NK STAT
CONSULTING
2025

Analysis and Visualization of Blood Donation Campaigns in Cameroon

Report Outline

INTRODUCTION

1. CONTEXT

In Cameroon, blood donation remains a critical public health issue. With a constantly evolving healthcare system and increasing transfusion needs, particularly due to obstetric emergencies, road accidents, diseases such as malaria and sickle cell disease, as well as surgical procedures, the country faces a constant challenge in supplying blood products. According to estimates from Cameroon's National Blood Transfusion Center (CNTS), the country only meets approximately 50% of its annual blood needs, leading to dramatic consequences for many patients.

CNTS
CENTRE NATIONAL DE TRANSFUSION SANGUINE
NBTS
NATIONAL BLOOD TRANSFUSION SERVICE

Blood donation campaigns in Cameroon face several obstacles, including unfavorable cultural beliefs, lack of awareness of procedures, and an uneven geographical distribution of blood donation centers between urban and rural areas. Furthermore, the low rate of regular donors (less than 20% according to the latest national statistics) further complicates the situation.

This project is part of a desire to optimize future blood donation campaigns in Cameroon by using data collected during previous initiatives, in order to improve the efficiency of collections and guarantee a more stable supply of blood products on a national scale.

2. GOALS

The dashboard developed as part of this challenge aims to optimize blood donation campaigns in Cameroon to increase the supply of blood products.

The three specific objectives logically group together the different analyses proposed:

- Geographic and demographic analysis (who and where)
- Analysis of eligibility and retention factors (why)

- Strategic recommendations based on temporal and qualitative analyses (how)

3. STRUCTURE OF THE REPORT

This report is structured around four main chapters. First, a presentation of the data and methodology is provided. Then, the main results are presented, along with the blood donation eligibility prediction model using the AI assistant. Finally, the conclusions and recommendations are presented.

Chapter 1

PRESENTATION OF DATA AND ANALYSIS METHODOLOGY

1. *Overview*

The database contains detailed information on potential blood donors in Cameroon, with a total of **1915 records** (individuals) and **39 columns** (variables). This collection represents a significant sample for analyzing donor characteristics and factors influencing blood donation eligibility in the Cameroonian context.

2. *Data structure and content*

Our dataset on blood donation in Cameroon consists of 1915 individual records described by 39 different variables. The vast majority of the data are categorical in nature (36 columns of type ' object '), supplemented by two numeric variables (of type 'float64') corresponding to physical measurements, and a temporal variable (in the format 'datetime64[ns]') for dates. This rich dataset covers several dimensions essential for understanding donor profiles and blood donation eligibility factors in the Cameroonian context.

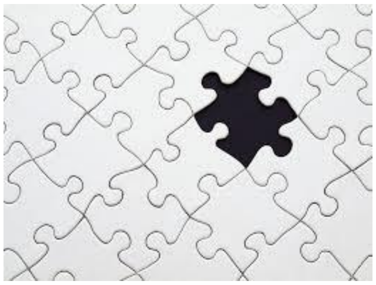
Demographic information constitutes a significant part of the variables collected, including personal data (date of birth, gender, height, weight, and marital status) that allow for the establishment of a basic profile of potential donors. Socioeconomic status is represented by educational level and occupation, while the geographic dimension is captured by the district and neighborhood of residence, providing valuable spatial granularity for territorial analysis. Cultural characteristics, including nationality and religion, complete this demographic portrait by providing an essential sociocultural dimension in the Cameroonian context.

Medical and eligibility information forms the analytical core of the database, with variables documenting donation history (previous experience and date of last donation), critical medical parameters such as hemoglobin level, and most importantly, the target donation eligibility variable that determines the candidate's ability to donate blood. This central variable is complemented by a detailed system for documenting reasons for ineligibility, structured into three main categories: temporary unavailability (such as ongoing antibiotic therapy, insufficient hemoglobin level, too recent a donation or a recent sexually transmitted infection), unavailability specific to women (related to the menstrual cycle, breastfeeding, recent childbirth, termination of pregnancy or ongoing pregnancy), and causes of permanent ineligibility (history of transfusion, positive serological status for HIV, hepatitis B or C, surgical interventions, sickle cell disease, diabetes, hypertension, asthma, heart problems, presence of tattoos or scarifications).

3. Methodology

This section presents the methodology adopted for data processing and analysis in this study. The main objective is to detail the different steps of preprocessing, cleaning, and analysis of key variables. The study relies on the use of Python and libraries such as Pandas, fuzzywuzzy , scikit-learn , and geopy to ensure rigorous data processing.

3.1. Value management missing



Missing values were identified and treated according to their nature.

- For numeric variables, imputation was performed using scikit-learn 's IterativeImputer algorithm , predicting missing values from other numeric variables in the dataset.

- For categorical variables, imputation was done by replacing missing values with the most frequent modality (mode).

3.2. Processing key variables

3.2.1. Processing of the occupation variable

The main problem encountered with this variable was the multiplication of writing forms for the same profession. To homogenize the data, the following steps were followed:

- Creation of a standardized variable ' **Profession_Commune** ' ;
- Use of a dictionary that associates the different variations of a profession with a single common category;
- Transformation of the variable into indicator variables (**dummy variables**) for the needs of quantitative analysis;

3.2.2. Processing of Nationality and Region variables

A function **(transform_nation)** is defined to clean and standardize the values of the 'Nationality' column (removing spaces and periods, converting to lowercase, specific replacements). An algorithm is thus built for this task and for the Religion variable.

- The values of the variables are cleaned and standardized:
    - Removed spaces and special characters.
    - Convert to lowercase.
    - Using replacement dictionaries to group similar terms.

This in fact reduces the number of modalities of this variable in order to allow good structuring of the analysis and modeling.

3.2.3. Processing of Temporal Data

Our approach to processing dates is based on several methodological principles:

- **Inconsistency Detection: We have developed a systematic method to identify invalid dates using** datetime conversion and identifying NaT (Not a Time) values.
- **Year Correction:** A specific transformation function has been designed to correct incorrect years. For example, for the *'Date of completion of the form'* , invalid years are systematically replaced by '2019'.
- **Normalization:** In case of completely invalid dates, we use the most frequent date (mode) among the valid dates as a replacement value.

For dates of birth, we have developed a correction strategy based on:

- Analysis of the last two figures of the year
- The attribution of a century (2000 or 1900) based on these figures
- Handling conversion errors with the errors = 'coerce' option

3.2.4 Residential area and district

First, the column is converted to lowercase and unwanted characters are removed. The fuzzywuzzy library is used to group similar neighborhood names based on a 90% similarity threshold because there may be neighborhood names that are very similar but actually designate two completely different areas. Then a dictionary (grouped_quartiers) is generated to represent these groups. A function (get_grouped_quartier) is then used to create a new column (' vrai_quartier ') containing the standardized neighborhood names. To check once again the Arrondissement variable, a data file downloaded from Wikipedia was used to check the consistency of neighborhoods to their arrondissement for the city of Douala.

3.2.3. Standardization and geolocation of the Neighborhood variable

The geographic coordinates are then retrieved using geopy 's Nominatim service , with a specific search targeting neighborhoods in Douala, Cameroon. Ultimately, this methodology enabled the precise geolocation of 211 distinct neighborhoods, transforming raw textual data into usable geographic information, with a latitude and longitude for each neighborhood. Using a pre-existing CSV file ('Vrai_Quartier.csv') as a geographic reference and removing unlocalizable neighborhoods ensures the quality and accuracy of the geolocated dataset.

In short, the variables Profession and Neighborhood of residence presented spelling problems and inconsistencies in the entry, for example, we could find 'Commerçant (e)', 'COMMERCANT', 'COMMERÇANTE'. For the neighborhoods, the dictionary made it possible to have a dictionary of this type: ' logbaba ': [' logbaba ', 'log-baba', 'log baba', ' logbaba ']

The following steps have been taken to resolve this issue:

- Convert all values to lowercase.
- Clustering similar names using the fuzzywuzzy library with a 90% similarity threshold.
- Geolocation of neighborhoods using geopy and the Nominatim service of OpenStreetMap .

- Export and import of geographic coordinates to reduce calculation time.
- Removal of non-geolocated neighborhoods.
- Calculation and aggregation of indicators by district.

3.3. Modeling and analysis

3.3.1. Textual and sentimental analysis

A sentiment analysis was carried out on the variable **Other to be specified** by applying the following methods:

- Removed **stop words** .
- Construction of a **word cloud** to visualize the most frequent terms.

4. Statistical analyses and modeling of the phenomenon of eligibility status

Statistical analyses :

- **Descriptive statistics** are calculated to understand the distributions of variables.
- **Chi-square tests** are performed to assess the independence between categorical variables and their relationship with donation eligibility.
- **Analyses of variance (ANOVA)** are performed to compare the means of numerical variables between different groups (e.g., age versus eligibility or blood type).
- annual and monthly **growth rates in the number of donations**
- **Kolmogorov-Smirnov test** to compare age distributions across gender and eligibility status.
- **Clustering : The K-** Means algorithm is applied to identify groups (clusters) of donors\candidates based on their demographic characteristics and donation history. The optimal number of clusters is determined using the elbow method and silhouette scoring. The clustering results are visualized and profiled.
- **Multiple Correspondence Analysis (MCA)** : MCA is used to analyze the relationships between multiple categorical variables simultaneously and to visualize the proximities between the modalities of these variables7.... Clustering techniques (Gaussian mixture and DBSCAN) are applied to the MCA coordinates.

Campaign Effectiveness Analysis : Donation history (date of last donation) is analyzed in relation to demographic factors to potentially assess the impact of past campaigns.

3.3.2 Modeling eligibility status

Preprocessing was conducted separately for numeric and categorical variables, using sophisticated techniques from the scikit-learn library.

For categorical variables ('Education Level', 'Gender', 'Marital Status (MS)', 'Common Occupation', 'Religion', 'Has he/she ever donated blood') a specific treatment was followed: imputation of missing values by the most frequent modality, then encoding via TargetEncoder. This method allows capturing the relationship between categorical variables and the target variable. The variable of interest being the donation eligibility status is transformed into two modalities: eligible and not eligible (definitively or temporarily).

Imputation of missing values of numerical variables was carried out according to the following steps:

- Selection of variables '**Hemoglobin level**' , '**Weight**' and '**Height**' .
- Application of **IterativeImputer** with 10 iterations to estimate missing values.

Data splitting was performed in an 80/20 ratio for training and testing sets, with a random_state of 42 ensuring reproducibility.

We then opted for a RandomForestClassifier classifier because of its power to predict unbalanced classes, integrated into a complete pipeline combining preprocessing and classification.

Evaluation with appropriate metrics: We highlighted the use of **F1 score** in addition to precision and classification ratio. F1 score is a harmonic mean of precision and recall, making it more sensitive to performance on the minority class than precision alone. The use of this metric suggests consideration of potential class imbalance in the target variable 'DONATION ELIGIBILITY.'

Stratified Cross-Validation: The StratifiedKFold implementation ensures that each cross-validation fold contains approximately equal proportions of each class. This is crucial when working with imbalanced data, as unstratified cross-validation could lead to folds with very little or no representation of the minority class, resulting in biased model evaluation.

Probability threshold optimization: This technique is commonly used to adjust the trade-off between precision and recall, which is particularly important in imbalanced class scenarios where one may want to favor the detection of the minority class (higher recall) at the expense of potentially increasing false positives (lower precision), or vice versa.

