionnelle de donneurs plus âgés. Cette visualisation offre des informations précieuses pour orienter les futures campagnes de sensibilisation au don de sang vers les groupes démographiques moins représentés.

Graphique 16: Distribution des phénotypes sanguins selon l'âge et le genre



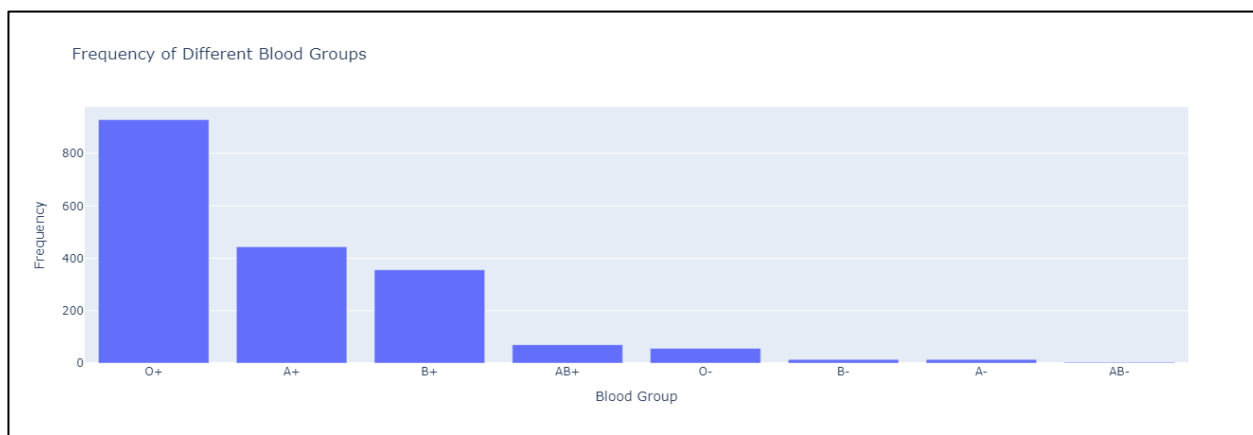
Source : Auteurs, à partir de python

5.2. Distribution des groupes sanguins des dons de sang au Cameroun en 2019

Le graphique illustre la distribution des groupes sanguins des dons de sang au Cameroun en 2019, montrant que le groupe O+ est le plus fréquent avec environ 900 dons, suivi par A+ (environ 450 dons) et B+ (environ 400 dons), tandis que AB+ et O- ont des fréquences bien plus faibles, autour de 100 dons chacun, et les groupes B-, A- et AB- sont les moins représentés, avec des fréquences proches de 0. Cette répartition reflète une prédominance du groupe O+, ce qui est cohérent avec les tendances génétiques en Afrique, où ce groupe est souvent majoritaire.



Graphique 17: distribution des groupes sanguins des dons de sang au Cameroun en 2019



