Blood Donation Report

AZONGO Neilla Audrey, Adou MOUSSA, Ibrahim KHALIL-LAH

March 25, 2025

Le travail soumis a notre etude consiste a faire une analyse sur l'eligibilite des individus pour le don de sang. Nous avons a notre disposition un fichier excel constitue de trois feuilles de calculs et donc nous nous consacrerons sur le fichier avec les informations sur le la date de naissance.
Pour mener a bien notre travail Nous commencons par une explorations de ces donnees puis le pretraitement et ensuite nous avons appliquer des techniques d'analyses superviser pour la prediction et les techniques non superviser pour les differents profil que ce jeu de donnees englobe.

1 Exploration et pretraitement des donnees

Notre dataset est constitué de 1879 lignes et 39 colonnes. Ces colonnes sont en effet des informations demographic, des conditions de sante, et des informations geographiques; 36 variables (colonnes) categorielles et 3 variables quantitatives. Le nombres de valeurs manquantes etant tres eleves pour la plupart des variables

1.1 Exploration et nettoyage

Pour une mise a jour de notre dataset, Nous commencons par verifier les doublons pour les supprimer. Avec 36 doublons, le nombres total de lignes de notre dataset revient a 1879. Nous definissons egalement une fonction qui nous permettra de normaliser les noms de nos variables, en enlevant les caracteres accentues, en supprimant les espaces, ramenant les caracteres tous en minuscule.

Lors de la visualisation du type de chaque variable, nous remarquons que la variable du taux d'hemoglobine est considere comme objet a cause des caracteres comme **kg**, **g/dl** et **,**. Apres suppression nous obtenons le type *float*.

- **Normalisation :** Nous continuons par une normalisations des colonnes *profession*, *religion*, *Nationalite*, *religion*, *arrondissement de residence* et *quartiers de residence*. Nous remarquons dans ces colonnes que la meme valeur peut etre ecrite de plusieurs maniere differentes, nous les ajustons donc. Plus encore, nous remarquons pour la variable comme profession et religion, des valeurs qui vont dans les meme categories, plutot que d'avoir une multitude, nous les regroupons par groupe.

Toujours en ce qui concernes les variables nous remarquons un ensembles de colonnes de type bouleenne, qui a un nombre asse important de valeur manquantes, ceux sont les colonnes de raisons d'indisponibilite temporaire et des raisons de non eleigibilite definitive. Nous remarquons egalement un fais sur les raiponse des individus, tous ceux qui repondent **non** pour une raison d'eligibilite temporaire sont *eligible* ou definitivement non eligible. et tous les oui sont temporairement non eligible.Pour des raisons de non eligibilite definitive, seulement ceux definitivement non eligible ont repondu **oui**. (NB: Tous temporairement non eligible n'ont pas donne de reponse a l'indisponibilite et tous les definitivement non eligible n'ont pas repondu a la non eligibilite definitive). A cause de cette raison, nous ne pourions repondre oui ou non chez ces individu qui devrait forcement avoir un probleme. Pour mieux gere cela, nous avons ramener toutes ces deux groupe de colonnes en deux seules. Pour y parvenir, nous avons recupere les reponses *oui* et avons changer a la variable correspondante de telle enseigne qu'il etai facile de savoir la cause de non eligibilite temporaire/ non eligibilite definitive de l'individu, et pour les individu n'ayant pas repondu, il etait facile de precise que cette personne a une autres indisponibilite/raison de non eligibilite. Tous les individus avait donc pour raison *non applicable*.

- **Datetime :** Pour ce qui concerne les differentes date, notamment la date de remplissage, l'annee de naissance, la date de derniere regles et la date de dernier dons, nous remarquons que le type de ces dates n'est pas datetime car, nous remarquons des format d'erreur differents.

1. **Date de remplissage :** Cette date contient deux valeurs manquantes que l'on parvient a remplacer facilement a travers les dates voisines, et les autres colonnes comme annee de naissance et niveau scolaire profession. En plus de cela il contient des date erronee qui sont parfois egale a la date de naissance de l'individu, ou une annee 1988, etc. et 2020 aussi. Pour ces donner nous les repechons apres avoir remplacer la date en datetime. Ces format d'annee non correspondant se retrouve enlever

du data (20 exactement), avec les deux valeurs manquantes de depart on revient a 22. Nous ecrivons donc le code qui change pour les dates lontaines comme preciser ci-dessus, l'annee en 2019 a travers une fonction **lambda** et remplace les valeur manquantes par le 12 Decembre 2019.