Although these elements indicate some sensitivity to the problem of unbalanced classes, we could also mention the use of explicit class balancing techniques such as:

- **Oversampling of the minority class .**
- **Undersampling of the majority class .**
- **Generation of synthetic data for the minority class (e.g., SMOTE).**

Chapter 2

PRESENTATION OF RESULTS

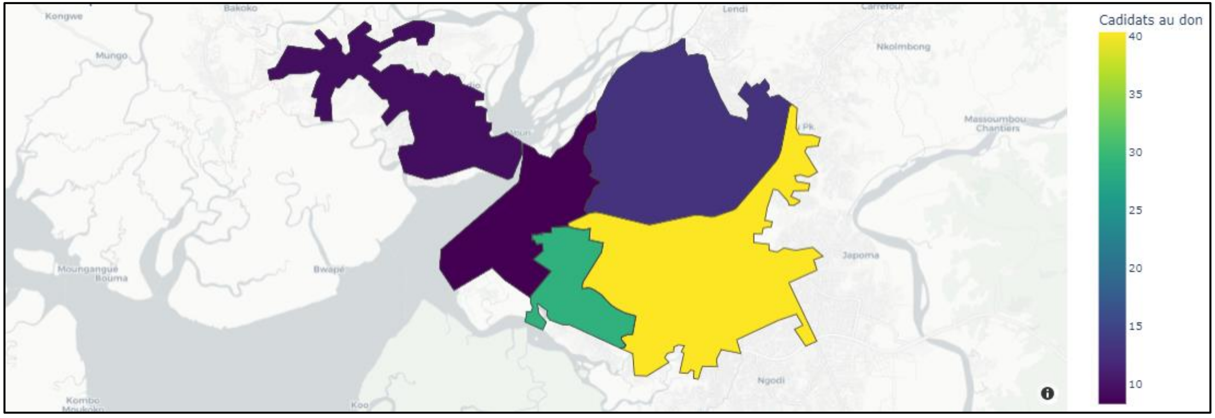
This section presents the results of the analyses.

I. Mapping the distribution of blood donations

I.I. Distribution of blood donor candidates in Douala

This map illustrates the geographical distribution of blood donor candidates in 5 districts of Douala, revealing significant territorial disparities: the yellow areas (Douala 3) in the east and southeast concentrate the largest number of donors (around 40), followed by the green sectors (Douala 5) in the south-center (25-30 donors), while the dark purple areas (Douala 4 and 1) in the west and northwest have the lowest rates (less than 10 donors), with an intermediate blue-purple area (Douala 2) in the north (15-20 donors). These significant variations could be explained by several factors such as differences in population density, accessibility of collection centers, the effectiveness of awareness campaigns or even socio-economic and cultural factors specific to each district, highlighting the need to develop targeted strategies to improve participation in blood donation in currently underrepresented areas.

Graph 1: Mapping of the distribution of blood donation candidates by district (Douala)

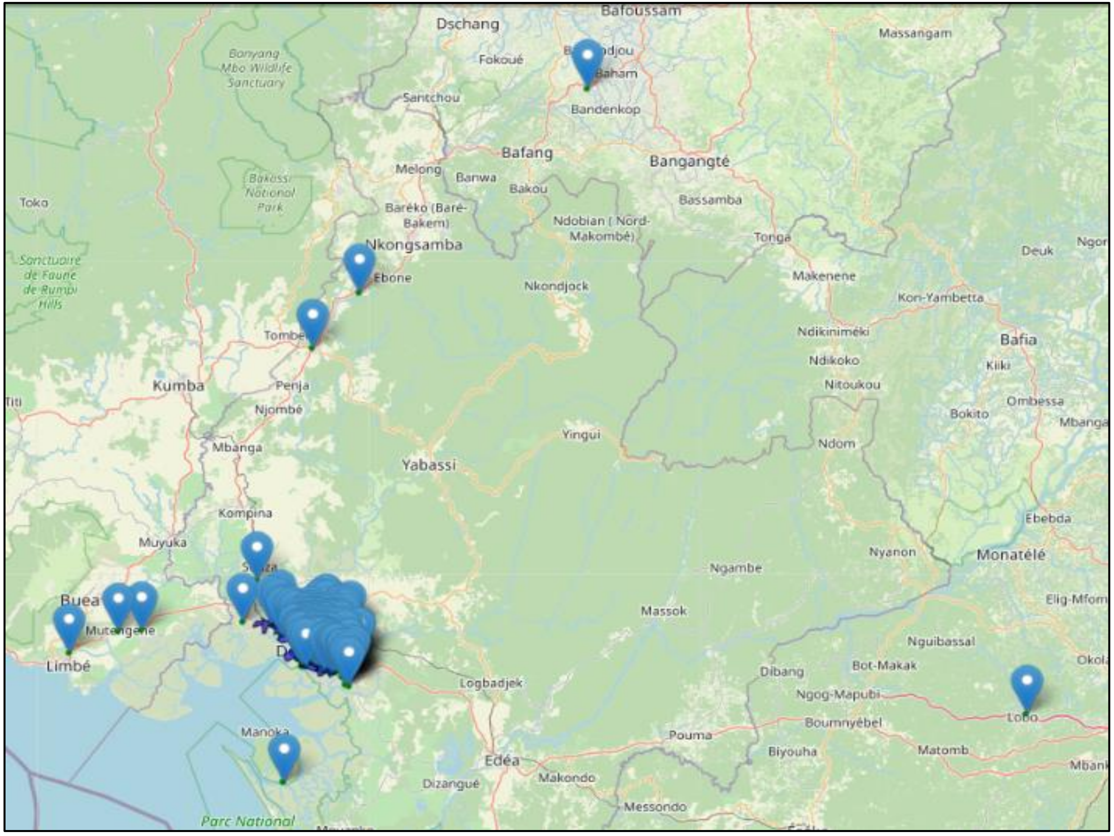


Source: Authors, from python

I.2. Location of blood donor candidates across the country

This map illustrates the geographic location of blood collection centers or donors, revealing a very marked concentration in the urban region of Douala (south-central part of the map) where a dense cluster of blue markers is visible, while other points are scattered much more sparsely across the territory: a few in the north towards Bafoussam and Bangangté , several along the west coast near Limbe and Buea, aligned markers appearing to follow a main road axis, and isolated points such as the one near Lobo in the east. This unbalanced distribution suggests a strong centralization of blood donation activities in the main urban areas, probably reflecting disparities in terms of population density, available health infrastructure and accessibility, which raises questions about the equity of access to blood donation services for populations in rural or remote areas of the country.

Graph 2: Location of blood donor candidates in Cameroon



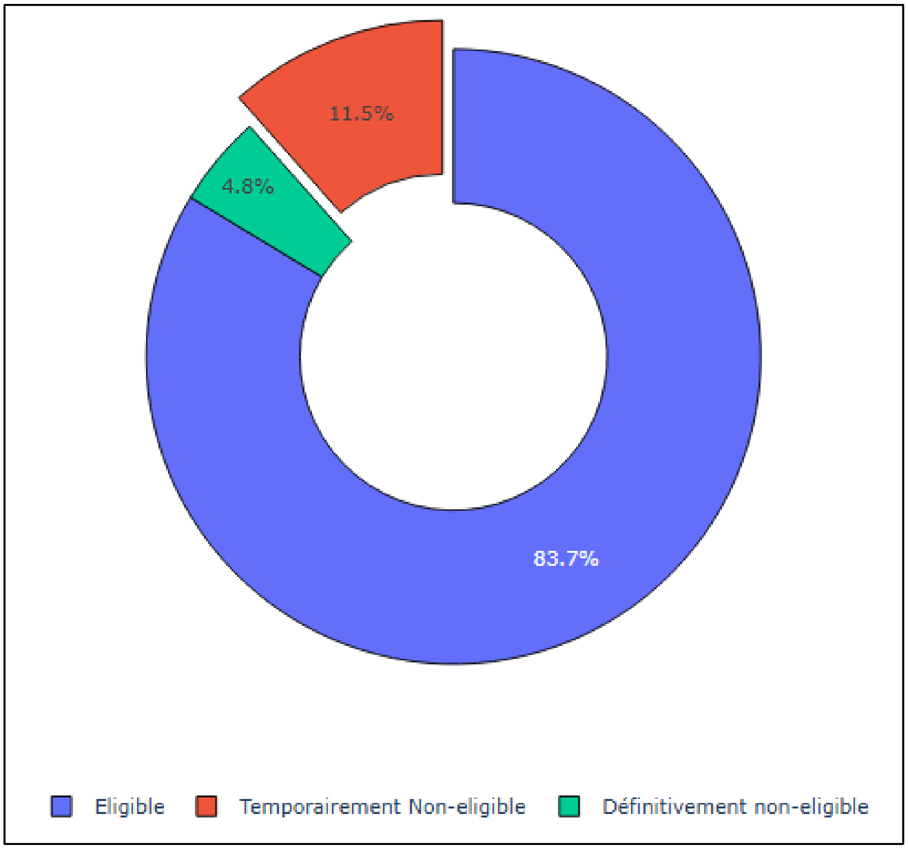
Source: Authors, from python

2. Health Conditions & Eligibility

2.1. Distribution of the 1915 respondents according to their eligibility to donate blood

Analysis of this graph reveals the distribution of the 1915 respondents according to their eligibility to donate blood, showing that the vast majority (1602 people, or 83.7%) are eligible to donate, while a moderate proportion (221 people, or 11.5%) face temporary ineligibility that may be lifted later, and a minority (92 people, or 4.8%) face definitive ineligibility that will permanently prevent them from donating blood, which overall indicates a strong donor potential in this sample studied.

Graph 3: Distribution of the 1915 respondents according to their eligibility to donate blood

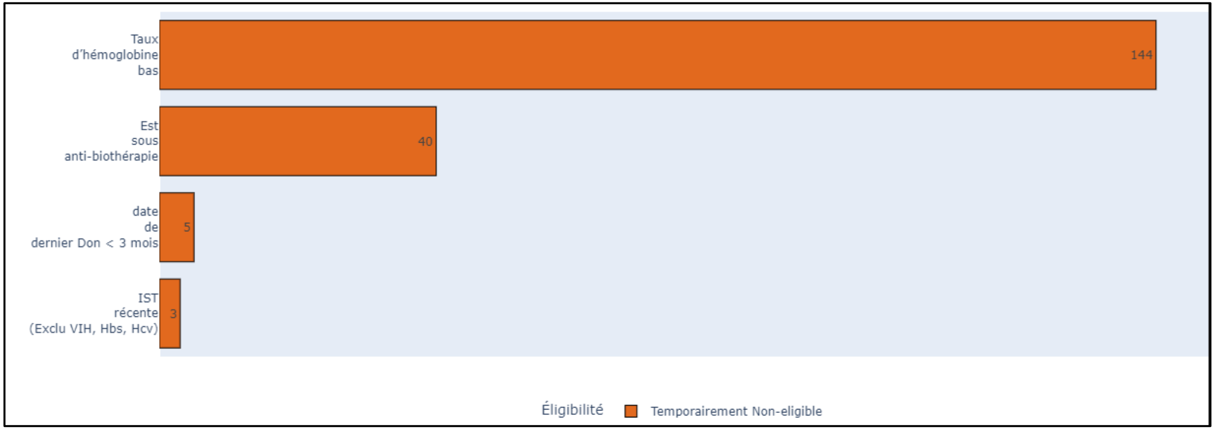


Source: Authors, from python

2.2. Distribution of temporarily ineligible candidates according to the reasons given

This graph shows the reasons why people are temporarily ineligible to donate blood in Cameroon, as well as the number of cases for each reason. Reasons for temporary exclusion include a recent STI (3 cases, excluding HIV, hepatitis B and C), a last blood donation less than 3 months ago (5 cases), current antibiotic treatment (40 cases), and low hemoglobin level (144 cases). Data suggest that anemia (low hemoglobin level) is the main cause of temporary exclusion, followed by antibiotic treatment, recent donation, and recent STI.

