université PARIS-SACLAY | FACULTÉ DES SCIENCES D'ORSAY

Détection de chute à l'aide d'un capteur de sol intelligent

Tia ZOUEIN

En vue de l'obtention du
Master 1 : Mathématiques Appliquées

Le 25 Août 2022

Encadré par
Madame Christine KERIBIN
Maître de conférences à l'Université Paris-Saclay,
Responsable du Master2 Data Sciences et Master1 Mathématiques Appliquées à l'Université Paris-Saclay

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Les chutes constituent un problème de santé publique. Les personnes âgées sont particulièrement exposées au risque de chute. Dans certains cas, elles peuvent être fatales. De plus, même sans blessure grave, elles peuvent déclencher chez la personne une peur de tomber à nouveau. Ce qui entraîne alors une réduction des interactions et des activités sociales. Pour cela, le secteur de la santé recommande une surveillance de ces personnes.

Les études de l'organisation mondiale de la santé affirment que le nombre de chutes mortelles s'élève à 684 000 par année. Ainsi, c'est la deuxième cause de décès accidentiel pour les personnes de plus de 60 ans [2]. 37,3 millions de chutes nécessitent des soins médicaux chaque année. Les coûts financiers des blessures sont importants. L'âge et l'état de santé de l'individu peuvent influer la gravité de la blessure. Les données provenant du Canada suggèrent que la mise en œuvre de stratégies de prévention efficaces avec une réduction de 20 % de l'incidence des chutes pourrait créer une économie nette de plus de 120 millions de dollars américains chaque année.

En france, les chutes entraînent 12 000 décès [3]. 40 % des personnes âgées hospitalisées après une chute ne peuvent pas retourner à domicile et doivent être accueillies en établissement. Il est prévu que le nombre de personnes de plus de 65 ans augmentera de 2,4 millions d'ici à 2030. Le coût pour couvrir les chutes s'élève à 2 milliards d'euros dont 1,5 milliard pour la seule Assurance Maladie. Pour faire face à ce problème, le ministre des Solidarités et de la Santé, Olivier Véran, et le ministre déléguée chargée de l'Autonomie, Brigitte Bourguignon, ont lancé un plan national dont le but est de réduire de 20 % des chutes mortelles [3]. Les principaux objectifs sont les suivants :

— Repérer les personnes qui risquent de tomber.

— Aménager le logement (bon éclairage, tapis non glissantes...).

— Activité physique : la ministre des Sports, Brigitte Bourguignon a décidé de favoriser le recours à l'activité physique adaptée pour les personnes âgées.

Dans ce but, il existe des systèmes de détection de chute rapides et fiables qui améliorent les chances de survivre à l'accident et de faire face à ses conséquences physiques et psychologiques. On cite : les appareils portables, les systèmes basés sur des capteurs d'ambiance et des caméras. Malgré leur grande précision concernant la détection des chutes, ces derniers souffrent d'importants inconvénients. En fait, les appareils portables doivent toujours être attachés à la personne, ce qui en fait un mauvais choix pour leur soin car leurs porteurs peuvent oublier ou simplement refuser de les porter. D'autre part, les caméras ont l'inconvénient d'être très intrusives dans la vie du patient, ce qui entraîne des problèmes de confidentialité.

Récemment, en octobre 2021, Bruno Duperrier, un entrepreneur originaire de Limoges - France, a développé un dispositif innovant de détection des chutes [4]. Il n'a pas besoin d'être porté ni d'une caméra. Il se présente sous la forme d'une colonne blanche, installée au mur. Le système détecte les mouvements et permet d'alerter en cas de chute. Il peut aussi vérifier l'état de santé de la personne, comme la fréquence respiratoire ou d'autres signes annonciateurs d'un changement de comportement. Son innovation est très demandée, y compris à l'étranger : Japon, Canada, Etats-Unis, et un certain nombre de pays européens. Sa fabrication est totalement faite en France.

Le principal avantage de la détection de chute, c'est qu'il permet d'assurer des services d'urgences dans les plus brefs délais et de manière automatique.