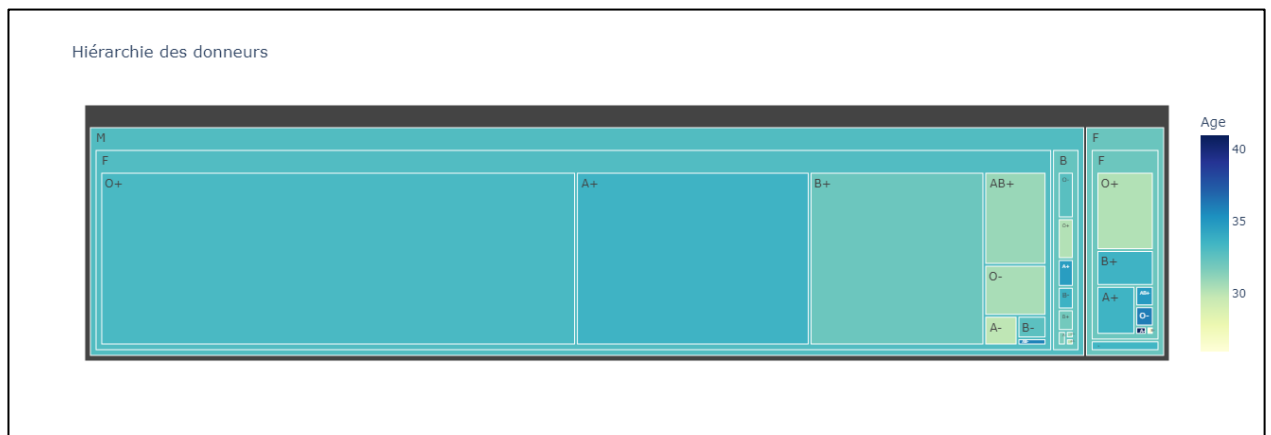
Source : Auteurs, à partir de python

5.3. Hiérarchie des donneurs en 2019

La treemap "Hiérarchie des donneurs" illustre la distribution des donneurs de sang en 2019 selon leur genre et groupe sanguin, où la taille des rectangles représente la proportion de chaque catégorie et la couleur indique l'âge moyen (du bleu foncé pour les plus âgés au jaune clair pour les plus jeunes). On observe une prédominance de donneurs familiaux (F) masculins (M), avec le groupe O+ représentant la plus grande proportion de donneurs, suivi par les groupes A+ et B+, tandis que les groupes Rhésus négatifs (A-, B-, AB-, O-) sont nettement moins représentés que leurs équivalents positifs, révélant ainsi des opportunités de ciblage pour les futures campagnes de don de sang, particulièrement vers les groupes sous-représentés. Cette même tendance est également observée chez les femmes tant familiales et bénévoles.



Graphique 18: Hiérarchie des donneurs



Source : Auteurs, à partir de python



Chapitre 3

PREDICTION D'ELIGIBILITE ET AI ASSISTANT



Dans le cadre de ce projet, nous avons développé un modèle de prédiction basé sur l'apprentissage automatique afin d'évaluer l'éligibilité des nouveaux donneurs de sang à partir de données démographiques et de santé. Ce modèle, conçu pour répondre à un besoin concret d'automatisation et de précision, vise à faciliter le processus de sélection des donneurs tout en garantissant la sécurité et la fiabilité des résultats. En intégrant ce modèle dans une API, nous offrons une solution pratique et adaptable, permettant une prédiction en temps réel directement depuis un tableau de bord. Ce chapitre présente les étapes clés de la conception du modèle, les choix techniques effectués, ainsi que ses performances et son potentiel d'application.

1. Résumé de l'approche adoptée pour la prédiction

Pour prédire l'éligibilité au don de sang, un modèle a été développé en Python à partir des données filtrées (`df_filtre`). Les variables explicatives, incluant des caractéristiques telles que la taille, le poids, le taux d'hémoglobine, l'âge, le genre ou encore la profession, ont été séparées de la variable cible (éligibilité au don), puis les données ont été divisées en ensembles d'entraînement et de test (80/20) avec `train_test_split`. Un prétraitement a été appliqué via un pipeline : les valeurs manquantes des variables numériques ont été imputées par la médiane et normalisées avec `StandardScaler`, tandis que celles des variables catégorielles ont été imputées par le mode et encodées avec `TargetEncoder`. Le modèle `RandomForestClassifier`, intégré dans le pipeline, a été entraîné sur les données d'entraînement, puis évalué sur les données de test, fournissant une précision via `accuracy_score` et un rapport détaillé des performances avec `classification_report`.