2. **Date de naissance :** Cette date egalement montre plusieurs differents format par exemple les annees 0093, 0001, etc. Pour rendre le format uniforme, nous le transformons en date time et utilisons une fonction qui arrange les annees en remplacant ajoutant 1900 ou 2000 a l'annee de naissance.

3. **Date de dernier don :** Nous avons 781 valeur contre 800 qui ont repondu **oui**, a deja fait le don; seulement, parmi ces 800 oui, 28 n'ont pas donne de date. Cela dit en se basant rien que sur les **oui**, il y aurait 772 valeur non nulles. D'ou vient le surplus de 9?, Nous remarquons dans le data que tous ceux qui ont repondu non (1070 valeurs manquantes) n'ont pas tous des valeurs manquants ($1070 + 800$ is not 1879), alors pour ces 9 individus, nous changons leur non en oui. Nous transformons la variable en datetime., nous continuons a arranger les dates avec des annees erronees.

4. **Date de derniere regles :** Cette variable concernant uniquement les femmes (188femmes contre 1691 hommes) est deja en datetime, elle ne contient que 39 valeurs non manquantes.

Apres avoir fait ces explorations,nous utilisons l'information de date de remplissage pour creer trois nouvelle variables notamment l'age, nombre de jours depuis les dernieres regles, nombre de mois depuis le dernier don. Pour obtenir l'age, nous faisons une soustraction que nous divisons par 365 pour obtenir le resultat en annee, pour la date de dernier don une division de 30 pour obtenir le resultat en mois. La date de derniere regle ne sera pas divise. Pour evitedes problemes pour ces variables nous mettons apres soustraction avec la date de remplissage le resultat en dt.days, et a la fin de chaque calcul nous arrondissont pour obtenir des valeurs numeriques entieres. Ainsi, pour emputer les valeurs manquantes nous remplacerons pour tous les individus n'ayant pas d'entree, un nombre de mois alleatoire mais superieur a 3 mois pour que cela n'impacte pas sur l'eligibilite temporaire.

- **Gestion des valeurs manquantes :** Commencons par les dates de dernieres regles et e derniers dons.

- Pour les 28 valeur manquantes des individus ayant repondu oui pour avoir deja fait le don sont essentiellement des personnes eligibles d'apres notre exploration.

- La date de derniere regle ne concernant que la femme, nous remarquons que ce sont les 3 categories d'heligibilite chez les femmes qui n'ont pas donne leur date de derniere regle. Donc 105 eligibles, 23 definitivement non eligible et 21 temporairement non eligible, soit 149 valeur manquantes reelles pour celle-ci. Donc pour les femmes eligible nous avons emputer les valeurs manquantes par une valeur superieur a 14 jours. Pour les temporaires et les definitive, nous avons remplacer par une valeur inferieur a 14 bien que d'autres parametres peuvent entrainer la non eligibilite temporaire (chez les femmes temporairement eligible) et ou la non eligibilite definitive).

Enfin, pour tous les hommes, nous remplacons la date de derniere regle par une valeur aleatoire superieur a 14 jours. et tous les individus ayant repondu non pour le don de sang, les valeurs manquantes du nombre de mois depuis le dernier don devront etre remplacer par une valeur de 0 normalement, mais ceci pourrait impacter car $0 < 3$ donc nous avons ete oblige de prendre une valeur aleatoire superieur a 3 mois.

- Sur **le taux d'hemoglobine,** nous n'avons que 1040 valeur non manquantes, parmi lesquelles 150 eligibilite temporaire, 838 eligibles et 50 definitivement non eligibles. Soit 742 valeurs manquantes pour les eligibles, 37 pour les definitivement non eligibles et 60 pour les temporaires. Nous emputons cettes variables par categories par la valeur meadiane de chacune des trois categories pour chacune d'elle (de par la distribution normal normal).