Chart 4: Distribution of temporarily ineligible candidates according to the reasons given

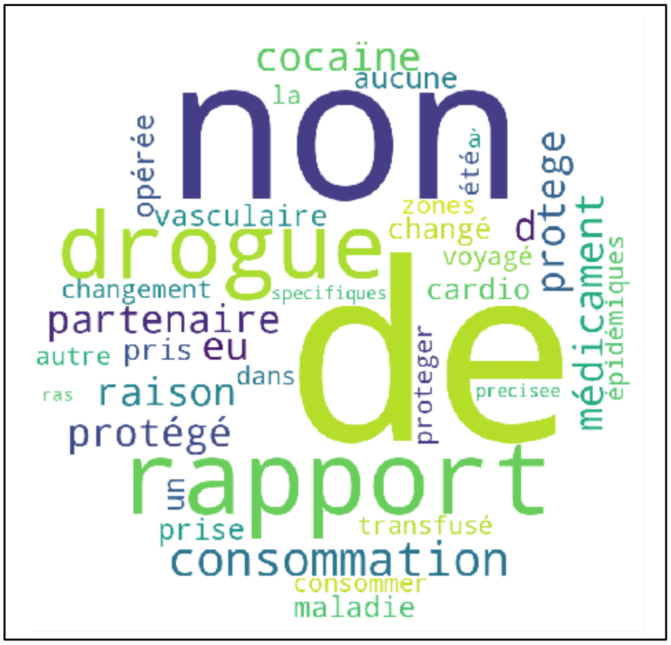


Source: Authors, from python

1.3. Other reasons for temporary ineligibility

This word cloud illustrates the other main causes of temporary ineligibility for blood donation, highlighting factors such as drug use (especially cocaine), unprotected sex or sex with new partners, as well as various medical considerations such as recent operations, cardiovascular problems, taking specific medications, or a history of transfusions. The prominent terms "non," "de," "rapport," and "drug" highlight the particular importance of risky behaviors in the temporary exclusion criteria, these restrictions aimed at ensuring the safety of both donors and recipients in the blood donation process.

Chart 5: Other reasons for ineligibility

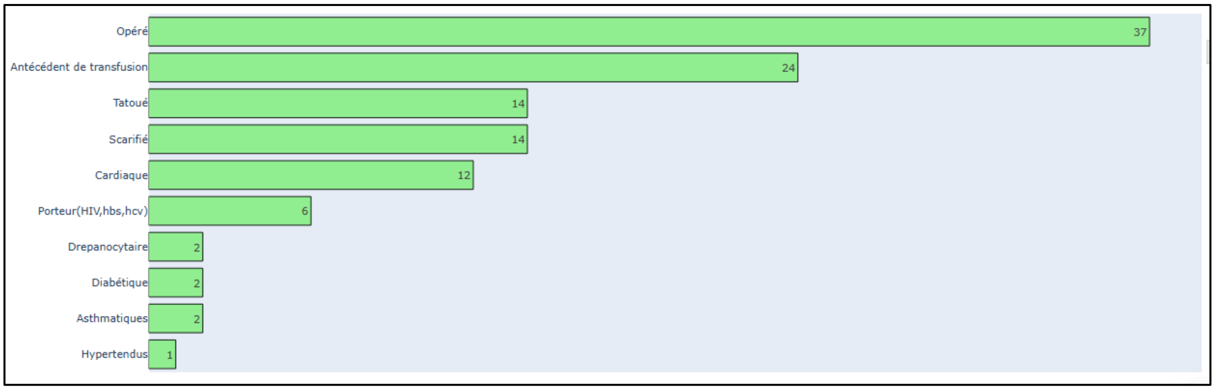


Source: Authors, from python

2.3. Distribution of respondents according to the different reasons for permanent ineligibility to donate blood

This graph shows the different reasons for permanent ineligibility to donate blood, classified by number of cases. We observe that among those permanently ineligible, the most frequent reason is related to a history of surgery (37 cases), followed by a history of blood transfusion (24 cases). Tattooed and scarified people each represent 14 cases. Heart problems concern 12 people. Virus carriers (HIV, HBs , HCV) account for 6 cases. Less represented medical conditions include sickle cell disease, diabetics and asthmatics (2 cases each), as well as hypertensive patients (1 case). All these conditions lead to a permanent exclusion from donating blood according to the criteria established for this population.

Chart 6: Distribution of respondents according to reasons for permanent ineligibility

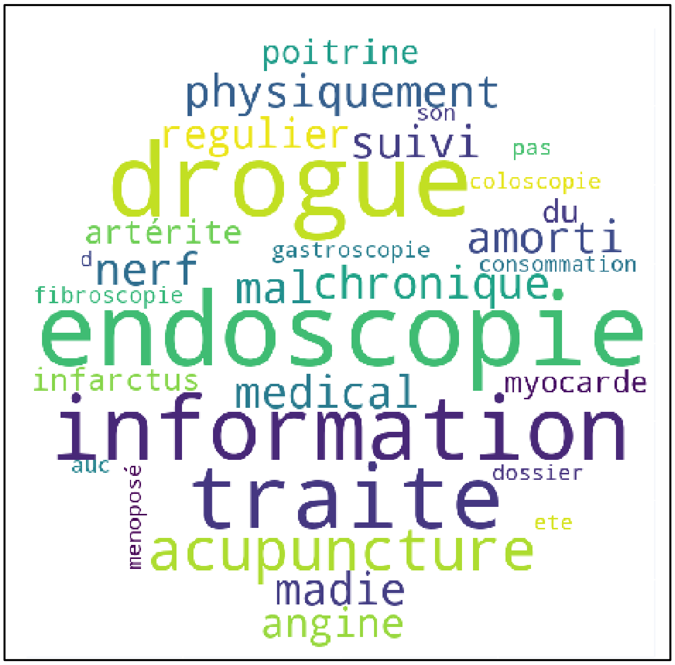


Source: Authors, from python

1.4. Other reasons for permanent ineligibility

This word cloud illustrates the additional reasons for permanent ineligibility to donate blood, where the terms "information", "treated", "endoscopy" and "drug" dominate, revealing the importance of medical history in disqualifying potential donors; there are various categories of causes including invasive procedures (endoscopy, fibroscopy, colonoscopy, gastroscopy), cardiac conditions (infarction, myocardium, angina, arteritis), specific medical treatments (acupuncture), neurological problems (nerve, chronic) and substance use (drug, consumption), indicating that these medical factors constitute rigorous criteria for permanent exclusion to preserve the safety of the transfusion system.

Chart 7: *Additional reasons for permanent ineligibility to donate blood*

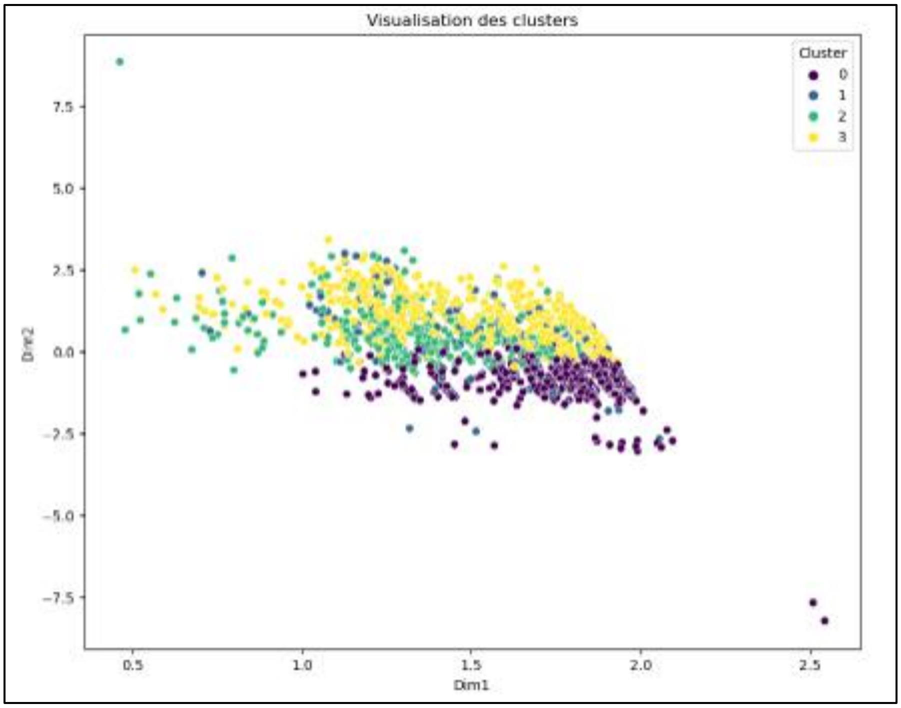


Source: Authors, from python

3. *Profiling Ideal Donors*

To profile ideal blood donors, we implemented a clustering approach based on the K- means algorithm . Our analysis incorporated both numerical (age, occupation, and residential area) and categorical (gender, marital status, religion, educational level, and previous donation experience) variables. Rigorous preprocessing was applied via a pipeline combining standardization of numerical variables and one-hot encoding of categorical variables. To determine the optimal number of segments, we jointly employed the elbow method and silhouette scoring, analyzing configurations of 2 to 8 clusters. This double validation allowed us to identify 4 clusters as the optimal configuration. Visualization of the results was facilitated by dimensional reduction using TruncatedSVD , allowing the multidimensional data to be projected onto a two-dimensional plane. This segmentation allowed us to distinctly identify the dominant characteristics of donors most likely to contribute regularly to blood donation campaigns.

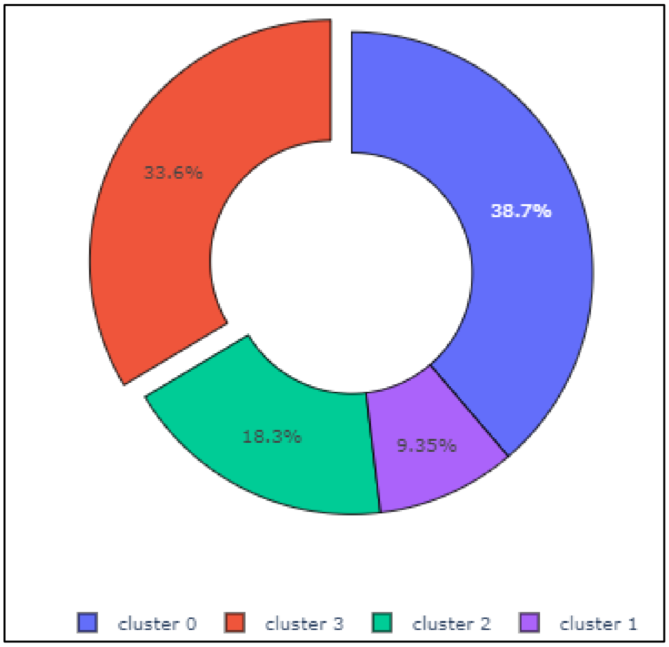
Graph 8: Cloud of individuals according to the clusters created



Source: Authors, from python

According to the descriptive statistics, our clustering segmentation generated four distinct profiles of potential donors, with an uneven distribution between the segments. Cluster 0 appears to be the majority group (with 38.7% of the sample), closely followed by cluster 3, which comprises 33.6% of individuals. Clusters 2 and 1 are less represented, with 18.3% and 9.35%, respectively.