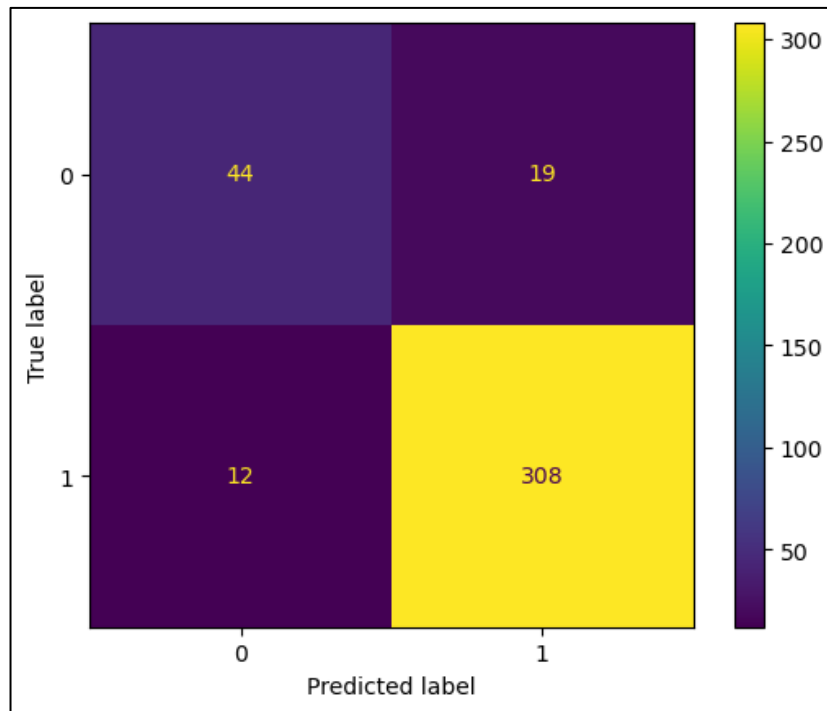
2. Performance du modèle

Cette matrice de confusion montre les performances du modèle de prédiction d'éligibilité au don de sang, pour les deux classes : 0 (non éligible) et 1 (éligible). Sur 383 individus, le modèle a correctement prédit 44 cas comme non éligibles (vrai négatif) et 308 cas comme éligibles (vrai positif), ce qui indique une bonne capacité à identifier les éligibles. Cependant, il a incorrectement prédit 19 cas comme éligibles alors qu'ils ne l'étaient pas (faux positif) et 12 cas comme non éligibles alors qu'ils l'étaient (faux négatif), suggérant une légère tendance à sur-prédire l'éligibilité.