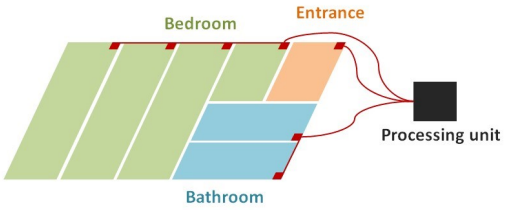

FIGURE 2 – Exemple schématique du capteur [1]

Pour cela, j'aborde dans mon TER le sujet de détection d'événements dans les signaux temporels pour le suivi des personnes âgées grâce à l'utilisation d'un capteur installé directement sous le sol. Il est fabriqué par l'entreprise française Tarkett [5], dont l'activité principale est la production de revêtements de sol. Ce système de détection est conçu en premier lieu pour être installé dans des maisons de retraites [6], avec un potentiel développement vers l'utilisation dans les foyers particuliers. Plusieurs scientifiques ont poussé leur recherche sur ce moyen de détection de chute, et récemment par le *Data Scientist* Ludovic Minvielle qui a rédigé sa thèse sur ce sujet [1] et sur laquelle je me base. Il s'agit d'un capteur de sol sensible, associé à une unité de traitement, discret, qui n'a pas besoin que le patient porte quoi que ce soit et qui peut être mis en place sur une grande surface avec une facilité d'installation par des bandes. Il est fabriqué à partir d'un polymère piézoélectrique, un matériau qui émet un champ lorsqu'il est soumis à une pression. Donc, le capteur n'a pas besoin d'alimentation pour générer un signal. Dès qu'une anomalie est enregistrée, le personnel sera notifié pour permettre une intervention rapide [5].

Le capteur s'installe comme des bandes de largeur fixe (60 cm) et de longueur variable, permettant ainsi de couvrir assez facilement différentes surfaces [1]. Il a été installé dans des chambres individuelles d'une maison de retraite, dans trois zones : l'entrée, la salle de bain et la chambre.

La figure 2 représente ces zones. La chambre est dans ce cas équipée de quatre bandes. Chaque bande émet son propre signal qui sera pré-traité individuellement (filtration, élimination de sa tendance linéaire et convertion de la charge en tension) par l'unité de traitement. Après avoir été prétraités, tous les signaux sont additionnés génèrant ainsi un seul signal pour cette zone [1]. Le signal est alors une des trois entrées qui alimentent l'unité de traitement local. Lorsqu'il est acheminé à l'unité, le signal est d'abord passé à travers un amplificateur de charge analogique qui convertit la charge du signal en tension (entre 0 volt et 3.3 volt), puis en séries chronologiques numériques. Chaque canal d'entrée est relié à quelques rouleaux de capteurs qui couvrent une zone spécifique.

L'objectif de mon travail est d'appliquer des algorithmes de Machine Learning sur mon jeu de données, dans le but de faire des classifications et des prédictions. La première partie consiste à présenter les données. Puis, j'applique des méthodes en apprentissage non supervisé (Analyse en Composantes Principales (ACP) et le Clustering) pour faire une classification des données. Ensuite, j'applique des méthodes en apprentissage supervisé dans le but de faire des prédictions de chute.

2 Présentation des données

On s'intéresse dans cette partie à présenter le jeu de données sur lequel on appliquera des algorithmes de classification.

Le jeu de données a été collecté avec des capteurs de sol conçus par Tarkett. Le système de capteurs de sol de Tarkett a été installé dans des chambres individuelles dans une maison de retraite (voir la figure 2 pour une installation schématique). Les signaux collectés ont été divisés sur des périodes d'environ une heure et étudiés avec un outil algorithmique et de visualisation de Tarkett pour trouver des événements de chute. Les chutes ont lieu dans différentes zones des chambres surveillées. Le début et la fin d'une chute, ainsi que sa durée, sont inconnus. En plus des chutes, un large éventail d'activités quotidiennes de personnes âgées ont également été enregistrés (mais non étiquetées). Les séries chronologiques fournies mesurent la tension des capteurs piézoélectriques.

Le jeu de données est déséquilibré :7.13% des cas sont de chute, ce qui peut causer un problème lorsque l'on souhaite mettre en place une méthode de classification. En fait, un modèle d'apprentissage fonctionne mieux si les proportions des classes sont proches.

D'autre part, deux autres transformations du signal ont été prises en considération :

— Sa dérivée première, car les variations des signaux et leur vitesse peuvent avoir une influence sur la détection des évènements.
— la transformée de Fourier, en supposant que la réponse en fréquence entre une chute et d'autres événements est différente.

Dans le but de transformer le signal pré-traité en un vecteur, il est d'abord transformé en un vecteur de caractéristiques statistiques, calculées sur une fenêtre de taille fixe du signal T. En fait, pour la détection de chute, les signaux à traiter étant assez complexes, les méthodes ont tendance à utiliser des extraction, ou au moins quelques transformations du signal de sortie. Au lieu de traiter donc le signal lui-même, il sera représenter par ses caractéristiques statistiques. Elles peuvent varier d'une expérience à l'autre. Par exemple, pour reconnaître un son, les caractéristiques souhaitées sont sa longueur, le rapport de bruit, et la puissance relative. Dans notre cas, les mesures statistiques sur la fenêtre sélectionnée se présentent sous la forme de mesures simples (par exemple, minimum, maximum, moyenne, variance...), à plus complexes (par exemple moment normalisé, énergie de Shannon...). Le tableau ci-dessous présente les caractéristiques statistiques considérées dans notre jeu de données.