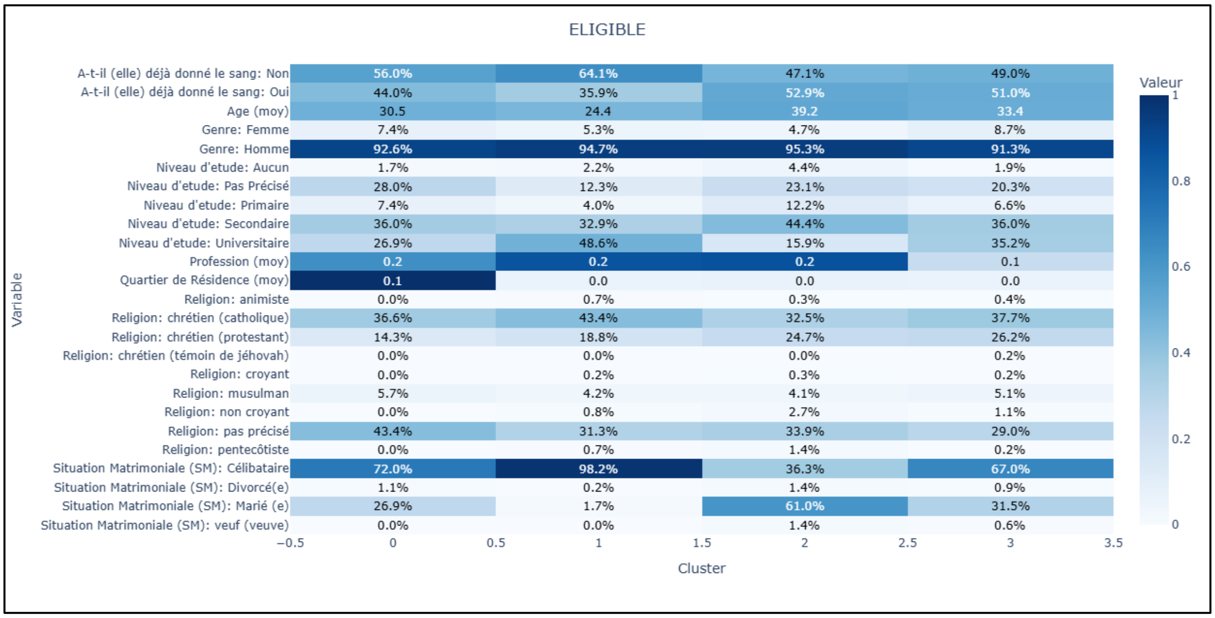
Graph 9: Distribution of respondents according to clusters



Source: Authors, from python

Now we move on to characterizing the different clusters.

Chart 10: Characterization of clusters



Source: Authors, from python

The visualization above allows us to characterize in detail the four clusters of potential donors according to various demographic and behavioral variables. By systematically analyzing the graph:

Cluster 0 (741 individuals - 38.7%):

- High presence of individuals who have never given blood (56.0%)
- Over-representation of people with secondary education (36.0%)
- High prevalence of Christian religions (Catholics 3 6.6% and Protestants 14.3%)
- Mostly men (92.6 %), single (26.5%)
- Average characteristics for other variables

Cluster 1 (179 individuals - 09.35%):

- relatively low rate of individuals having already donated blood (35.9%)
- Higher proportion of higher education level (4 8.6%)
- High rate of single people (98.2%)
- Notable presence of Catholics (43.4%) but more religiously diverse

Cluster 2 (350 individuals - 18.3%):

- Relatively high rate of people having already donated blood (52.9%)
- High concentration of secondary education level (44.4%)
- High prevalence of Christian religions (Catholic 32.5% and Protestant 24.7%)
- Significant proportion of married (61.0%) and single (36.3%) people
- Mostly men

Cluster 3 (645 individuals - 33.6%):

- High rate of people having already donated (51.0%)
- Balanced level of education between secondary (36.0 %) and higher (35.2%)
- High proportion of Catholics (37.7%)
- High rate of married people (3 1.5%) and single people (67%)

Recommendations Strategic

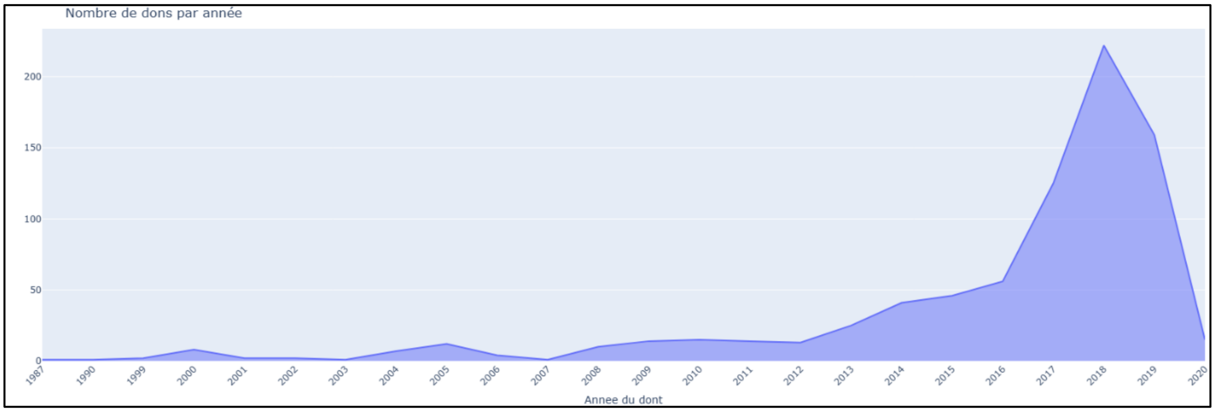
- **Build loyalty** in Clusters 2 & 3 through recognition programs (donor cards, SMS reminders).
- **Convert** Cluster 1 through partnerships with schools/businesses.
- **Mobilize** Cluster 0 through local campaigns (religious leader, local posters).

4. Campaign Effectiveness Analysis

4.1. Evolution of the number of blood donations per year

This graph of the evolution of the number of blood donations per year reveals a relatively stable and modest trend between 1997 and 2012, with occasional slight peaks but without major variation. From 2013 onwards, we observe the beginning of a gradual growth that accelerates markedly from 2016 onwards, peaking in 2018 with more than 200 annual donations, a dramatic increase compared to previous years. The year 2019 shows a slight decrease from the 2018 peak, while maintaining a significantly higher number of donations than before 2016, while 2020 marks a sharp drop, probably attributable to the COVID-19 pandemic that disrupted healthcare systems and blood drives worldwide, bringing the number of donations back to a level comparable to that observed before 2013.

Graph 11: *Change in the number of blood donations per year*

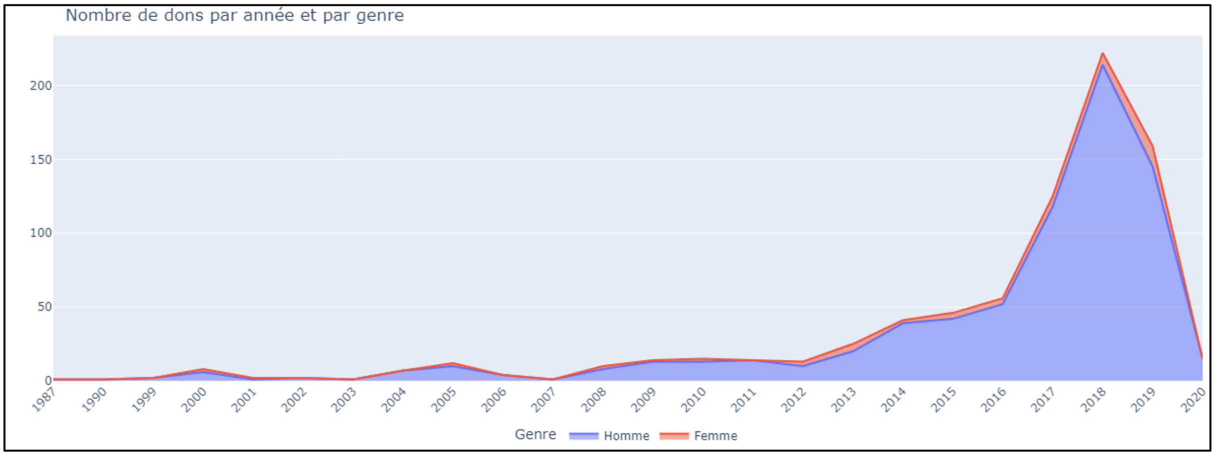


Source: Authors, from python

4.2. Evolution of the number of blood donations per year and by gender

This graph shows the evolution of the number of blood donations by year and by gender between 1987 and 2020. A similar trend is observed to the previous graphs: relatively stable and low donations until 2013, a slight increase until 2016, then a spectacular increase culminating in 2018 (around 210 donations) before a sharp drop in 2019-2020. The dominant blue area represents donations by men, while the thin red band above shows the marginal contribution of women. This marked disproportion between the genders confirms the observations of the previous diagram, indicating a very predominantly male participation in blood donation in Cameroon over the entire period studied.

Graph 12: *Change in the number of blood donations per year and by gender*

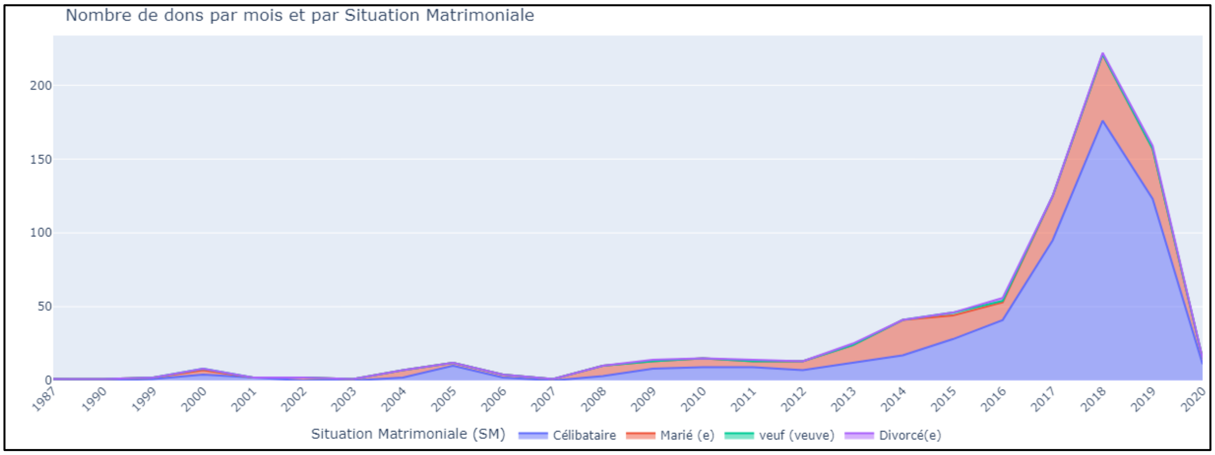


Source: Authors, from python

4.3. Number of donations per month and by marital status

Analysis of the graph "Number of donations per month and by Marital Status" reveals a striking evolution of blood donations between 1998 and 2020, characterized by a period of relative stability at low levels until 2015, followed by a slight gradual increase and then a spectacular rise between 2017 and 2018, peaking at around 200 monthly donations in 2018 before experiencing a drastic drop in 2019-2020. Single donors (represented by the blue curve) clearly constitute the majority of contributors, particularly during the 2018 peak, followed by married donors (red zone), while widowed and divorced donors (green and purple curves respectively) represent a tiny proportion of donations, suggesting either particularly effective awareness campaigns targeting single people in 2017-2018, or significant changes in blood collection policies that influenced this remarkable dynamic.