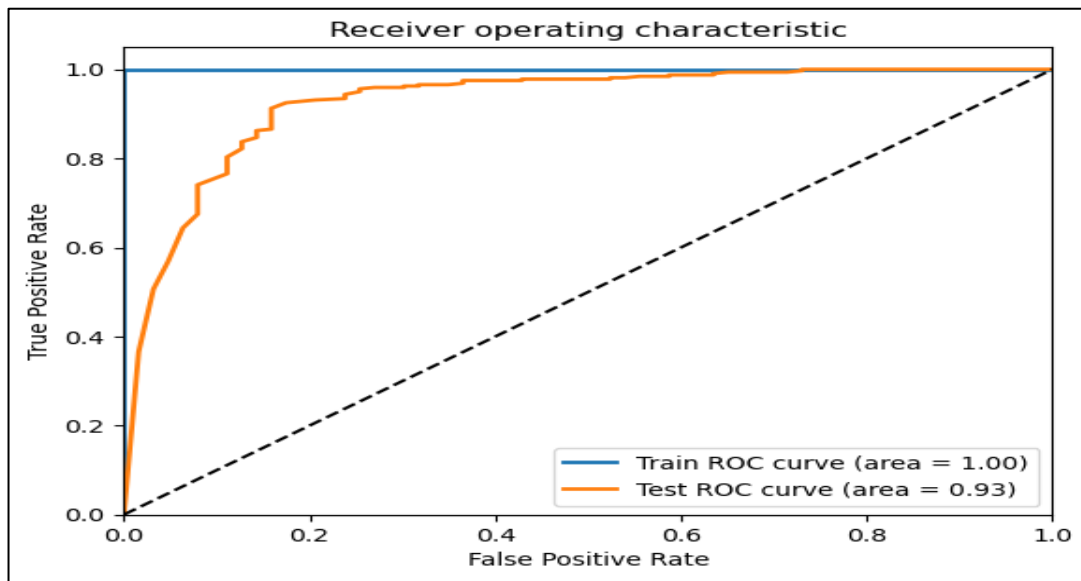
Graphique 19 : Matrice de confusion



Source : Auteurs, à partir de python

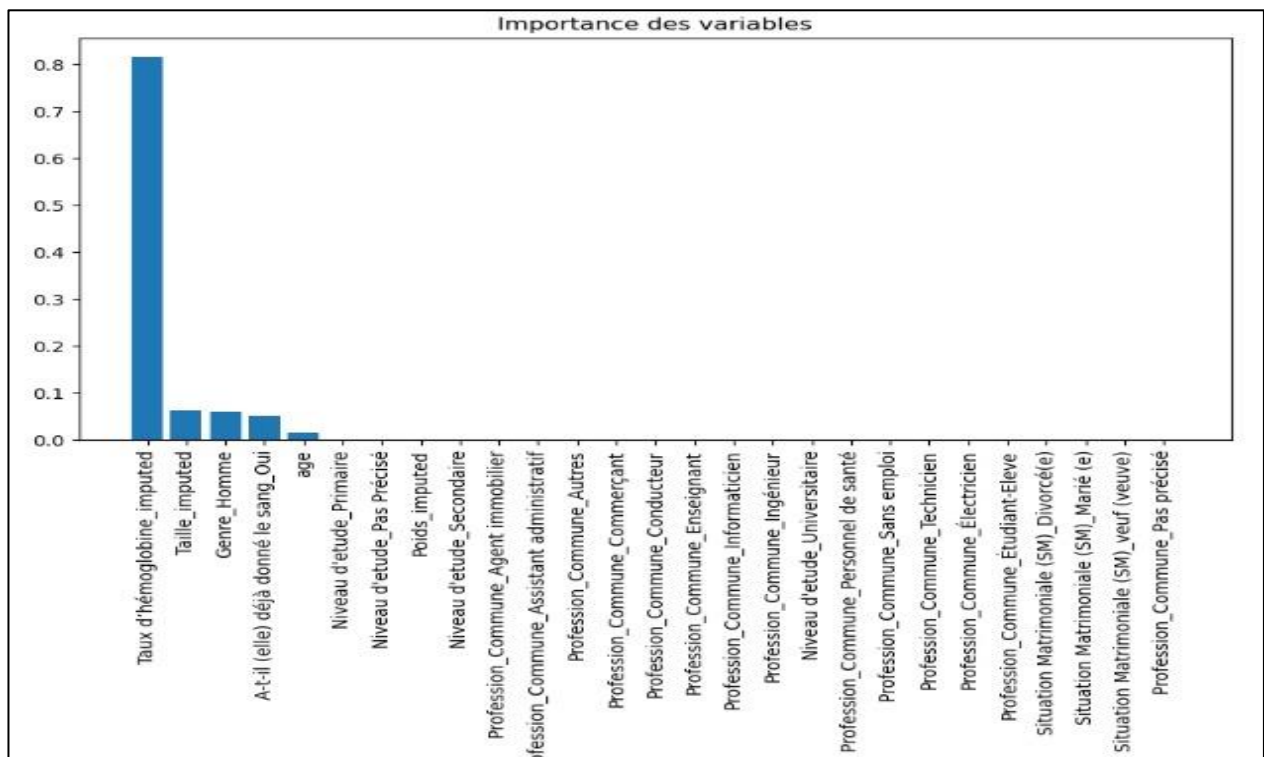
La courbe ROC du modèle de prédiction d'éligibilité au don de sang montre une excellente performance, avec une AUC de 1.0 sur les données d'entraînement (indiquant une classification parfaite) et une AUC de 0.93 sur les données de test, ce qui reste très bon et reflète une forte capacité à discriminer les éligibles des non éligibles, avec un faible taux de faux positifs.





Source : Auteurs, à partir de python

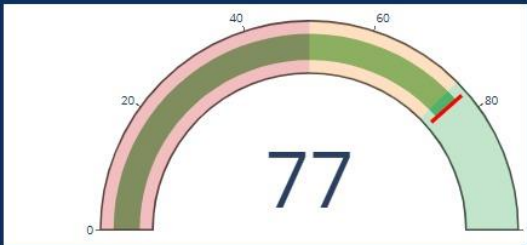
L'analyse de l'importance des variables dans notre modèle de prédiction d'éligibilité au don de sang révèle que la variable taux d'hémoglobine est de loin la plus influente (importance ≈ 0.8), suivie par la taille (≈ 0.1) et le poids (≈ 0.1), soulignant l'importance des critères physiologiques et de l'expérience passée dans la prédiction, "A-t-il (elle) déjà donné le sang ?".





3. Simulation de prédiction

Prédiction d'Éligibilité au Don de Sang

Données Démographiques		Résultats de la Prédiction
Genre	Âge	 <p>Éligible au don de sang</p> <p>Éligible au don de sang. Bon candidat pour le don de sang.</p> <p>Conseils et Recommandations</p> <p>Hydratez-vous bien avant votre don.</p> <p>Prenez un repas équilibré dans les heures précédant votre don.</p>
Femme	30	
Niveau d'études	Situation Matrimoniale	
Universitaire	Divorcé(e)	
Religion		
Pas Précisé		
Profession	Nationalité	
Agent immobilier	Cameroonais	
Taille (cm)	Poids (kg)	
168	89	
Arrondissement de résidence	Quartier de résidence	

Source : Auteurs, à partir de python

Ici, le modèle prédit que l'individu est éligible au don à 77%.