1	Maximum	$\max_t s(t)$
2	Minimum	$\min_t s(t)$
3	Delta-min-max	$\max_t s(t) - \min_t s(t)$
4	Median	median(s)
5	Mean	$\mu = \frac{1}{T}\sum_{t=1}^{T} s(t)$
6	Variance	$\sigma^2 = \mu_2 = \frac{1}{T}\sum_{t=1}^{T}(s(t)-\mu)^2$
7	Standard deviation	σ
8	Normalised momentum for n=3	$\frac{\mu_n}{\sigma_n}$
9	Normalised momentum for n=4	$\frac{\mu_n}{\sigma_n}$
10	Normalised momentum for n=5	$\frac{\mu_n}{\sigma_n}$
11	Normalised momentum for n=10	$\frac{\mu_n}{\sigma_n}$
12	Energy	$\frac{1}{T}\sum_{t=1}^{T} s(t)^2$
13	Log-Energy	$\frac{1}{T}\sum_{t=1}^{T} log(1+s(t)^2)$
14	Shannon-Energy	$\frac{1}{T}\sum_{t=1}^{T} s(t)^2 log(1+s(t)^2)$
15	Max-n-derivative	$\max_t(s^n)$
16	Energy-derivative	Energy(s')
17	N-greater-threshold	$\sum_{t=1}^{T} 1_{t\|s[t]>\alpha}$
18	Peak-count	$\sum_{t=1}^{T} 1_{t\|s[t]>\alpha, s'(t)>\beta, -s'(t-1)<-\beta}$
19	Derivative-before-max	$s'(t_{max}-1)$
20	Derivative-after-max	$s'(t_{max})$
21	Derivative-before-min	$s'(t_{min}-1)$
22	Derivative-after-min	$s'(t_{min})$
23	Proportion-abs-lower	$\frac{1}{T}\sum_{t=1}^{T} 1_{t\|s[t]<\alpha}$
24	Mean-segment-lower	$\frac{1}{K}\sum_{k=1}^{K}(b_k - a_k)$
25	Percentile	percentile(s, α) (with α the percentage)
26	Interpercentile	percentile(s, α) - percentile(s, 1 - α)
27	log −mean-Peak(s)	$\log(1+\frac{1}{N}\sum_{k=1}^{N} \|s(t)\|\|s(t)\| > percentile(\|s\|, \alpha)$
28	log −mean-Valley(s)	$\log(1+\frac{1}{N}\sum_{k=1}^{N} \|s(t)\|\|s(t)\| < percentile(\|s\|, \alpha)$
29	log −mean-Dif	log −mean-Peak(s) - log −mean-Valley(s)

TABLE 1 – Les variables - Caractéristiques Statistiques extraites sur une fenêtre de taille fixe

Avec :
- t désigne le temps
- s(t) désigne un signal prétraité au temps t
- t_{min} (respectivement t_{max}) est le temps auquel le signal est minimum (maximum) dans la fenêtre
- μ la moyenne sur s : $\mu = \frac{1}{T}\sum_{t=1}^{T}(s(t))$.
  Le moment central d'ordre n d'une variable aléatoire X est défini comme $E[(X-\mu)^n]$. Nous désigne par μ_n sa version discrète, exprimée en : $\mu_n = \frac{1}{T}\sum_{k=1}^{T}(s(t)-\mu)^n$
- On note C l'ensemble de sous-signaux c_k de s tels que |s(t)| est supérieure à un seuil α : $C = \{s[t]_{a_k}^{b_k} / |s[t]| > \alpha, 1 < a_k < b_k < T\}$ et la longueur d'un segment c_k est alors $b_k - a_k$.

- La dérivée de s est calculée avec la différence discrète : s'(t) = s(t + 1) - s(t). Et la dérivée d'ordre n est obtenue de la même manière : $s(t)^n = s(t+1)^{(n-1)}$ - $s(t)^{(n-1)}$

Ces caractéristiques sont calculées sur le signal prétraité s, sa première dérivée s', et la valeur absolue de son transformée de Fourier. Ainsi, le jeu de données est constitué de 88 variables (87 va-

riables quantitatives et une variable qualitative nominale) et de 2817 observations. Les événements ont été enregistrés à l'aide du système de détection installé dans les maisons de soins infirmiers [1], d'où la collecte d'une quantité importante de données. Les variables sont centrées réduites et on n'a pas de valeur manquante.

