

HEMOBOARD-DSWB Dashboard Report

Introduction and Project Overview

The project originated from a national challenge focused on improving blood donation systems through the use of data science and interactive dashboards. The core objective was to build a comprehensive, user-friendly, and insightful dashboard named HEMOBOARD for monitoring and analyzing trends in blood donation data. The target users included health professionals, campaign organizers, hospital staff, and policy decision-makers.

From the beginning, the project team was provided with an Excel file named Updated Challenge dataset.xlsx containing three sheets, two of which included redundant information about donors and donation candidates. The primary datasets extracted from these sheets were:

1. **"Candidat au don 2019 (avec anne)"** – which detailed candidate donors including demographics, clinical data, eligibility status, reasons for ineligibility, and prior donation history.
2. **"Donneurs 2019"** – which included actual blood donations, donation dates, blood types, and types of donations labeled as "F", "B", or blank for familial, bénévole (voluntary), or undefined respectively.

Recognizing the inconsistency and messiness in raw data, the very first step was an extensive data cleaning and standardization phase. This phase ensured uniform column names, resolved missing values, standardized categorical labels, and made all dates and numerical entries consistent for analysis. This cleaning was executed in Python using pandas, with additional use of re, unicodedata, and datetime formatting tools to correct string anomalies and convert date formats.

For instance, hemoglobin levels, which were inconsistently entered with units like "g/dl" or using commas instead of dots, were cleaned, normalized, and transformed into a float numeric format. Outliers and implausible values were removed, and strategic imputations were performed. Hemoglobin levels missing for women or men were filled using gender-specific and condition-specific averages, calculated based on whether they were marked as "indisponibles pour hémoglobine basse".

This preparatory work created two clean CSV files `Candidat_au_don_2019_cleaned.csv` and `Donneurs_2019_cleaned.csv` which became the foundation for building the data models and the interactive dashboard. The cleaned datasets were rich and multidimensional, allowing the team to explore several analytical perspectives including geospatial distribution, health eligibility criteria, donor retention, campaign effectiveness, sentiment analysis, and machine learning predictions.

1: Dashboard Architecture and Technical Stack

The dashboard was implemented using Python and the Dash framework by Plotly. Dash was chosen due to its simplicity, powerful data visualization capabilities, and seamless integration with pandas and machine learning models. To provide a polished user interface, the project integrated `dash_bootstrap_components`, which allowed for aesthetically appealing layouts using Bootstrap 4 themes.

The project was structured around a central controller file named `central_app.py`, which acted as the router and root container for all the dashboard pages. Each analytical component mapped donor distribution, health conditions, clustering, campaign effectiveness, donor retention, sentiment analysis, and eligibility prediction was developed as a separate Python script within an `objectives` folder. These scripts encapsulated their respective layouts and functionalities in isolated functions, ensuring modular design and maintainability.

To enhance navigability, a fixed vertical sidebar was placed on the left of the dashboard. This sidebar included a custom banner and a menu linking to all the analytical tabs. The rest of the page layout followed a two-column layout where the sidebar remained static while the main content dynamically updated based on user interaction with the menu or filters. A consistent styling theme was applied across the application through a dedicated `style.css` file located in the `assets` folder. This stylesheet defined padding, fonts, card aesthetics, legend blocks, and unified the look of the headers and sections.

The dashboard supported dynamic data filtering using date pickers that updated each module in real time. This filtering system ensured that users could explore data for specific periods, a crucial requirement for tracking donation patterns over time or evaluating the impact of specific campaigns.

The sidebar used Dash's `dcc.Location` and callback mechanisms to update the content pane (page-content) based on the selected path, and all pages were wrapped in a consistent layout

using Bootstrap rows and cards. KPIs were displayed using Dash Bootstrap Components' Card objects, presenting key metrics in visually appealing blocks with different color codes for quick interpretation.

2: Geographical and Health-Based Insights

The **Donor Mapping** module served as the first analytical component of the dashboard. Its objective was to visualize the spatial distribution of blood donor candidates across various neighborhoods and districts. It leveraged the address information specifically the `Arrondissement_de_résidence` and `Quartier_de_Résidence` from the cleaned candidate dataset. Since geographical coordinates were not originally included in the dataset, a simplified choropleth-style visualization was adopted using Plotly Express bar charts, displaying the number of donors per district. This visualization helped stakeholders identify which geographical zones contributed the most donors and where awareness campaigns could be intensified.

The tab included a `DatePickerRange` that filtered data by the date of form submission. This allowed users to observe changes in donation participation across neighborhoods over time. These filters dynamically triggered updates using Dash callbacks to redraw the map view accordingly.

The second major module, **Health Conditions**, focused on the eligibility status of donors based on clinical and biological metrics. From the cleaned data, a derived variable named `ÉLIGIBILITÉ_AU_DON` was used to separate eligible from ineligible donors. The ineligibility reasons were diverse and captured through multiple binary variables indicating whether the individual had conditions like low hemoglobin, recent illnesses, breastfeeding status, recent delivery, pregnancy, or infectious diseases.

To make these health exclusions interpretable, the dashboard presented a bar chart showing the frequency of each ineligibility factor. Additionally, visualizations grouped data by gender, age groups, and eligibility status to uncover disparities in health-based eligibility, such as the prevalence of low hemoglobin in women.

These two tabs geographical mapping and health conditions set the stage for the deeper analytical features.

3: Donor Profiling and Clustering

The **Donor Clustering** module was designed to create a visual and analytical profile of ideal blood donors using unsupervised learning, specifically the KMeans clustering algorithm. This approach allowed the system to identify natural groupings within the donor population without prior labels, based on six critical features: age, weight, height, hemoglobin level, gender (encoded), and donation history (binary).

