DATA PRE-PROCESSING TECHNIQUES.

The following techniques were used to pre-process the data;

- Removing the accents related to French language for easy coding i.e (à, â, ä, é, è, ê, ë, î, ï, ô, ö, ù, û, ü, ÿ, ç).
- Removing unnecessary spaces before and after words.
- Matching similar names by changing texts to sentence case , removing separating characters such as (-)and aligning typing differences. This was common in columns such as;
    - Profession (AGENT DE SECURITE and Agent de securite)
    - Quartier de residence e.g (Akwa-Nord and Akwa nord)
    - Religion e.g (Pentecostiste and Pentecotiste).
    - Arrondissement de residence (Ras for RAS, R A S). e.t.c.
- Aligning names in Quartier de residence to actual names that exists for the locations: e.g Bepanda for Bependa, Ndogpassi for Bdogpassi,Ndog-Passi, Ndobasssi, Ndogbong for Ndogbon, Ndogbom, e.t.c.
- Modifying year of birth to reasonable dates: 1980 for 0080,1989 for 2089,1992 for 2092, 1992 for 1882, 1993 for 0093 and replacing year 0000 and 0001 by 1988(average age, 31 years from 2019).
- Taux d'hemoglobine (Haemoglobin count); Missing values were replaced by the average of the given values, 13.4. Here we also removed measurement units like g/dl and replaced (,) by (.) where comma had been used as the separator.
- Poids (weight); Missing values were imputed randomly by taking the median of the available values, 85 as the maximum value and 50 as the minimum (informed by WHO lower weight requirements for blood donation).
- Age: the starting age was given as 18 years, thus all participants below this age were omitted from the analysis.
- Taille; This variable was exclusively used in building the model for predicting eligibility. The missing values were imputed by randomizing between the minimum value , 158 and maximum value, 192, that were provided in the dataset.

The initial dataset had **1915** observations. After applying the above mentioned techniques, the data used had **1886** observations.