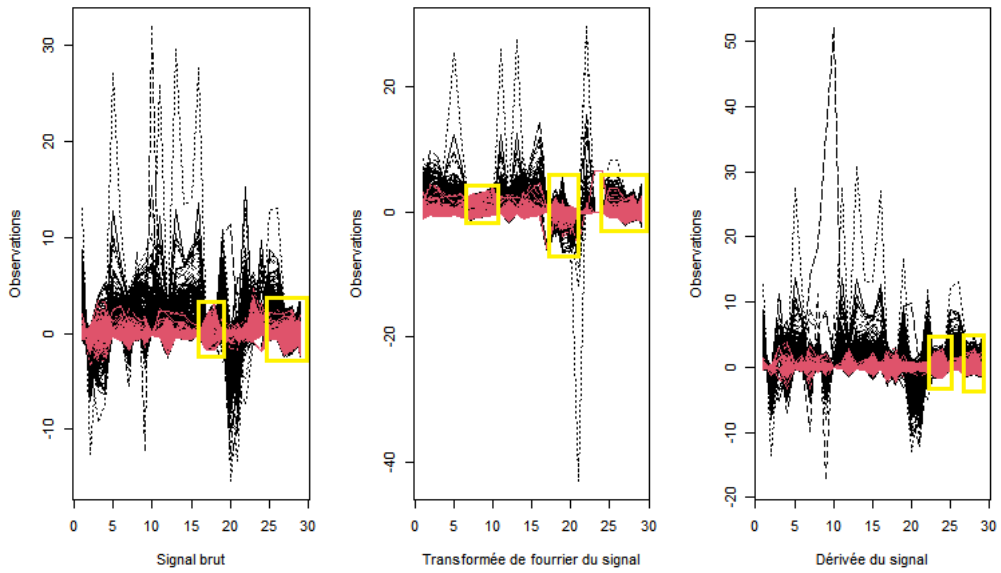

FIGURE 3 – Visualisation des données - Les signaux en rouge correspondent au cas de chute, et ceux en noir au cas de non chute - En abscisse, on a l'indice des variables décrites dans la table 1, et ceci pour les trois grands type de variables (le signal brut, la transformée de fourier du signal et sa dérivée)

D'après la figure 2, on remarque que l'amplitude des signaux en cas de non chute s'étend de -10 à 10 alors que celle des cas de chute est entre -2 et 2. En fait, les valeurs aberrantes sont fortement présentent dans des cas de non chute. Cela est vérifié aussi dans la figure 5.

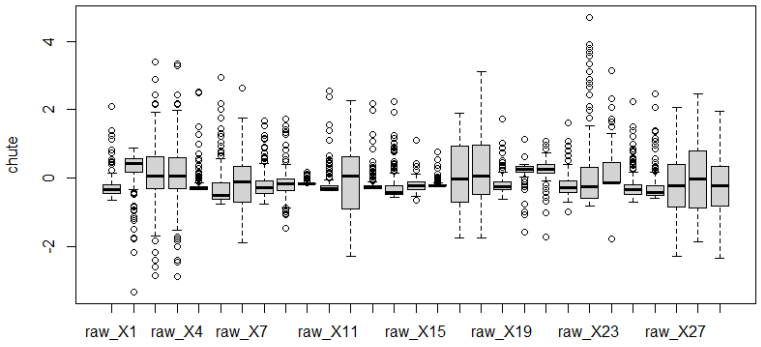

FIGURE 4 – Boîtes à moustache des variables centrées réduites définies sur le signal brut

