# DEMARCHE DE TRAVAIL

## OBJECTIF MESURABLE

\*\*OBJECTIF\*\*

Prédire si une personne est atteinte d'un accident vasculaire cérébral ou pas à partir des données personnelles et cliniques fournies (problème de classification). Il est essentiel de noter qu'il est plus urgent, dans ce cas de figure, de détecter toutes les personnes malades (et non ceux qui sont vraiment malades parmi ceux qui sont identifiés comme malade). Sinon un patient malade peut être diagnostiquer comme sain alors que ce n'est pas le cas. Nous préférerons ainsi la sensibilité/spécificité à la précision.<hr size = "2">

\*\*METRIQUE ET SCORE A ATTEINDRE\*\*

La proportion de la classe positive est largement inférieure à la proportion de la classe négative (nous notons une forte déséquilibre de classes). Dans ce cas de figure, la métrique accuracy ne sera pas assez performante pour évaluer notre modèle de classification. A la place, nous allons utiliser les métriques sensibilité et précision pour valider notre modèle :

- Sensibilité : $\frac{VP}{VP+FN}$. Elle permet de calculer le pourcentage de tests positifs parmi les patients réellement atteints d'AVC ;

- Spécificité : $\frac{VN}{VN+FP}$. Elle permet de calculer le pourcentage de tests négatifs parmi les patients réellement sains. C'est la sensibilité pour les test négatifs ;

- Précision : $\frac{VP}{VP+FP}$. Elle permet de calculer le pourcentage de patients réellement malades parmi ceux dont le test est positif ;

- F1-score : $\frac{2\times Sensibilite \times Precision}{Sensibilite+Precision}$ ;

-

Avec :

- $VP$ : Nombre de Vrais Positifs ;

- $VN$ : Nombre de Vrais Négatifs ;

- $FN$ : Nombre de Faux Négatifs ;

- $FP$ : Nombre de Faux Positifs ;

Nous nous fixons une sensibilité/spécificité supérieure à \*\*80%\*\*. Nous pouvons également utiliser des outils supplémentaires comme \*\*les courbes ROC et Precision-Recall\*\* pour affiner nos choix de performance.

## EXPLORATION DES DONNEES

### ANALYSE DE LA FORME

\*\*CIBLE\*\*

La variable dépendante est la variable \*\*stroke\*\* qui contient des données discrètes. Cette variable indique si une personne est atteinte d'un accident vasculaire cérébral (1 pour atteinte) ou pas (0 pour non atteinte). Nous allons abréger, par la suite, accident vasculaire cérébral par AVC pour faciliter l'écrit. Nous devons transformer, plus tard, la variable cible en variable catégorielle non ordinale (cela nous permettra de différencier la catégorie "être atteint(e) d'AVC" de la catégorie "être non atteint(e) d'AVC"). Nous devrons ainsi utiliser un modèle de classification (classification model) pour déterminer si un patient est malade ou sain.

\*\*NOMBRE DE LIGNES ET DE COLONNES\*\*

Nous avons identifié 5110 observations et 12 variables (dont la cible). Le nombre d'observations est inférieur à 10000, donc notre dataset ne contient pas énormément d'observations mais assez pour entraîner un modèle.

\*\*TYPES DE VARIABLES\*\*

Parmi les variables explicatives, nous avons identifié 7 variables catégorielles (dont deux contiennent des valeurs discrètes et les autres contiennent des valeurs de type chaîne de caractères) et 3 variables non catégorielles.

- Les variables de type object ont pour valeurs possibles les suivantes :

- \*\*gender\*\* (genre) : contient les valeurs \*Male\* (Homme), \*Female\* (Femelle) ou \*Other\* (ni Homme, ni Femme) ;

- \*\*ever\_married\*\* (jamais marié(e)) : peut prendre les valeurs \*Yes\* (Oui) ou \*No\* (Non) ;

- \*\*work\_type\*\* (type de travail) : peut prendre les valeurs \*Private\* (Privée), \*Self\_employed\* (Auto emploi ou Travail autonome), \*Govt\_job\* (Travail au gouvernement), \*children\* (enfant), \*Never\_worked\* (N'a jamais travaillé(e)) ;

- \*\*Residence\_type\*\* (type de résidence) : peut prendre les valeurs \*Urban\* (Urbaine), \*Rural\* (Rurale) ;

- \*\*smoking\_status\*\* (statut de fumeur) : contient les valeurs \*formerly\_smoked\* (a fumé(e) dans le passé), \*never\_smoke\* (n'a jamais fumé), \*smokes\* (fume), \*Unknow\* (donnée non recueillie).