AI ASSISTANT

Nous avons intégré une IA assistant au sein du dashboard afin qu'elle accompagne les utilisateurs dans l'interprétation des différents indicateurs, les aide à comprendre les concepts liés au don de sang et les guide dans la prise de décision en fonction des résultats affichés. L'ajout d'un chatbot dans un dashboard dédié au don de sang au Cameroun présente un intérêt majeur : il offre une assistance personnalisée et immédiate, permettant aux utilisateurs, qu'ils soient professionnels de santé ou organisateurs de campagnes, de mieux comprendre les données, d'identifier les profils de donneurs potentiels, et d'optimiser les stratégies de collecte de sang, tout en tenant compte des spécificités locales, comme les barrières culturelles ou logistiques, pour améliorer l'efficacité et l'impact des initiatives de don de sang dans le pays.

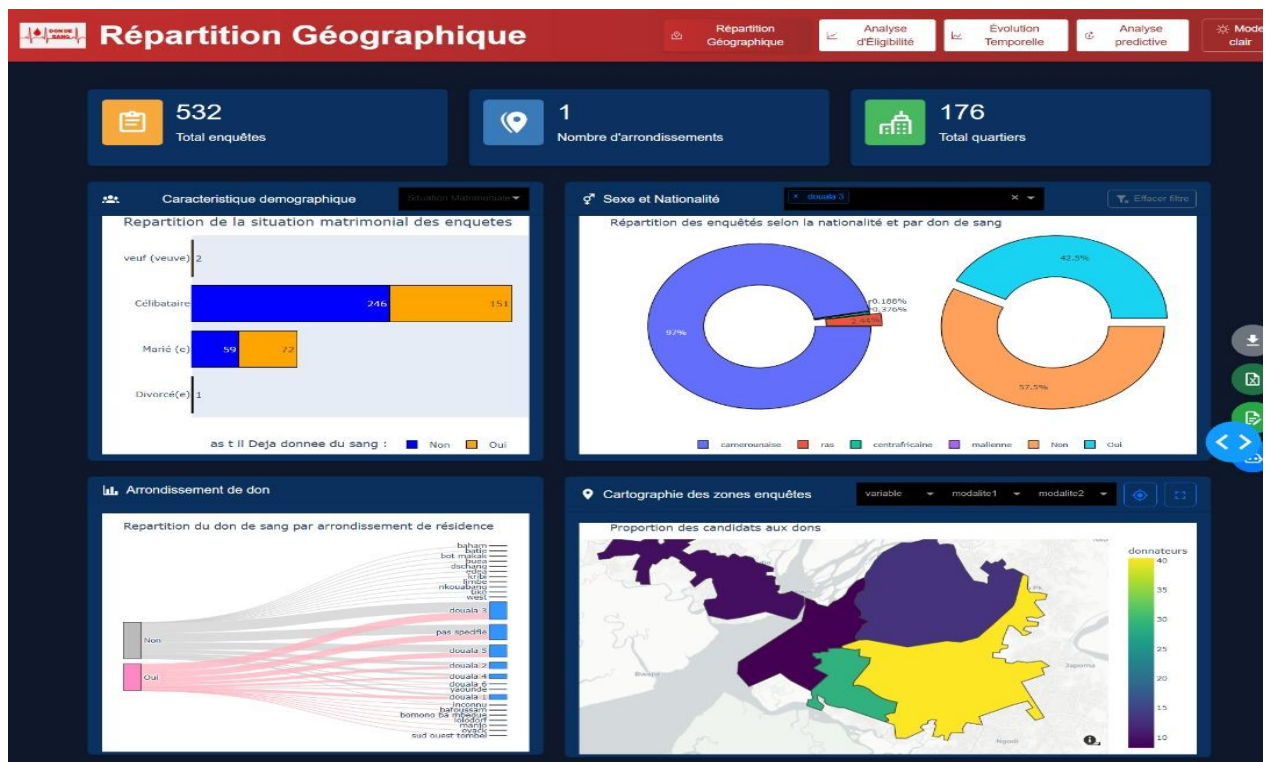


PRESENTATION DU DASHBOARD

Le Dashboard que nous avons mis en place est composé des 4 onglets suivants :

Pour accéder au Dashboard, [cliquer ici](#).