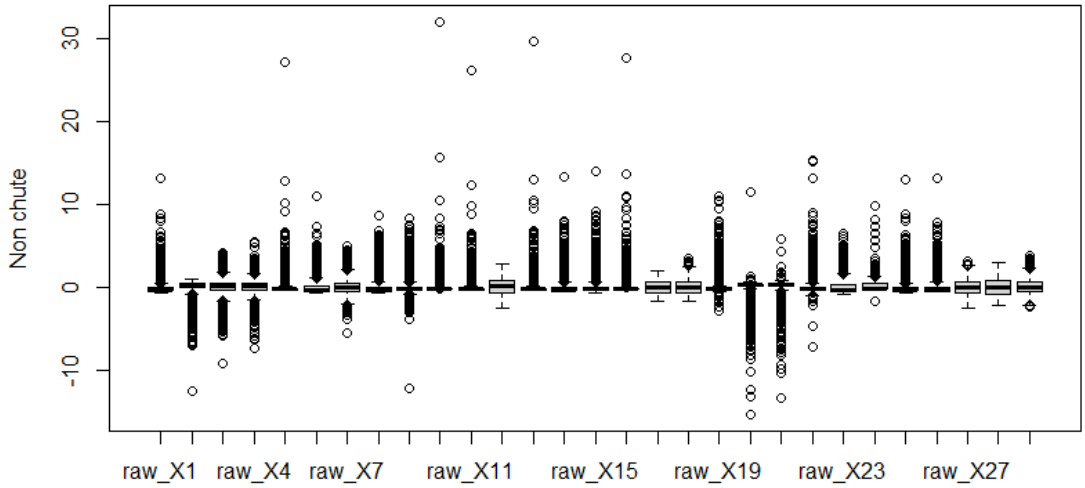

FIGURE 5 – Boîtes à moustache des variables centrées réduites définies sur le signal brut

Les boîtes à moustache des variables centrées réduites définies sur la transformée de fourier du signal et sa dérivée se trouvent en annexe. Le comportement des variables dans les cas de chute et non chute se distingue peu. Ces cas sont marqués en jaune dans la figure 3. Cependant, il existe des variables qui ne vérifient pas ceci, par exemple : $fourrier_{X7}$ et $deriv_{X29}$.

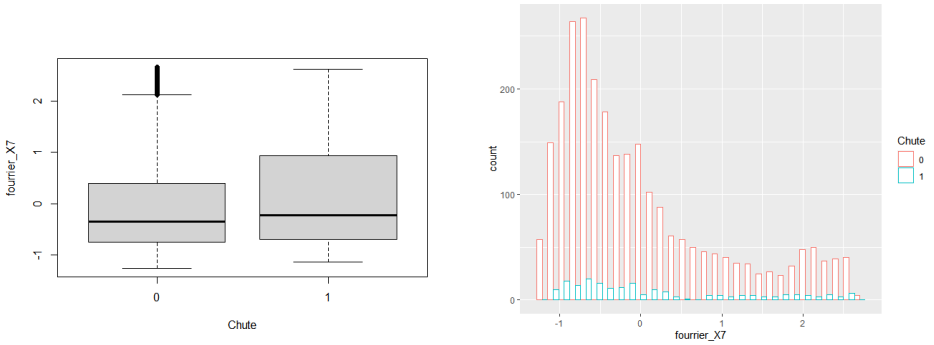

FIGURE 6 – La variable centrée réduite $fourrier_{X7}$ en fonction de la variable chute

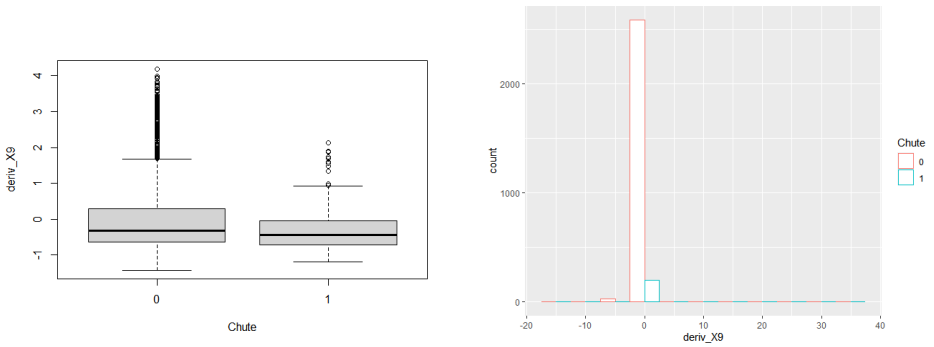

FIGURE 7 – La variable centrée réduite $deriv_{X9}$ en fonction de la variable chute

Il existe des signaux de cas de chute qui possèdent des valeurs pour la variable $fourrier_{X7}$ supérieures à celles des cas de non chute. Au contraire pour la variable $deriv_{X9}$. Cela est intéressant du point de vue de prédiction, il permettra à un algorithme de classification de bien discriminer les observations selon la variable chute.

2.1 Comparaison des moyennes

on s'intéresse dans cette partie à faire une comparaison de moyennes entre deux groupes d'observations par rapport à la variable chute. Pour cela, je décompose chaque variable en deux échantillons : chute et non chute.
Etant donnée la taille de l'échantillon, nous utiliserons l'approximation normale. Ainsi, on applique le test de Welsh pour comparer les moyennes des deux échantillons. Les résultats pour le signal brut et ses deux transformées se trouvent dans la tableau deux.