- Quant aux variables catégorielles \*\*hypertension\*\* et \*\*heart\_disease\*\* (maladie de coeur), elles peuvent prendre les valeurs 1 (pour \*positive\*) ou 0 (pour \*négative\*).

- Nous identifions une variable discrète non categorielle (\*\*age\*\*) dont les valeurs varient de 0.08 (enfant de 9 mois) à 82 (adulte de 82 ans). Sa valeur moyenne est de 43 (adulte de 43 ans).

-

Le dataset ne contient que deux variables continues, \*\*avg\_glucose\_level\*\* (niveau moyen de glucose dans le sang) et \*\*bmi\*\* (indice de masse corporelle). Ces deux variables ne s'expriment pas avec les mêmes unités :

- La variable avg\_glucose\_level a pour valeur maximale 271.74 et pour valeur minimale 55.12. Sa valeur moyenne est de 106.15.

- La variable bmi a pour valeur maximale 97.60 et pour valeur minimale 10.30. Sa valeur moyenne est 28.89.

- Nous définirons plus en détail ces variables dans la partie analyse du fond.

\*\*IDENTIFICATION DES VALEURS MANQUANTES\*\*

Nous remarquons que seule la variable bmi contient des valeurs manquantes qui sont un peu dispersées dans le dataset (Un peu entassées au niveau des premières observations). Mais en sachant que le nombre de valeurs manquantes ne représente que 16% dans l'ensemble des données présentes dans la colonne bmi et que nous allons choisir un échantillon aléatoire pour l'entraînement du modèle (ainsi que pour l'évaluation) donc on peut supposer que le remplacement des valeurs manquantes par la valeur la plus fréquente dans la variable bmi constitue une bonne stratégie.

\*\*IDENTIFICATION DES VALEURS REDONDANTES\*\*

Les données ne contiennent pas d'observations dupliquées.

### ANALYSE DU FOND

\*\*VISUALISATION DE LA CIBLE\*\*

Nous remarquons que 95.13% des patients sont atteintes d'AVC contre seulement 4.87% des patients atteintes d'AVC. La proportion de patients non malades est largement supérieure au nombre de patients non malades.

\*\*COMPREHENSION DES DIFFERENTES VARIABLES\*\*

- id : La variable id contient des valeurs discrètes et identifie chaque observation de manière unique. Elle n'apporte aucune information supplémentaire et donc n'influence pas le fait qu'une personne soit malade ou non. La variable id doit être supprimée.

- Variables catégorielles :

- Hypertension : Cette variable indique si oui ou non, le patient souffre d'hypertension. Un patient est testé positif à l'hypertension si on constate à deux reprises, et pas dans le même jour, une tension systolique supérieure ou égale à 140 mm Hg et/ou une tension diastolique supérieure ou égale à 90 mm Hg. Selon les tests cliniques, une hypertension peut souvent être la cause d'AVC. Nous verrons par la suite si les données recueillies de cette variable sont fiables. La colonne hypertension contient 90% de tests négatifs et 10% de tests positifs.

- heart\_disease : La variable heart\_disease indique si le patient est atteint de cardiopathie (maladie cardiaque) ou pas. Il y a différents types de cardiopathies et nous verrons si ces derniers peuvent influencer le risque d'attraper un AVC. La colonne heart\_disease contient 95% de tests positifs contre seulement 5% de tests négatifs. Cependant, nous savons que l'hypertension non traitée peut causer la cardiopathie. Donc on peut supposer, d'ores et déjà, que ces deux variables sont fortement corrélées (hypothèse à vérifier).

- gender : La variable gender indique à quel sexe appartient le patient. On doit vérifier si le sexe du patient peut influencer le risque qu'il soit atteint d'AVC. La colonne gender est composée de 58% de femmes, de 41% d'hommes et très peu (presque 0%) de sexe de type autre.

- ever\_married : Cette variable indique si le patient a déjà été marié. L'analyse de cette variable doit nous permettre de dire si le patient a plus de chance d'attraper un AVC en s'étant déjà marié (dans le temps présent ou passé) ou pas. Nous notons 66% des patients qui se sont jamais mariés et 34% qui se sont déjà mariés.