Chart 13: *Number of donations per month and by marital status*

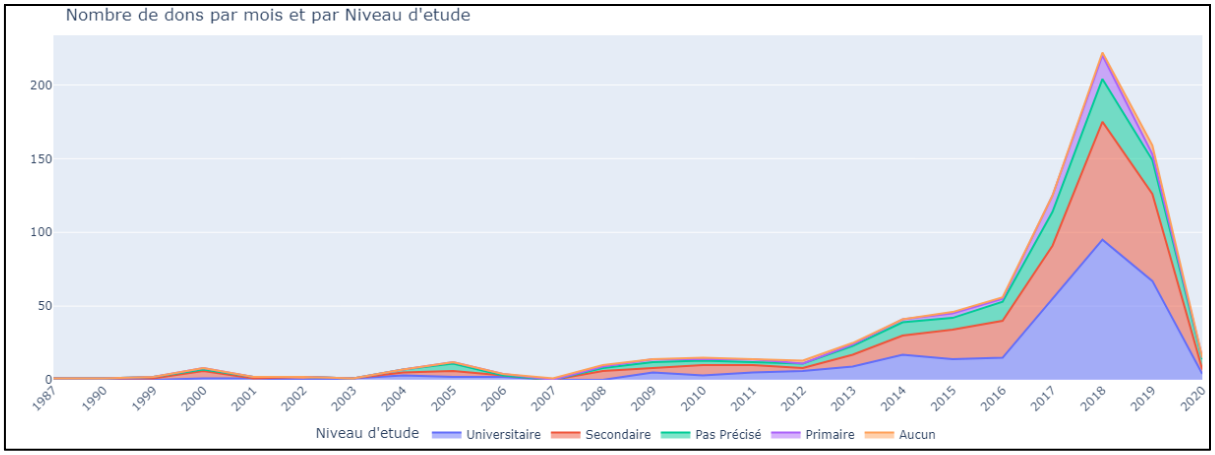


Source: Authors, from python

4.4. Monthly changes in blood donations by level of education between 1998 and 2020

This graph shows the monthly evolution of blood donations by level of education between 1987 and 2020. We observe a similar trend to the previous graph: relatively stable and low donations until 2013, followed by a slight increase until 2016, then a dramatic increase peaking in 2018 (around 220 monthly donations) before a sharp drop in 2019-2020. University donors (blue zone) constitute the largest proportion, followed by secondary school donors (red zone) and those whose level is not specified (green zone). Donors with primary school education or without formal education are almost non-existent. This distribution suggests that blood donation campaigns were particularly effective among educated populations, particularly university students, during the peak period.

Graph 14: *Monthly changes in blood donations by level of education between 1987 and 2020*

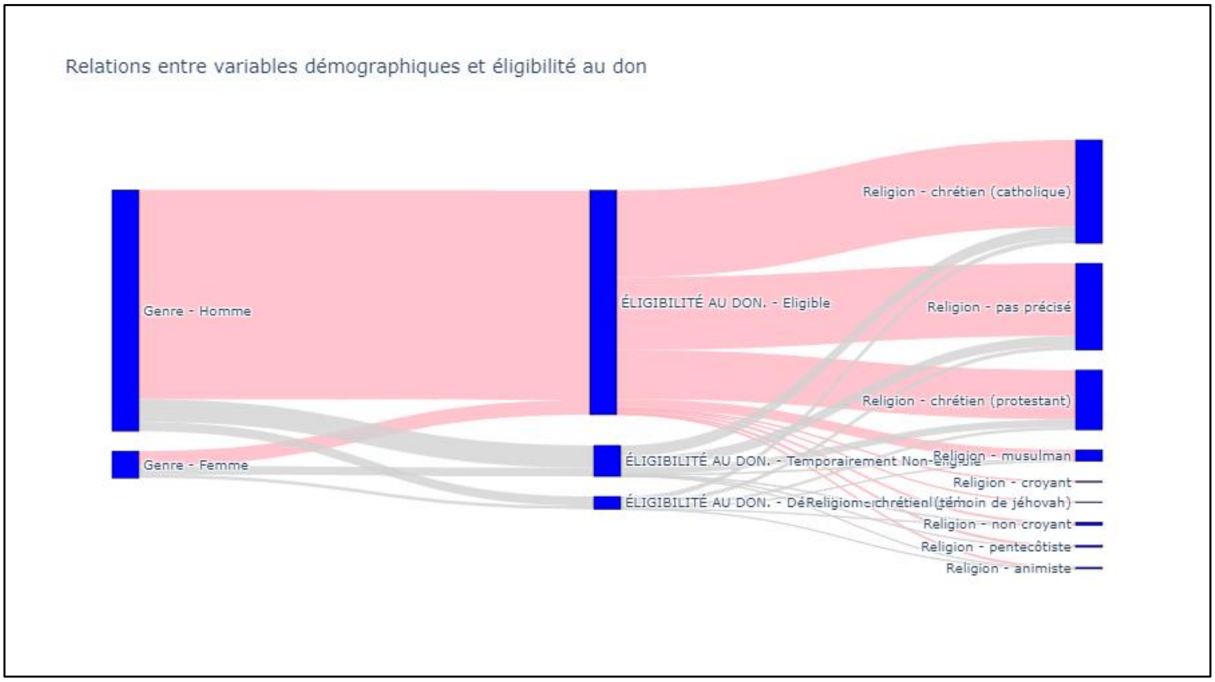


Source: Authors, from python

4.5. Relationship between demographic variables and blood donation eligibility

The flow diagram illustrates the relationship between demographic variables and blood donation eligibility, revealing a clear predominance of men among eligible donors, with a strong representation of Catholic Christians as the majority religious group, followed by those whose religion is not specified and Protestants. Conversely, women are overrepresented in the "Temporarily Ineligible" and "Referred" categories, suggesting specific barriers to their participation, while certain religious affiliations (Muslims, Jehovah's Witnesses, non-believers, Pentecostals, and animists) appear in marginal proportions among donors, highlighting the determining influence of gender and religious affiliation on blood donation eligibility and participation patterns.

Figure 15: *Relationship between demographic variables and eligibility to donate blood*



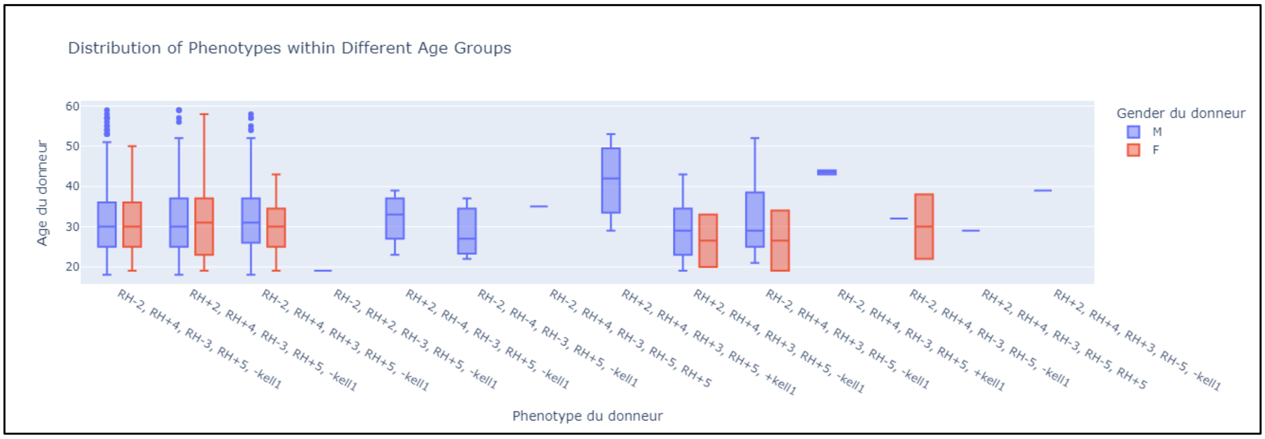
Source: Authors, from python

5. *Analysis of the characteristics of actual blood donors*

5.1. Distribution of blood phenotypes by age and gender

The 2019 blood donation campaign in Cameroon reveals a varied distribution of blood phenotypes by age and gender. Donors are predominantly between 25 and 40 years old, with a relatively balanced representation between men (in blue) and women (in pink). Rh+/- phenotypes are the most common, while some rarer phenotypes show a more restricted age distribution or gender predominance. A few outliers are observed in older age groups for certain phenotypes, suggesting occasional participation of older donors. This visualization provides valuable information to guide future blood donation awareness campaigns towards less represented demographic groups.

Figure 16: Distribution of blood phenotypes according to age and gender

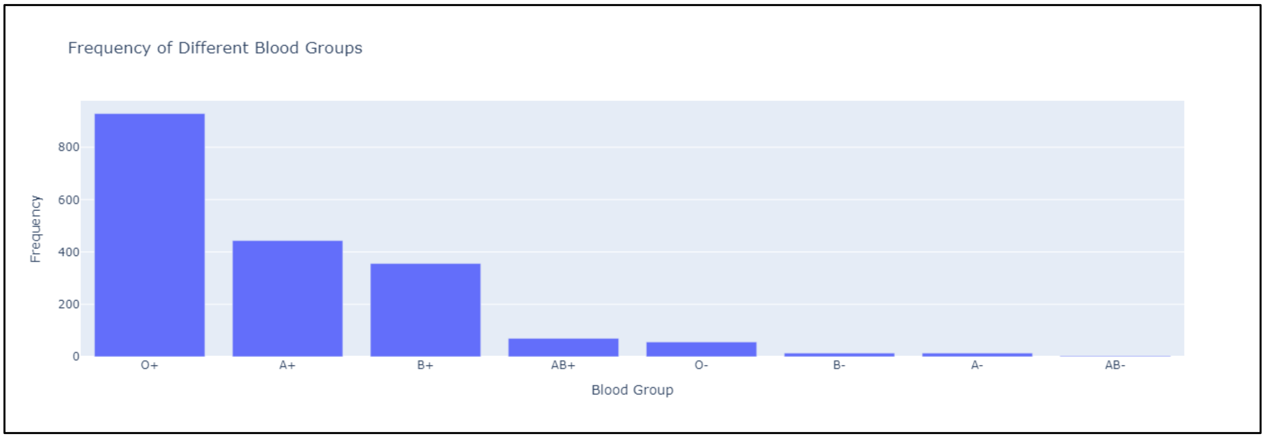


Source: Authors, from python

5.2. Distribution of blood groups in blood donations in Cameroon in 2019

The graph illustrates the distribution of blood groups in blood donations in Cameroon in 2019, showing that group O+ is the most frequent with approximately 900 donations, followed by A+ (approximately 450 donations) and B+ (approximately 400 donations), while AB+ and O- have much lower frequencies, around 100 donations each, and groups B-, A- and AB- are the least represented, with frequencies close to 0. This distribution reflects a predominance of group O+, which is consistent with genetic trends in Africa, where this group is often in the majority.