L'hypothèse nulle du test, H0, est acceptée si la p-value > 0.05 avec une erreur de seconde espèce inconnue. Sinon, on décidera que les moyennes sont différentes avec une erreur de première espèce de 5%.

$$\begin{cases} \text{H0 :Les moyennes sont égales} \\ \text{H1 :les moyennes ne sont pas égales} \end{cases}$$

	P-Value		
Variable	Signal brut	Transformée de fourrier du signal	Dérivée du signal
Maximum	$6.532e-13$	0.01668	$7.545e-16$
Minimum	$4.731e-10$	0.1096	$< 2.2e-16$
Delta-min-max	0.03572	0.05582	0.03344
Median	0.0637	$6.291e-09$	0.9755
Mean	$2.846e-10$	$1.867e-09$	$< 2.2e-16$
Variance	$1.605e-08$	$7.214e-07$	$1.822e-13$
Standard deviation	0.02661	0.03797	$1.104e-05$
Normalised momentum	$1.009e-09$	0.03875	$1.037e-06$
Energy	$9.93e-07$	0.03842	$1.487e-06$
Log-Energy	$8.031e-15$	0.04624	0.09495
Shannon-Energy	$7.497e-10$	$7.497e-10$	$< 2.2e-16$
Max-n-derivative	0.0651	0.004655	0.002217
Energy-derivative	$6.521e-11$	$1.237e-09$	$< 2.2e-16$
N-greater-threshold	$3.468e-13$	0.01679	$< 2.2e-16$
Peak-count	$< 2.2e-16$	0.001454	0.02913
Derivative-before-max	$< 2.2e-16$	0.0007855	$< 2.2e-16$
Derivative-after-max	0.2179	0.3625	0.02632
Derivative-before-min	0.02054	0.049	0.0321
Derivative-after-min	$5.903e-12$	0.0007169	0.0001474
Proportion-abs-lower	$3.171e-11$	0.8543	$2.404e-05$
Mean-segment-lower	$5.453e-12$	0.07121	$5.823e-09$
Percentile	$5.547e-10$	0.0007992	$1.193e-06$
Interpercentile	0.4658	0.3625	0.0634
Log-mean-Peak	0.8404	0.3625	0.08377
Log-mean-Valley	$6.57e-12$	$5.718e-06$	$1.779e-14$
Log mean Dif	$4.253e-12$	$5.31e-06$	≤2.2e-16
Normalised momentum (n=4)	0.0003961	$2.341e-05$	$1.546e-05$
Normalised momentum (n=5)	0.7048	0.05841	0.04231
Normalised momentum (n=10)	$7.903e-06$	0.001264	$9.679e-08$

TABLE 2 – Comparaison des moyennes

Le test rejette l'égalité sauf pour les variables : Delta-min-max, Median, Derivative-after-max, Interpercentile, Log-mean-Peak, Normalised momentum, Minimum, Proportion-abs-lower, Mean-segment-lower et Log-Energy. On remarque que, dans la majorité des cas, les moyennes ne sont pas égales. Ces variables sont donc intéressantes pour distinguer la chute de la non chute.

3 Machine Learning

Le machine learning (ML) est une composante de l'intelligence artificielle (IA). Il est très utilisé pour la Data Science et l'analyse des données. Il permet de développer, tester et d'appliquer des algorithmes d'analyse sur différents types de données afin de faire des prédictions. Il existe deux types de ML : supervisé et non supervisé.

3.1 Apprentissage Non Supervisé

L'objective de l'apprentissage non supervisé est de modéliser la structure des données non étiquetées en les regroupant, afin d'en découvrir des informations sur les données.

3.2 Analyse en Composantes Principales

L'Analyse en Composantes Principales (ACP) est une méthode de l'apprentissage non supervis [7]. Elle synthétise les informations importantes contenues dans une table de données multivariées en quelques nouvelles variables appelées composantes principales. Il s'agit donc de réduire la dimensionnalité d'une donnée multivariée à quelques composantes principales (le nombre dépend de la dimension des données), qui peuvent être visualisées graphiquement, avec une perte minimale d'information. Ces nouvelles variables correspondent à une combinaison linéaire des variables initiales. Dans ce but, la réduction de la dimension est obtenue en identifiant les premières composantes principales.
La première étape de l'ACP consiste à centrer (moyenne=0) les variables pour translater l'origine du repère au centre de gravité du nuage, et à réduire (variance = 1) les variables pour qu'elles aient le même poids dans l'analyse. L'ACP sera dite normée.

Les valeurs propres mesurent la variance de chaque composante principale. Elles sont grandes pour les premières composantes principales et petites pour les suivantes. Nous examinons alors les valeurs propres pour déterminer le nombre de composantes principales à prendre en considération.