- work\_type : Elle indique le type de travail effectué par un patient. On identifie 57% de patients travaillant dans le secteur privé (professions et secteurs d'activité ne dépendant pas de l'Etat), 16% des patients effectuant du travail autonome (ils sont leurs propres employés), 13% des patients sont des enfants (on n'a pas plus d'informations sur la catégorie children mais on suppose pour l'instant que le patient est un enfant et donc qu'il n'a pas besoin de travailler), presque 13% des patients travaillent pour le gouvernement (dans le secteur public) et seulement 0.4% n'ont jamais travaillé. Pour les patients dont le type de travail est children, nous devons vérifier si leurs âges indiquent que ce sont des enfants ou pas (c'est-à-dire certains sont des adultes).

- residence\_type : Cette variable indique le type de résidence du patient. 51% des patients résident dans un milieu urbain (ville) et 49% des patients résident dans un milieu rural (campagne). Les proportions de ces deux classes sont presque similaires.

- smoking\_status : Elle indique si le patient est/était un fumeur ou pas. 37% des patients n'ont jamais fumé, 30% des patients n'indiquent pas s'ils fument, 17% des patients ont fumé auparavant et 15% des patients disent fumer au moment où on les interrogeait. La catégorie unknow (inconnue) peut constituer un problème car elle n'apporte aucune information utile. Nous vérifierons par la suite si cette variable apporte de l'information à notre modèle.

- Variables non catégorielles :

- age : Nous constatons que les patients qui sont âgés entre 37 et 63 ans sont plus nombreux dans la base de données, suivis des patients qui sont âgés entre 78 et 82 ans tandis que les plus jeunes sont les moins nombreux.

- avg\_glucose\_level : Le niveau moyen de glucose dans le sang indique si, oui ou non, le patient est atteint de diabète. Cette variable s'exprime en mg/dL (milligrammes par décilitre). Les patients qui souffrent de diabète ont un niveau moyen de glucose supérieur ou égale à 200 mg/dL. Par contre, un niveau moyen de glucose, inférieur à 140 mg/dL, est considéré comme normal. Nous constatons que la plupart des patients n'ont pas de diabète car une grande partie des patients ont un niveau moyen de glucose tournant autour de 75-87 mg/dL de sang. Nous remarquons qu'environ 10 à 20% des patients ont un niveau moyen de glucose tournant autour de 210-225 mg/dL. On peut considérer que ces derniers sont atteints de diabète.

- bmi : L'indice de masse corporelle (IMC) permet d'indiquer la corpulence d'un patient. Il s'exprime en kg/m2 (kilogrammes par mètre carré). Une personne est considérée comme obèse si son IMC dépasse 30 kg/m2. La plupart des patients examinés ont un IMC situé autour de 29 kg/m2. Ces derniers présentent un cas de surpoids ou d'obésité modérée (dont les IMC sont respectivement situés entre 25 et 30 kg/m2 et entre 30 et 35 kg/m2).

\*\*ETUDE PLUS POUSSEE DES VARIABLES AVEC PANDAS-PROFILING\*\* ##

\*\*VISUALISATION DES RELATIONS ENTRE LES VARIABLES EXPLICATIVES ET LA CIBLE \*\*

- Relations Cible - variables catégorielles :

- Pour la variable work\_type : Nous remarquons que les patients qui ont pour type de travail children ont plus de chance de ne pas attraper d'AVC que les patients qui effectuent d'autres types de travail. On a peu de patients non travailleurs dans la variable work\_type donc on ne peut rien dire par rapport à cette catégorie (mais il est possible qu'elle influence aussi le fait qu'on soit atteint d'AVC ou pas).

- Pour la variable smoking\_type : Nous remarquons que les patients ne disant pas s'ils fument/fumaient présentent une proportion de malades inconsistante par rapport aux autres types de fumeurs. Cela est dû au fait que la catégorie unknow n'est pas fiable (elle se comporte comme une valeur manquante).

- Relations Cible - Variables non catégorielles :

- Parmi les variables non catégorielles, seule la variable âge influence le fait qu'un patient soit atteint d'AVC ou pas. Il y a plus de risque d'attraper un AVC chez les patients plus âgés ;

- Pour les deux autres variables restantes nous constatons que les distributions sont presque les mêmes pour les patients malades et non malades ;

- Nous vérifierons plus amplement ces hypothèses à travers un test de student.

\*\*VISUALISATION DES RELATIONS ENTRE VARIABLES QUANTITATIVES\*\*

Les variables quantitatives ne partagent aucune forte corrélation entre elles.