Graph 17: *Distribution of blood groups in blood donations in Cameroon in 2019*

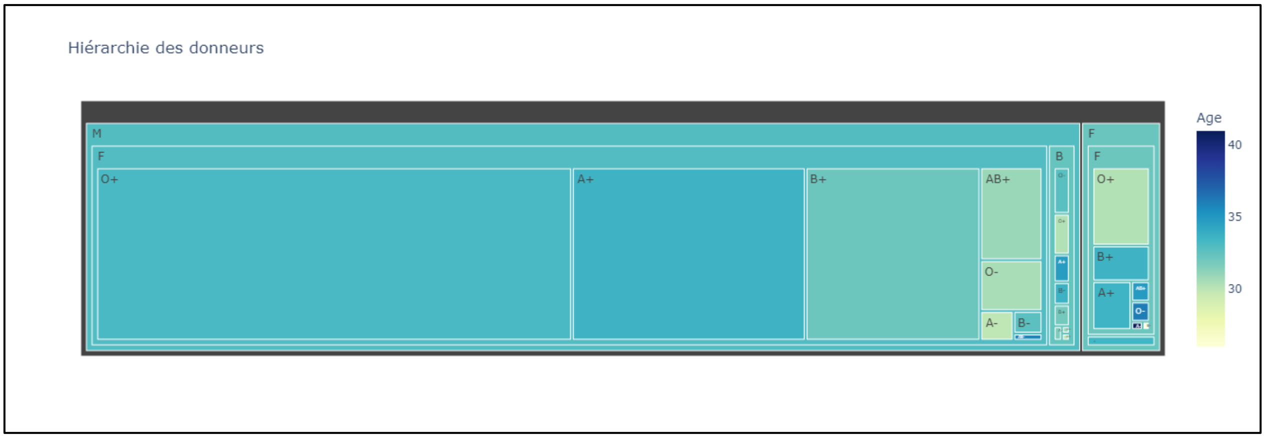


Source: Authors, from python

5.3. Donor hierarchy in 2019

The "Donor Hierarchy" treemap illustrates the distribution of blood donors in 2019 by gender and blood type, where the size of the rectangles represents the proportion of each category and the color indicates the average age (from dark blue for the oldest to light yellow for the youngest). A predominance of male (M) family donors (F) is observed, with the O+ group representing the largest proportion of donors, followed by the A+ and B+ groups, while Rh negative groups (A-, B-, AB-, O-) are significantly less represented than their positive counterparts, thus revealing targeting opportunities for future blood donation campaigns, particularly towards underrepresented groups. This same trend is also observed among female family donors and volunteers.

Chart 18: *Hierarchy of donors*



Source: Authors, from python

Chapter 3

ELIGIBILITY PREDICTION AND AI ASSISTANT

In this project, we developed a machine learning-based prediction model to assess the eligibility of new blood donors based on demographic and health data. This model, designed to meet a concrete need for automation and accuracy, aims to facilitate the donor selection process while ensuring the safety and reliability of results. By integrating this model into an API, we offer a practical and adaptable solution, enabling real-time prediction directly from a dashboard. This chapter presents the key steps in the model's design, the technical choices made, as well as its performance and application potential.

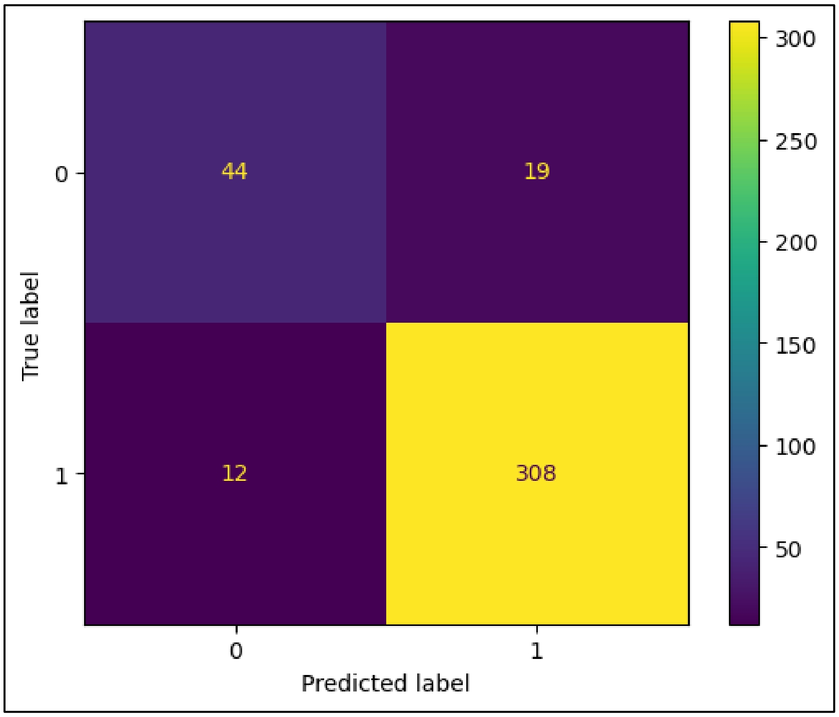
1. Summary of the approach adopted for prediction

To predict blood donation eligibility, a model was developed in Python from the filtered data (df_filtre). The explanatory variables, including characteristics such as height, weight, hemoglobin level, age, gender, or profession, were separated from the target variable (donation eligibility), and then the data were split into training and test sets (80/20) with train_test_split . Preprocessing was applied via a pipeline: missing values for numeric variables were imputed by the median and normalized with StandardScaler , while those for categorical variables were imputed by the mode and encoded with TargetEncoder . The RandomForestClassifier model , integrated into the pipeline, was trained on the training data and then evaluated on the test data, providing accuracy via accuracy_score and a detailed performance report with classification_report .

2. Model performance

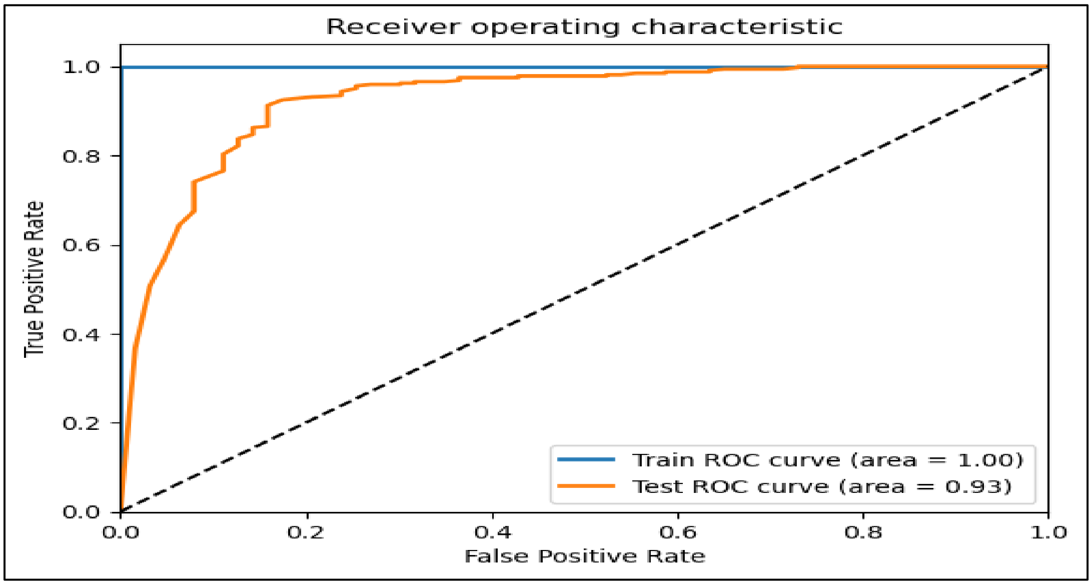
This confusion matrix shows the performance of the blood donation eligibility prediction model, for the two classes: 0 (ineligible) and 1 (eligible). Out of 383 individuals, the model correctly predicted 44 cases as ineligible (true negative) and 308 cases as eligible (true positive), indicating a good ability to identify eligible individuals. However, it incorrectly predicted 19 cases as eligible when they were not (false positive) and 12 cases as ineligible when they were (false negative), suggesting a slight tendency to overpredict eligibility.

Chart 19 : Confusion matrix



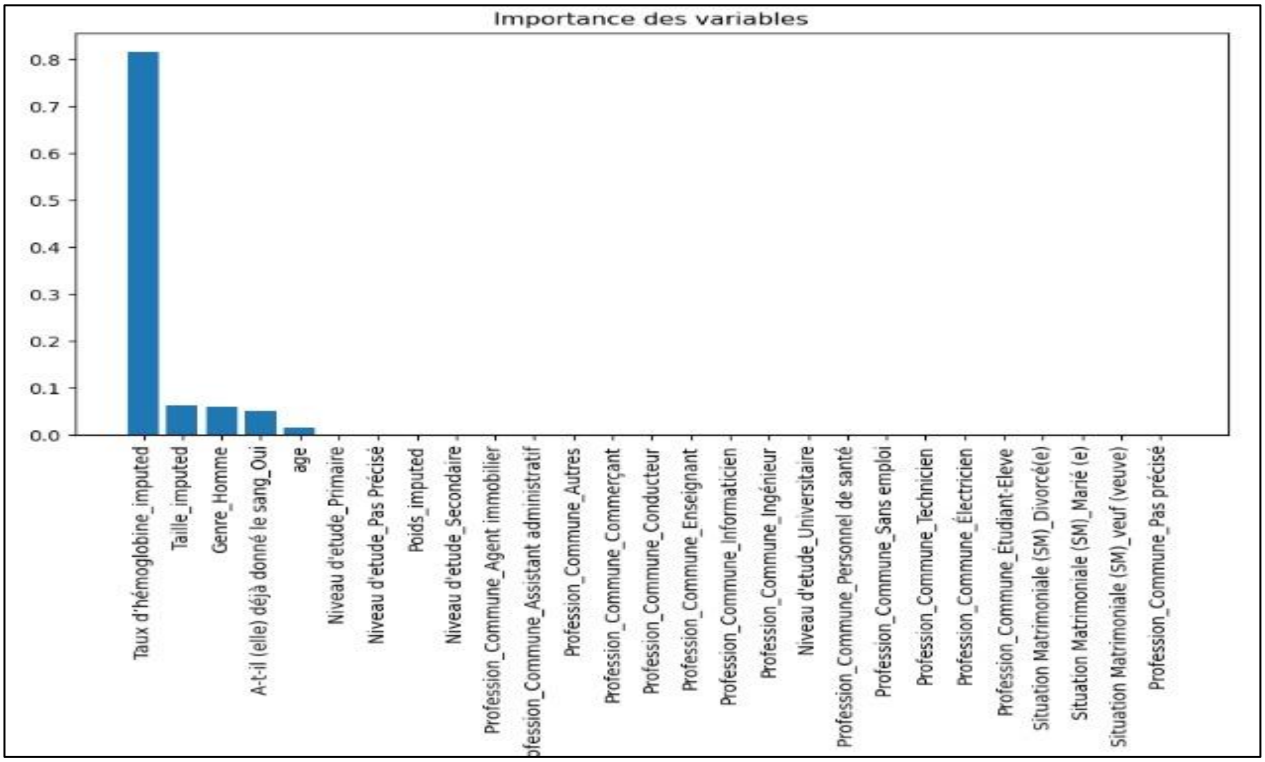
Source: Authors, from python

The ROC curve of the blood donation eligibility prediction model shows excellent performance, with an AUC of 1.0 on the training data (indicating perfect classification) and an AUC of 0.93 on the test data, which remains very good and reflects a strong ability to discriminate eligibles from ineligibles, with a low false positive rate.