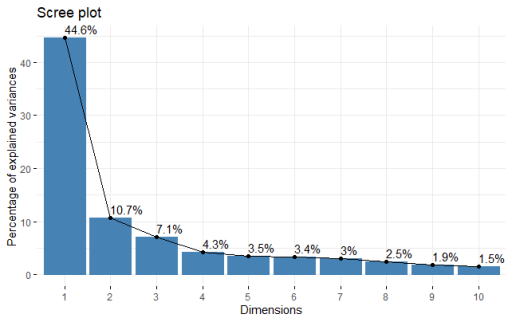

FIGURE 8 – Pourcentage d'inertie

Les 2 premiers axes de l'analyse expriment 55.99 % de l'inertie totale du jeu de données ; cela signifie que 55.99% de la variabilité totale du nuage des individus (ou des variables) est représentée dans ce plan. Du graphique de la figure 8, nous pourrions vouloir nous arrêter à la 3ème composante principale. 62.4% de la variance est expliquée par les trois premiers axes. En fait, d'après le critère du Coude, sur l'éboulis des valeurs propres, on observe un décrochement suivi d'une décroissance régulière. On choisit les axes avant le décrochement. Ces trois premiers axes présentent un éventuel intérêt pour la réduction des données.

3.2.1 Nuage des individus

On commence à ajuster le nuage des individus. les observations sont représentées par leurs projections, et la distance entre les individus projetés traduit leur différence ou similitude. Deux individus qui sont proches répondent d'une façon similaire au variables. Deux individus qui s'opposent ou qui sont loins prennent des valeurs différentes pour les variables. D'autre part, un individu qui est proche du centre de gravité a tendance à avoir des valeurs moyennes pour les

variables synthétisées. Par contre, les individus qui sont loins du centre prennent des valeurs différentes de la moyenne.
Examinons le nuage des individus : Les individus sont répartis selon le cas de chute ou non. On remarque qu'il existe des individus qui correspondent à des cas de non chute qui sont éloignés les uns des autres. Alors qu'ils existent des cas de chute qui sont très proches aux cas de non chute, ils ont donc des valeurs proches pour les variables.

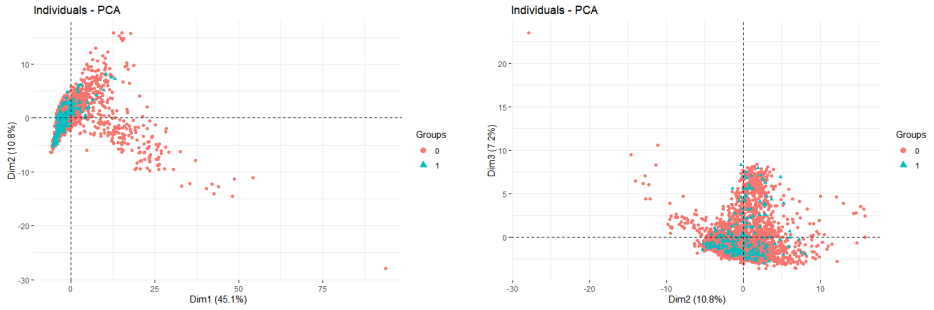

FIGURE 9 – Nuage des individus - Les individus sont colorés selon leur appartenance aux modalités de la variable Chute.

3.2.2 Nuage des variables

L'étude du nuage des variables suit la même démarche que celle du nuage des individus. Leur représentation diffère de celle des observations : les observations sont représentées par leurs projections sur les axes principaux, alors que les variables qui décrivent les données sont placées sur un cercle de rayon 1 qui représente leurs corrélations.

La figure 10 illustre le nuage des variables après le choix des composantes et des axes principaux.

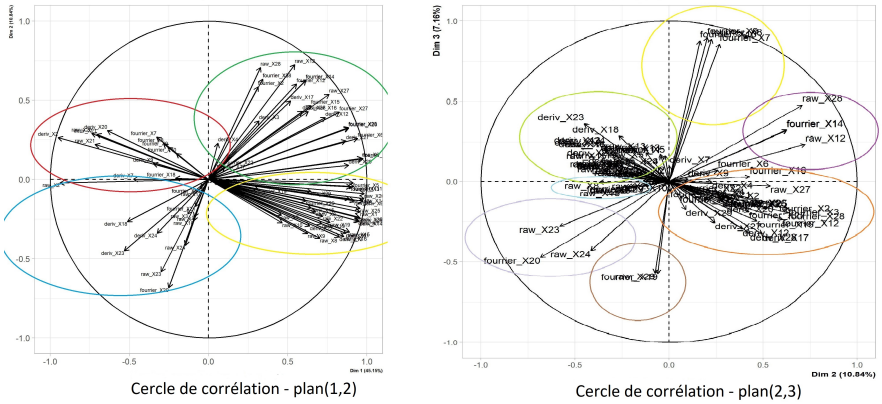

FIGURE 10 – Nuage des variables

Le cosinus de l'angle entre deux flèches correspond au coefficient de corrélation entre les deux variables correspondantes.