\*\*VISUALISATION DES RELATIONS ENTRE VARIABLES CATEGORIELLES ET QUANTITATIVES\*\*

Nous constatons que quelques variables catégorielles ont de fortes corrélations avec des variables quantitatives.

- La variable \*\*age\*\* est fortement corrélée avec les variables \*\*work\_type\*\*, \*\*smoking\_status\*\* et \*\*ever\_married\*\*. Pour sa relation avec les autres variables catégorielles nous remarquons de légères corrélations ou quasiment pas de corrélations pour certaines.

- La variable \*\*bmi\*\* est, pour son cas, fortement corrélée avec les variables \*\*ever\_married\*\* et \*\*work\_type\*\*.

- En revenant au niveau de pandas profiling, nous constatons que nos analyses sont soutenues par les remarques faites par la librairie (aucune incohérence n'est à noter).

\*\*IDENTIFICATION DES VALEURS ABERRANTES\*\*

Nous constatons que quelques valeurs dépassent la limite et peuvent être considérées comme aberrantes pour le moment.

- La variable age ne comporte presque pas de valeurs aberrantes.

- Les variables bmi et avg\_glucose\_level contiennent par contre des valeurs aberrantes.

- La variable avg\_glucose\_level ne contient des valeurs aberrantes que du côté de la classe \*test negatif\* de la variable stroke alors que la variable bmi en comporte pour les deux classes (mais beaucoup plus du côté de la classe \*test negatif\*).

- Cependant pour la variable avg\_glucose\_level nous constatons que ces valeurs coincident avec les valeurs désignant le groupe de patients atteints de diabète. Donc nous devrons traiter ces valeurs avec précaution car elles peuvent être utiles à l'identification de patients atteints d'AVC.

\*\*PREMIERE CONCLUSION\*\*

- On note une forte désiquilibre de classes au niveau de la cible. Le pourcentage de patients non atteints dépasse les \*\*90%\*\*. Ce problème peut-être traité avec l'équilibrage des classes à l'aide de quelques techniques de rééchantillonnage (sous échantillonnage, sur échantillonnage ou les deux).

- On constate une dépendance entre la variable cible et la variable age.

- La variable id ne comporte aucune information utile donc il est préférable de la supprimer de la base de données.

- La base de données contient des valeurs manquantes que nous allons remplacer par la valeur la plus fréquente de la variable bmi ;

- La catégorie unknow de la variable smoking\_status n'apporte aucune information supplémentaire donc nous devons la remplacer par une catégorie plus intéressante (\*never smoked\* par exemple) ;

- Les variables quantitatives ne sont pas corrélées entre elles ;

- La variable age est fortement corrélée avec certaines variables catégorielles. Notamment les variables work\_type, ever\_married et smoking\_status ;

- La variable bmi est aussi fortement corrélée avec les variables catégorielles ever\_married et work\_type ;

- Nous devons garder les variables age et bmi et supprimer les variables smoking\_status, work\_type et ever\_married ;

- Des tests supplémentaires seront réalisé pour étayer certaines hypothèses ou les rejeter ;

- Les variables quantitatives comportent des valeurs 'anormales'. Nous allons vérifier s'il est nécessaire de les remplacer ou de les supprimer. Ces données peuvent être utiles car ils caractérisent un groupe de patients.

- Nous allons retenir pour l'instant les variables hypertension et heart\_disease qui sont cliniquement très intéressantes pour nos analyses.

\*\*TESTS D'HYPOTHESES\*\*

Ces tests nous permettrons de valider ou de rejeter certaines hypothèses.

- Testons si les variables quantitatives influencent le fait qu'un patient soit atteint d'AVC ou pas (nous noterons H0 l'hypothèse selon laquelle une variable n'a pas d'influence sur la cible) :

- Le test de student nous indique que les variables \*\*age\*\* et \*\*avg\_glucose\_level influencent le fait qu'un patient soit atteint d'AVC ou pas\*\*.

- La variable \*\*bmi n'entretient pas de dépendance avec la variable stroke\*\*.

\*\*SECONDE CONCLUSION\*\*

- Les variables \*\*age\*\* et \*\*avg\_glucose\_level\*\* apportent de l'information contrairement à la variable bmi qui, elle, peut-être supprimée.

- Les variables à supprimer sont donc : \*\*bmi\*\*, \*\*work\_type\*\*, \*\*ever\_married\*\*, \*\*smoking\_status\*\*, \*\*id\*\*. Pour le moment nous ne supprimerons que ces variables. Nous ferons une sélection antérieure de variables si on constate un sur entraînement du modèle.

</div>