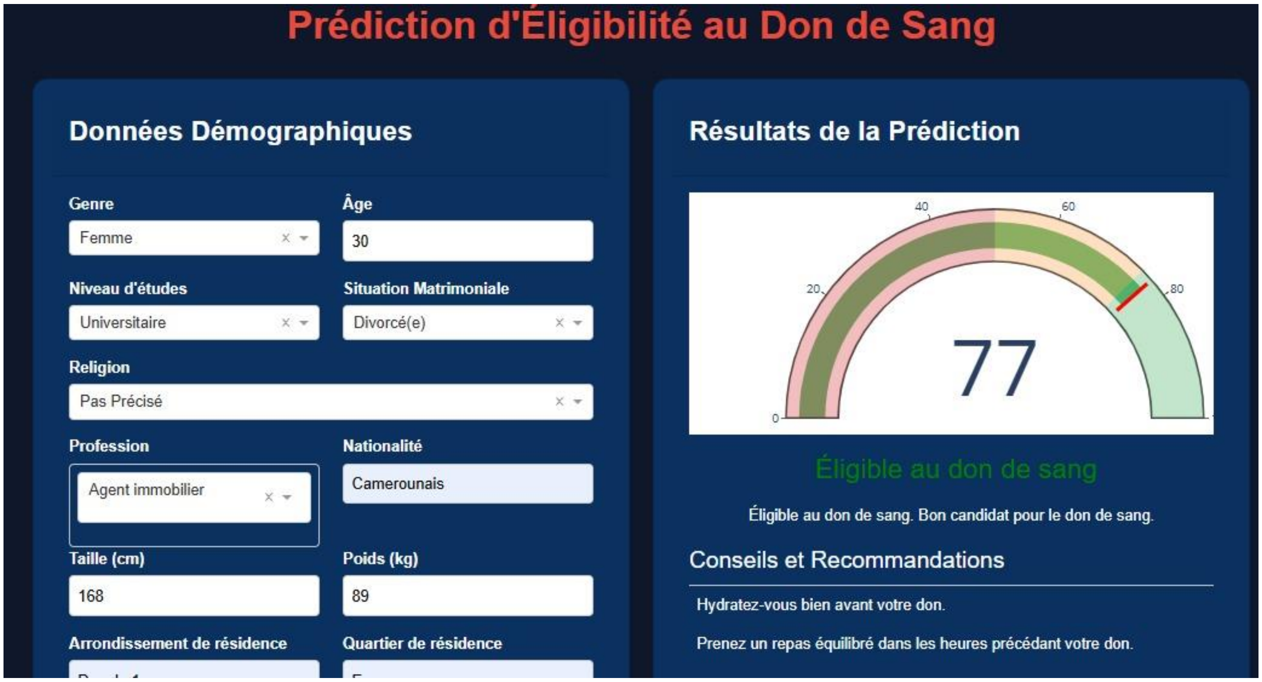


Source: Authors, from python

The analysis of the importance of variables in our blood donation eligibility prediction model reveals that the hemoglobin level variable is by far the most influential (importance ≈ 0.8), followed by height (≈ 0.1) and weight (≈ 0.1), highlighting the importance of physiological criteria and past experience in predicting, "Has he (she) already donated blood?".



3. *Prediction simulation*



Prédiction d'Éligibilité au Don de Sang

Données Démographiques

Genre: Femme

Âge: 30

Niveau d'études: Universitaire

Situation Matrimoniale: Divorcé(e)

Religion: Pas Précisé

Profession: Agent immobilier

Nationalité: Camerounais

Taille (cm): 168

Poids (kg): 89

Arrondissement de résidence

Quartier de résidence

Résultats de la Prédiction

Éligible au don de sang

Éligible au don de sang. Bon candidat pour le don de sang.

Conseils et Recommandations

Hydratez-vous bien avant votre don.

Prenez un repas équilibré dans les heures précédant votre don.

Source: Authors, from python

Here, the model predicts that the individual is 77% eligible to donate.

AI ASSISTANT

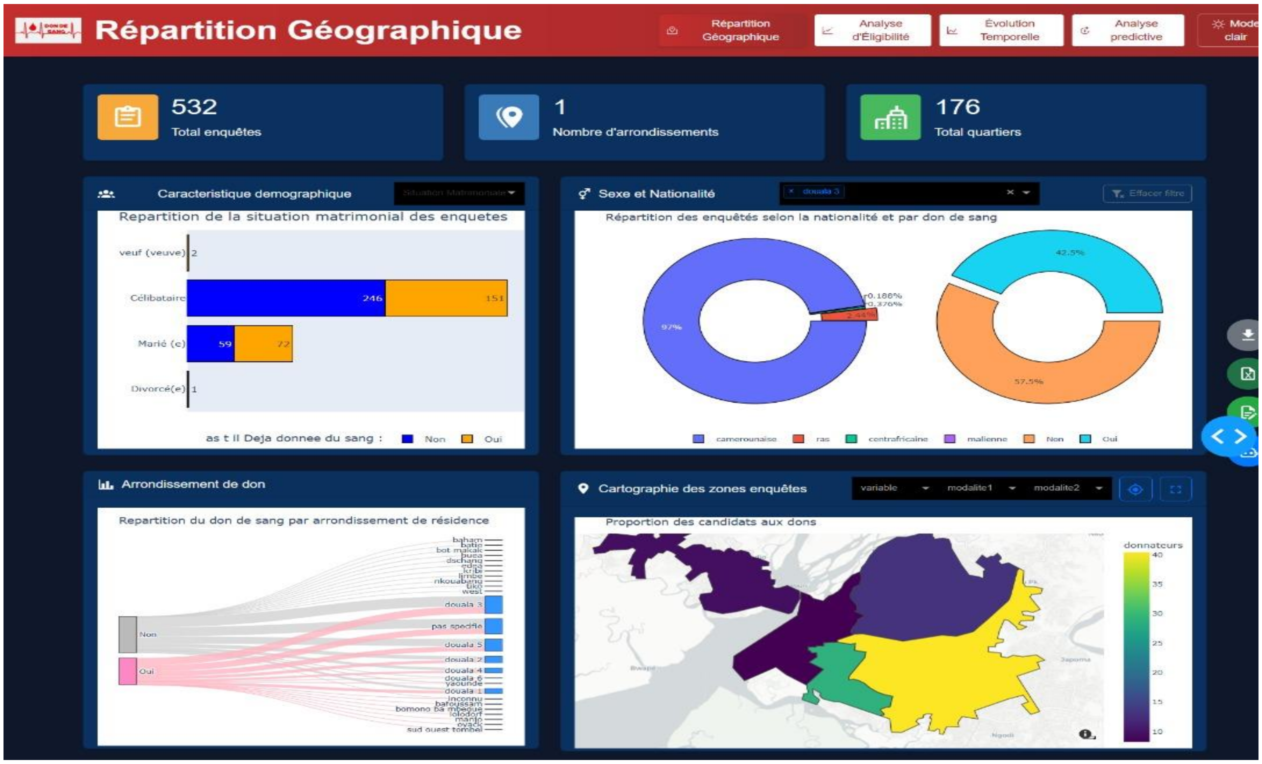
We integrated an AI assistant into the dashboard to assist users in interpreting the various indicators, help them understand concepts related to blood donation, and guide them in decision-making based on the results displayed. Adding a chatbot to a dashboard dedicated to blood donation in Cameroon has a major advantage: it offers personalized and immediate assistance, allowing users, whether healthcare professionals or campaign organizers, to better understand the data, identify potential donor profiles, and optimize blood collection strategies, while taking into account local specificities, such as cultural or logistical barriers, to improve the efficiency and impact of blood donation initiatives in the country.

DASHBOARD PRESENTATION

The Dashboard that we have set up is made up of the following 4 tabs:

To access the Dashboard, click here .