Onglet 1 : Répartition géographique



Onglet 2 : Analyse d'éligibilité



Onglet 3 : Evolution temporelle





Onglet 4 : Analyse prédictive

The screenshot shows a web application titled 'Analyse prédictive' with a red header. Below the header, there are four tabs: 'Répartition Géographique', 'Analyse d'Éligibilité', 'Évolution Temporelle', and 'Analyse prédictive' (which is active). A 'Mode clair' button is also present. The main content area is titled 'Prédiction d'Éligibilité au Don de Sang'. It is divided into two columns. The left column, 'Données Démographiques', contains several input fields: 'Genre' (Homme), 'Âge' (30), 'Niveau d'études' (Secondaire), 'Situation Matrimoniale' (Célibataire), 'Religion' (Chrétien (Protestant)), 'Profession' (Agent immobilier), 'Nationalité' (Etranger), 'Taille (cm)' (149), and 'Poids (kg)' (56). The right column, 'Résultats de la Prédiction', features a semi-circular gauge with a score of 82. The gauge has a green section (0-40), a yellow section (40-60), and a red section (60-80). Below the gauge, it says 'Éligible au don de sang' and 'Éligible au don de sang. Bon candidat pour le don de sang.' At the bottom, there is a section 'Conseils et Recommandations' with the text 'Hydratez-vous bien avant votre don.' On the right side of the dashboard, there are several icons: a download icon, a print icon, and a share icon.

Questionnaire Kobotoolbox

Pour éviter d'avoir des problèmes d'incohérence dans les bases de données et faciliter les collectes de données futures, nous avons prévu un questionnaire électronique crée à l'aide de Kobotoolbox. Le lien vers le questionnaire a été intégrée dans le dashboard. Pour accéder au questionnaire : [cliquez ici](#)

Ce chapitre a présenté un modèle de prédiction d'éligibilité au don de sang, développé en Python avec un RandomForestClassifier sur des données filtrées, montrant une excellente performance (AUC de 1.0 en entraînement, 0.93 en test) et une bonne identification des éligibles (308 vrais positifs sur 383), malgré quelques erreurs (19 faux positifs, 12 faux négatifs). Les variables clés sont taux d'héoglobine, l'expérience passée de don (importance ≈ 0.1), la taille (≈ 0.1) et le poids (≈ 0.1). Une IA assistante intégrée au dashboard (quatre onglets : Répartition géographique,



Analyse d'éligibilité, Évolution temporelle, Analyse prédictive) aide à interpréter les résultats et à optimiser les stratégies de collecte au Cameroun, en tenant compte des spécificités locales.





Chapitre 4

LIMITES ET RECOMMANDATIONS





4. LIMITES DE L'ETUDE

Ce travail présente quelques limites. Tout d'abord, l'échantillon de 1915 individus, bien que significatif, est principalement centré sur Douala, ce qui limite la représentativité nationale, notamment pour les zones rurales où les infrastructures et les comportements diffèrent. Ensuite, le déséquilibre des classes (83,7 % d'éligibles) peut biaiser les prédictions, malgré les efforts d'optimisation du seuil de probabilité, et les barrières culturelles ou logistiques spécifiques au Cameroun (croyances, accès aux centres) n'ont pas été pleinement intégrées dans le modèle.

2. RECOMMANDATIONS

Les résultats de cette étude appellent à une **optimisation ciblée** des campagnes de don de sang au Cameroun. Sur le plan géographique, il est crucial de renforcer les collectes dans les zones sous-représentées (Douala 1, 4 et régions rurales) via des unités mobiles et des partenariats locaux, tout en décentralisant les centres de collecte le long des axes identifiés (Bafoussam, Limbé). Pour améliorer l'éligibilité, des actions spécifiques s'imposent : lutte contre l'anémie (supplémentation en fer, conseils nutritionnels) et sensibilisation aux contre-indications temporaires (guides clairs, application mobile de pré-vérification). L'analyse des clusters révèle que les hommes célibataires éduqués (Clusters 2 & 3) constituent le cœur des donneurs réguliers : leur fidélisation passe par des programmes de reconnaissance (badges, rappels personnalisés) et des collectes adaptées à leurs disponibilités. Les groupes à potentiel sous-exploité (Cluster 1 : universitaires) nécessitent des partenariats avec écoles et entreprises, tandis que les non-donneurs (Cluster 0) doivent être mobilisés via des campagnes communautaires, notamment dans les églises. Par ailleurs, une gestion proactive des stocks est essentielle, avec un ciblage des groupes sanguins rares (AB-, B-) et une anticipation des pénuries saisonnières. Enfin, la digitalisation (plateforme de suivi des donneurs, analyse prédictive) et des protocoles adaptés aux crises (ex : collectes décentralisées post-COVID) garantiront la résilience du système. Ces mesures, combinant analyse data-driven et ancrage local, permettront de répondre efficacement aux disparités identifiées et d'assurer un approvisionnement durable en sang.