Before clustering, missing values in height and weight were handled by replacing them with the population mean, ensuring the algorithm could operate on complete data. All variables were standardized using StandardScaler from scikit-learn to normalize units and magnitudes. The clustering was executed with a predefined value of $k=3$, chosen empirically to provide interpretable and distinct donor profiles.

Each cluster was visualized through 2D scatter plots (using PCA or t-SNE if necessary) to explore the separation between donor types. The interface also presented the cluster centers, translated into descriptive characteristics (e.g., "young males with high hemoglobin levels and prior donation history").

To enhance interactivity, the layout included a date filter to re-cluster donors dynamically based on time, enabling the monitoring of changes in donor profiles. Health institutions could thus better target recruitment efforts by identifying which donor clusters were more responsive over specific periods.

This machine learning integration was a key differentiator of the HEMOBOARD dashboard, bridging data science with public health needs in a usable interface. From unsupervised learning, the dashboard next moved toward behavioral analytics specifically, evaluating donor retention across time.

4: Donor Retention and Behavioral Insights

The **Donor Retention** module aimed to assess the continuity and commitment of blood donors, which is a critical factor in ensuring stable blood supply chains. This section relied primarily on the candidate dataset but derived key behavioral insights by examining the column `A-t-il_(elle)_déjà_donné_le_sang`. This field, when cleaned and converted into a binary variable, served as the foundation for defining donor loyalty.

The dashboard calculated the total number of candidates, their gender breakdown, and the number of loyal donors (those who reported having donated blood before). Another derived metric was the count of *eligible donors*, determined from the `ÉLIGIBILITÉ_AU_DON` field.

These indicators were summarized in a brief textual interpretation block to guide users on the meaning and impact of the figures.

Visualizations in this module included a time series of total versus eligible donors per month, a gender-disaggregated pie chart, and a bar chart showing donor loyalty by age group using custom bins (15–25, 26–35, etc.). To make trends more insightful, a linear trendline was added to the monthly bar chart of donation activity, helping users detect any gradual increases or decreases.

These figures were not only descriptive but actionable. They highlighted months or demographic groups with sharp declines or peaks in donation retention, allowing blood banks and health professionals to proactively plan engagement strategies, reminders, or incentives.

With behavioral trends analyzed, the dashboard transitioned to sentiment analysis to interpret donor feedback. The next page describes how Natural Language Processing (NLP) techniques were used to extract emotions from textual feedback.

5: Sentiment Analysis and Feedback Interpretation

The **Sentiment Analysis** tab brought a qualitative dimension to the dashboard by processing and interpreting donor feedback captured in the column `Si_autres_raison_préciser`. These open-text fields often contained valuable but unstructured information, offering insights into why some individuals declined to donate or how they perceived the donation experience.

The project employed the TextBlob library, a simple yet effective NLP tool, to calculate sentiment polarity scores for each feedback entry. Polarity values above +0.1 were classified as positive, values below -0.1 as negative, and values in between as neutral. This classification enabled a straightforward understanding of public sentiment toward the donation experience.

Two main visualizations were presented. The first was a bar chart showing the distribution of sentiments by profession, highlighting how different occupational groups responded to donation campaigns or reasons for non-participation. The second visualization was a pie chart summarizing the overall sentiment across all collected feedback.

To provide context, an interpretation block explained what the charts revealed. For example, negative sentiments among specific professions could signal the need for improved campaign communication, while a predominance of neutral feedback could suggest a lack of emotional engagement.

This sentiment layer complemented the quantitative analysis by amplifying the voice of donors and non-donors alike. Understanding not just *who* is donating, but *how they feel* about the process, significantly enhances campaign effectiveness and donor experience strategies.



The final module of the dashboard brought predictive power into play. Page 7 details the real-time eligibility prediction model integrated directly into the interface.

6: Predictive Modeling and Eligibility API

The most advanced feature of the HEMOBOARD dashboard was the **Eligibility Prediction** module. This section introduced a machine learning-powered tool designed to estimate, in real time, whether a new donor would be eligible based on their health and demographic characteristics. The model was trained using the cleaned candidate dataset and incorporated variables including age, weight, height, hemoglobin level, gender, and education level.

Prior to modeling, the system performed several preprocessing steps. Gender was encoded numerically (0 for male, 1 for female), and education levels were mapped into a dropdown menu for user-friendly selection. Missing values in height and weight were automatically imputed with the dataset mean, ensuring no interruption to the model's predictions. The target variable, eligibility, was binary and derived from the field `ÉLIGIBILITÉ_AU_DON.`, where "oui" was mapped to 1 and all other responses to 0.

The prediction model used was a `RandomForestClassifier`, trained after scaling the features with `StandardScaler`. Its performance was evaluated during development using cross-validation, and its feature importances were visualized in a horizontal bar chart, helping users understand which variables contributed most to the predictions.

The dashboard allowed users to manually enter new donor profiles through input fields. Once submitted, the system returned a prediction ("ÉLIGIBLE  " or "NON ÉLIGIBLE  ") along with a probability score displayed as both text and an intuitive gauge using Plotly's indicator chart. An interpretative message followed the result, helping non-technical users understand the prediction outcome in plain language.

This module simulated a real-world scenario where blood banks could instantly screen candidates using existing data, thereby improving operational efficiency and reducing administrative load.

Together, these seven modules formed an end-to-end digital tool for blood donation monitoring, strategy, and decision-making. The dashboard stands as a replicable model for health organizations aiming to integrate data science into community engagement and clinical logistics.