Tab 1: Geographic distribution



Répartition Géographique

532 Total enquêtes

1 Nombre d'arrondissements

176 Total quartiers

Caracteristique demographique

Repartition de la situation matrimonial des enquetes

Sexe et Nationalité

Répartition des enquêtés selon la nationalité et par don de sang

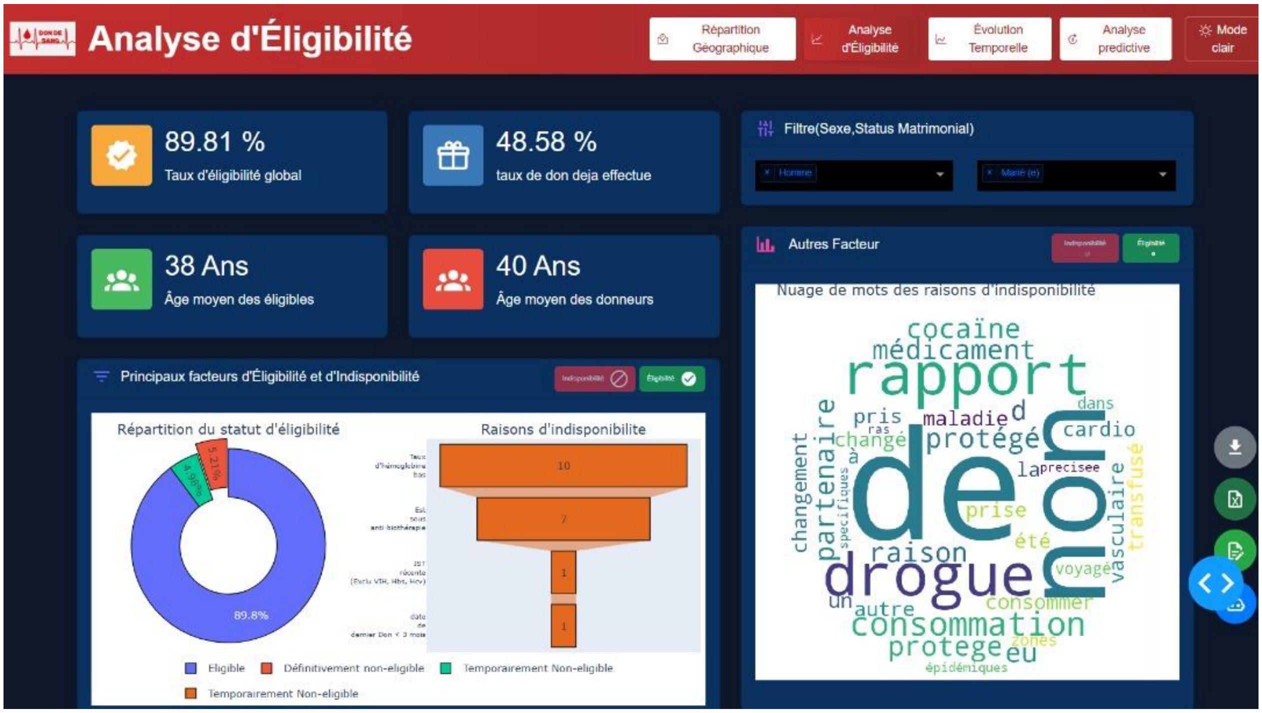
Arrondissement de don

Repartition du don de sang par arrondissement de résidence

Cartographie des zones enquêtes

Proportion des candidats aux dons

Tab 2: Eligibility Analysis



Analyse d'Éligibilité

89.81 % Taux d'éligibilité global

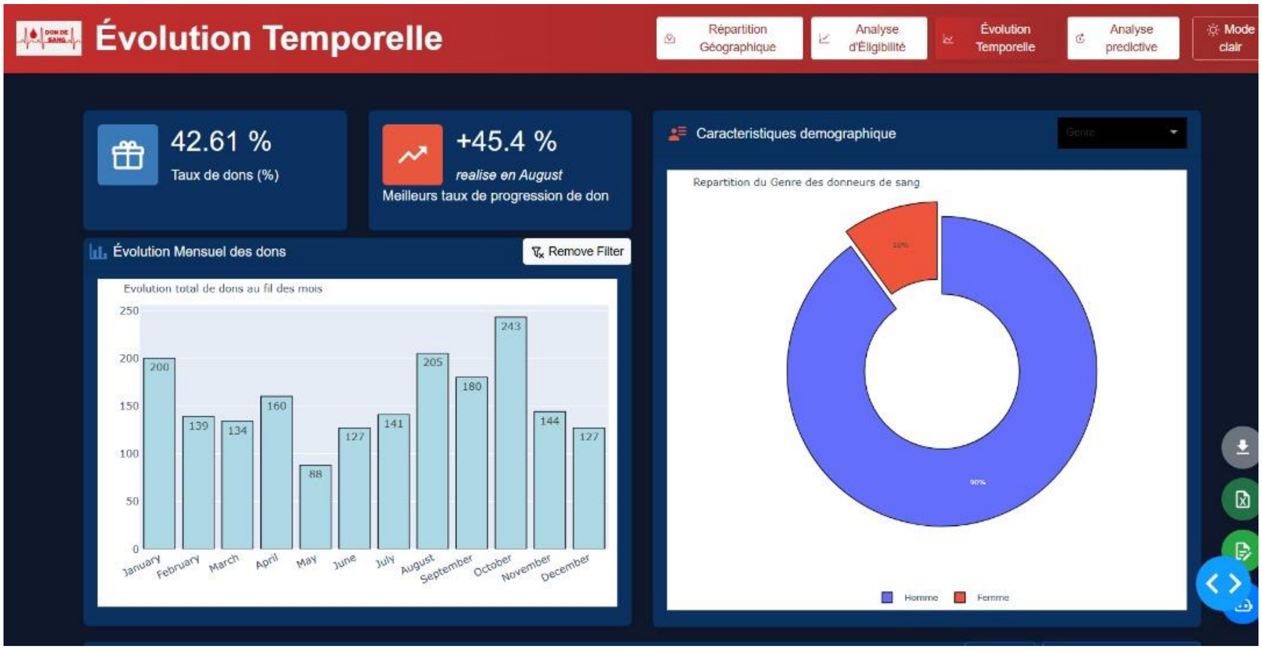
48.58 % taux de don deja effectue

38 Ans Âge moyen des éligibles

40 Ans Âge moyen des donneurs

Filtre(Sexe,Status Matrimonial): × Homme; × Marié (e)

Tab 3: Time Evolution

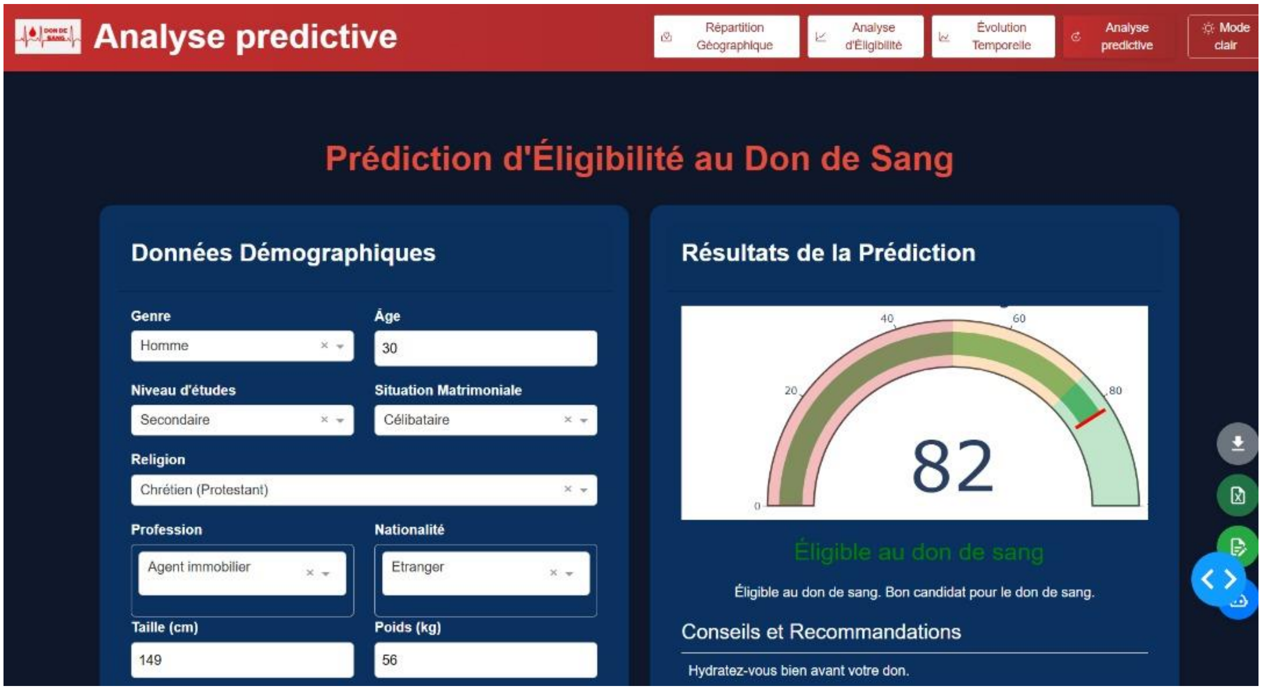


Évolution Temporelle

42.61 % Taux de dons (%)

+45.4 % *realise en August* Meilleurs taux de progression de don

Tab 4: Predictive Analysis



Prédiction d'Éligibilité au Don de Sang

Données Démographiques

Genre: Homme

Âge: 30

Niveau d'études: Secondaire

Situation Matrimoniale: Célibataire

Religion: Chrétien (Protestant)

Profession: Agent immobilier

Nationalité: Etranger

Taille (cm): 149

Poids (kg): 56

Résultats de la Prédiction

Éligible au don de sang

Éligible au don de sang. Bon candidat pour le don de sang.

Conseils et Recommandations

Hydratez-vous bien avant votre don.

Kobotoolbox Questionnaire

To avoid database inconsistency issues and facilitate future data collection, we have planned an electronic questionnaire created using Kobotoolbox . The link to the questionnaire has been integrated into the dashboard . To access the questionnaire: click here

This chapter presented a blood donation eligibility prediction model, developed in Python with a RandomForestClassifier on filtered data, showing excellent performance (AUC of 1.0 in training, 0.93 in testing) and good identification of eligibles (308 true positives out of 383), despite some errors (19 false positives, 12 false negatives). The key variables are heomoglobin level , past donation experience (importance ≈ 0.1), height (≈ 0.1) and weight (≈ 0.1). An AI assistant integrated into the dashboard (four tabs: Geographic distribution, Eligibility analysis, Time evolution, Predictive analysis) helps to interpret the results and optimize collection strategies in Cameroon, taking into account local specificities.

Chapter 4

LIMITATIONS ET RECOMMENDATIONS

4. LIMITATIONS OF THE STUDY

This work has some limitations. First, the sample of 1,915 individuals, although significant, is mainly centered in Douala, which limits national representativeness, particularly for rural areas where infrastructure and behaviors differ. Second, class imbalance (83.7% of eligible individuals) may bias predictions, despite efforts to optimize the probability threshold, and cultural or logistical barriers specific to Cameroon (beliefs, access to centers) were not fully integrated into the model.

2. RECOMMENDATIONS

The results of this study call for **targeted optimization** of blood donation campaigns in Cameroon. Geographically, it is crucial to strengthen collections in underrepresented areas (Douala 1, 4 and rural regions) via mobile units and local partnerships, while decentralizing collection centers along identified axes (Bafoussam, Limbe). To improve eligibility, specific actions are required: combating anemia (iron supplementation, nutritional advice) and raising awareness of temporary contraindications (clear guides, pre-verification mobile application). Cluster analysis reveals that educated single men (Clusters 2 & 3) constitute the core of regular donors: their loyalty is achieved through recognition programs (badges, personalized reminders) and collections adapted to their availability. Groups with underutilized potential (Cluster 1: academics) require partnerships with schools and businesses, while non-donors (Cluster 0) must be mobilized through community campaigns, particularly in churches. Furthermore, proactive inventory management is essential, with targeting rare blood groups (AB-, B-) and anticipation of seasonal shortages. Finally, digitalization (donor tracking platform, predictive analysis) and crisis-adapted protocols (e.g., decentralized post-COVID blood drives) will ensure the resilience of the system. These measures, combining data- driven analysis and local anchoring, will make it possible to effectively address the disparities identified and ensure a sustainable blood supply.

CONCLUSION

This report explored in-depth the dynamics of blood donation in Cameroon, revealing both persistent challenges and strategic opportunities to improve the supply of blood products. Analysis of data from 1,915 potential donors highlighted several key findings: a **high geographical concentration** of donations in Douala (particularly in districts 3 and 5), **high overall eligibility (83.7%)** but hampered by factors such as anemia (the main cause of temporary exclusion), and **distinct donor profiles** identified by clustering (with Clusters 2 and 3 emerging as priority targets). Gender (92.6% men among regular donors) and socio-cultural (influence of Christian religions) disparities underscore the need for a segmented approach.

The developed predictive model (Random Forest, AUC = 0.93) confirmed the critical importance of **hemoglobin level** , height and weight in eligibility, providing a reliable tool to optimize recruitment. However, the limitations of the study (urban sample centered on Douala, class imbalance) call for caution in generalizing the results.

The **strategic recommendations** provide a concrete roadmap:

- **Decentralize collections** through mobile units and local partnerships in under-represented rural and urban areas.
- **Target regular donors** (educated, single or married men) with loyalty programs (SMS reminders, badges).
- **Convert high-potential populations** (academics, women) through appropriate campaigns (awareness-raising in academic circles, nutritional advice against anemia).
- **Digitize the process** (tracking platform, support chatbot) to strengthen operational efficiency.

In summary, this study demonstrates that the **alliance between data science and socio-cultural contextualization** can transform blood donation management in Cameroon. The insights drawn from the data, combined with targeted interventions on the identified obstacles (geographic, medical, behavioral), will help build a more **equitable, resilient and sustainable transfusion system** , save lives and meet the country's growing needs. The implementation of these recommendations, in partnership with local stakeholders (CNTS, churches, universities), constitutes a decisive step towards blood self-sufficiency.

Outlook : A future extension of this research could incorporate broader national data and qualitative surveys to refine the analysis of cultural barriers, while strengthening the predictive AI with additional contextual variables (access to transportation, seasonality of donations).

BIBLIOGRAPHY

BIBLIOGRAPHIE

1. **WHO (2021)** . *Guidelines on Blood Donor Selection and Blood Collection* . World Health Organization.
    - *Global reference on eligibility criteria and best collection practices.*
2. **Pindyck, RS, & Rubinfeld, DL (2018)** . *Econometrics* (5th ed.). Pearson.
    - *Statistical methods applied to hypothesis testing (ANOVA, Chi²).*
3. **James, G. et al. (2021)** . *An Introduction to Statistical Learning* (2 ^{e} ed .). Springer.
    - *Foundations of predictive models (Random Forest, K- means clustering).*
4. **CNTS Cameroon (2020)** . *Annual Report on Blood Transfusion* .
    - *Local data on blood needs and shortages.*
5. **Nguyen, T. et al. (2019)** . *"Machine Learning for Blood Donor Eligibility Prediction"* , Journal of Medical Systems, 43(7).
    - *Application of machine learning to donor sorting.*