Ce rapport a exploré de manière approfondie les dynamiques du don de sang au Cameroun, révélant à la fois des défis persistants et des opportunités stratégiques pour améliorer l'approvisionnement en produits sanguins. L'analyse des données de 1 915 donneurs potentiels a mis en lumière plusieurs constats clés : une **forte concentration géographique** des dons à Douala (notamment dans les arrondissements 3 et 5), une **éligibilité globale élevée (83,7%)** mais freinée par des facteurs comme l'anémie (cause principale d'exclusion temporaire), et des **profils de donneurs distincts** identifiés par clustering (avec les Clusters 2 et 3 émergeant comme cibles prioritaires). Les disparités genre (92,6% d'hommes parmi les donneurs réguliers) et socio-culturelles (influence des religions chrétiennes) soulignent la nécessité d'une approche segmentée.

Le modèle prédictif développé (Random Forest, AUC = 0.93) a confirmé l'importance critique du **taux d'hémoglobine**, de la taille et du poids dans l'éligibilité, offrant un outil fiable pour optimiser le recrutement. Cependant, les limites de l'étude (échantillon urbain centré sur Douala, déséquilibre de classes) appellent à une prudence dans la généralisation des résultats.

Les **recommandations stratégiques** proposent une feuille de route concrète :

- **Décentraliser les collectes** via des unités mobiles et des partenariats locaux dans les zones rurales et urbaines sous-représentées.
- **Cibler les donneurs réguliers** (hommes éduqués, célibataires ou mariés) avec des programmes de fidélisation (rappels SMS, badges).
- **Convertir les populations à potentiel** (universitaires, femmes) via des campagnes adaptées (sensibilisation en milieu académique, conseils nutritionnels contre l'anémie).
- **Digitaliser le processus** (plateforme de suivi, chatbot d'assistance) pour renforcer l'efficacité opérationnelle.

En synthèse, cette étude démontre que l'**alliance entre data science et contextualisation socio-culturelle** peut transformer la gestion des dons de sang au Cameroun. Les insights tirés des données, combinés à des interventions ciblées sur les freins identifiés (géographiques, médicaux, comportementaux), permettront de construire un système transfusionnel plus **équitable, résilient et durable**, sauver des vies et répondre aux besoins croissants du pays. La mise en œuvre de ces





recommandations, en partenariat avec les acteurs locaux (CNTS, églises, universités), constitue une étape décisive vers l'autosuffisance en sang.

Perspective : Une extension future de cette recherche pourrait intégrer des données nationales plus larges et des enquêtes qualitatives pour affiner l'analyse des barrières culturelles, tout en renforçant l'IA prédictive avec des variables contextuelles supplémentaires (accès aux transports, saisonnalité des dons).





BIBLIOGRAPHIE

1. **OMS (2021).** *Guidelines on Blood Donor Selection and Blood Collection*. Organisation Mondiale de la Santé.
 - *Référence mondiale sur les critères d'éligibilité et bonnes pratiques de collecte.*
2. **Pindyck, R. S., & Rubinfeld, D. L. (2018).** *Économétrie* (5^e éd.). Pearson.
 - *Méthodes statistiques appliquées aux tests d'hypothèses (ANOVA, Chi²).*
3. **James, G. et al. (2021).** *An Introduction to Statistical Learning* (2^e éd.). Springer.
 - *Fondements des modèles prédictifs (Random Forest, clustering K-means).*
4. **CNTS Cameroun (2020).** *Rapport Annuel sur la Transfusion Sanguine*.
 - *Données locales sur les besoins et pénuries de sang.*



5. **Nguyen, T. et al. (2019).** *"Machine Learning for Blood Donor Eligibility Prediction"*, Journal of Medical Systems, 43(7).
 - *Application du machine learning au tri des donneurs.*