- Pour les nouvelles variables creer sur les raisons d'indisponibilite temporaire et les raisons de non eligibilite definitive, Pour la colonne indisponibilite, tous les individus eligibles et definitivement non eligibles auront pour valeur *non applicable*, de meme, la colonne *non eligibilite totale*, les individus eligibles et temporaires auront pour valeur *non applicable*. Cependant on remarque toujours les valeurs manquantes qui ont ete traite comme explique ci-dessus, en renplacant par autres raisons.

- Les variables **Taille et poids** ont des pourcentage de valeur manquantes tres eleves (plus de 95%). Pour leur imputation, nous avons commencer par imputer le poids. Nous avons renvoyer des valeurs aleatoires choisi entre 60kg et 120 kg pour les individus eligibles et une repartition entre 40kg et 59kg pour les autres individus. En ce qui concerne la taille, nous la repartissons entre 155 m et 192 m (les valeurs min et max), puis nous calculons leur imc entre 18.5 et 30 (indice de masse corporelle). Si l'imc d'un individu est inferieur a 18.5, sa taille sera modifier en ramenant son imc a 18.5, de meme, si l'imc est superieur a 30, la taille de l'individu sera modifier en ramenant son imc a 30. ces tailles sont donc modifier suivant l'operation:

$$t = \sqrt{\frac{p}{18.5} - \frac{t^2 \times (18.5 - imc)}{18.5}}, \qquad t = \sqrt{\frac{p}{30} + \frac{t^2 \times (imc - 30)}{30}}. \tag{1}$$

2 Pretraitement :

Pour une meilleure analyse des donnees, nous encodons les valeurs categorielles, celles comme profession qui ont plusieurs valeurs, ces valeurs ont ete rougroupe en different groupe et nous obtenons desormais un ensemble reduis de reponses. En plus de l'encodage des valeur categorielles, nous standardisons les valeurs numeriques pour les avons a la meme echelle.

3 Analyse

3.1 Analyse predictive :

Nous considerons pour les variables X, le niveau d'etude, le genre, la profession, taille, poids, situation matrimoniale, nationalite, religion, a-t-il deja donne le sang, taux d'hemoglobine, age, nombre de mois depuis dernier don, nombre de ours depuid dernieres regles.

Pour les modeles, nous choisissons les algorithmes tels que **Random Forest**, **Decision Tree**, **Logistic Regression**, **Naive Bayes**, **KNN**, **SVM**, et le **Gradient Boosting**. Nous choisissons premierement de diviser notre dataset en 80% pour l'entrainement et 20% pour la prediction. Dans un deuxieme temps nous appliquons le **K-fold cross validation** avec un nombre de $k = 5$. Les resultats obtenus sont presentes dans le tableau ci-dessous:

Models	**Train/test** accuracy	**K-fold CV** accuracy
Random Forest	0.94	0.93
Decision Tree	0.90	0.91
Logistic Regression	0.93	0.91
Naive Bayes	0.85	0.82
KNN	0.86	0.86
SVM	0.91	0.90
Gradient Boosting	0.94	0.94

3.2 Clustering technique

Pour detecter les differents types de profils d'individus qui participent au don de sang, nous appliquons la methodes de distance qui est le clustering. Nous commencons par appliquer la technique du dendogramme. A travers la visualisation du dendogramme, nous parvenons a identifier 3 groupe de profil. Nous utilisons ce nombre 3 pour obtimiser la methode du coude pour une meilleur observation a travers la techniques K-means.

4 Observation et remarque perspective:

Les perspective s'ettende au fur et a mesure a l'etape de pretraitement, pour les variables qui concernes uniquement les femmes, il serait minitieux de gerer deux modeles pour eviter une quelconque impact dans

le resultat, a tel enseigne que dans le dashbord, l'utilisateur devrait dabord entrer son genre et un modele sera applique en fonction de cela. Une autre observation est faite sur le record des dates, il sera judicieux d'utiliser un format automatique pour eviter les different erreurs lie au format de remplissage. Connaitre l'imc de l'individu. Pour les noms des quartiers egalement une strategie devrait etre elaborer pour pouvoir avoir les differentes localites selon les coordonnees geographiques (latitude, longitude).