- Si l'angle entre deux variables bien représentées est droit alors elles sont décorrélées.
- Deux variables bien représentées, proche l'une de l'autre et qui sont dans la même direction, sont positivement corrélées.

— Deux variables qui s'opposent sur le graphique sont corrélées négativement.

La distance entre les variables et l'origine mesure la qualité de représentation des variables. Les variables qui sont loins de l'origine sont bien représentées par l'ACP.
Dans le plan(1,2), les extrémités des vecteurs représentant les variables ne sont pas toutes très proches du cercle des corrélations, ce qui montre que pas toutes les variables sont bien représentées, parmi ces variables on cite : raw_{X23}, $fourrier_{X18}$, $deriv_{X4}$ et autres. En examinant le cercle de corrélation de ce plan, on peut distinguer 4 groupes de variables dont dans chacun les flèches pointent dans la même direction. Les variables bien projetées de chaque groupe sont alors corrélées.

— Groupe 1 (couleur rouge) contient les variables suivantes :
$deriv_{X2}, raw_{X2}, deriv_{X9}, raw_{X21}, deriv_{X20}, deriv_{X21}, raw_{X20}, deriv_{X7}$.

— Groupe 2 (couleur vert) contient les variables suivantes : raw_{X28}, raw_{X12}, $fourrier_{X28}$, $fourrier_{X12}$, $deriv_{X17}$, $fourrier_{X14}$, $fourrier_{X15}$, $fourrier_{X3}$, $deriv_{X3}$, $deriv_{X4}$, $fourrier_{X22}$.

— Groupe 3 (couleur bleu) contient les variables suivantes :
$fourrier_{X20}, deriv_{X23}, raw_{X24}, deriv_{X18}, deriv_{X24}, raw_{X18}, raw_{X3}, raw_{X4}$.

— Groupe 4 (couleur jaune) contient les variables suivantes :
$raw_{X8}, deriv_{X8}, deriv_{X14}, fourrier_{X29}, fourrier_{X5}, deriv_{X25}, raw_{X1}, raw_{X15}, raw_{X25}, deriv_{X6}$.
Les variables de ce groupe contribuent positivement à la formation du premier axe.

Au contraire des groupes 2 et 4, les groupes 1 et 3 ne sont pas bien représentés.
Le premier axe oppose les groupes 2 et 4 aux groupes 1 et 3. Alors que le 2ème axe oppose les groupes 1 et 2 aux groupes 3 et 4. Les variables des groupes 1 sont fortement corrélées (négativement) avec les variables des groupes 4. De même pour les variables des groupes 2 et 3. D'autre part, Les variables des groupes 2 et 4 sont décorrellées. On ne peut pas affirmer ceci pour les groupes groupes 1 et 3 car elles ne sont pas bien representées.

On peut interpréter la position d'un individu sur le graphique du nuage des individus en fonction de la position des variables sur le cercle de corrélation. Un individu se situe du côté des variables pour lesquelles il possède des valeurs élevées. Prenons par exemple l'individu (2775) qui est le plus éloigné dans le nuage des individus. Il a vraisemblablement des valeurs fortes pour les variables du groupe 4 et des valeurs faibles pour les variables du groupe 3.

On remarque aussi que les variables raw_{X2}, raw_{X6} et $deriv_{X29}$ sont très corrélées au premier axe et contribuent le plus à sa formation : elles sont proches du cercle et proches de l'axe. Les individus ayant une forte valeur sur cet axe ont une forte valeur pour les variables corrélées positivement à lui et une valeur faible pour les variables qui sont corrélées négativement avec ce premier axe. Celles qui expliquent le mieux le deuxième axe sont : raw_X28 et $fourrier_X20$ pour les mêmes raisons.

D'autre part, on peut distinguer plus de groupes de variables dans le plan (2,3). Cependant, la majorité des variables ne sont pas bien représentées dans ce plan car ils sont proches du centre. Les variables les bien représentées sont remarquables : $fourrier_{X7}$, $fourrier_{X8}$, $fourrier_{X9}$, $fourrier_{X10}$, $fourrier_{X20}$ et raw_{X28}.

Plan	(1,2)	(2,3)
Projection moyenne des individus	48.70975	15.65758
Qualité de représentation moyenne des individus	0.3876301	0.2487993
Qualité de représentation moyenne des variables	0.5598822 2	0.1799721

TABLE 3 – Projection sur les plans

On constate donc que le plan formé par les axes 1 et 2 est celui qui la plus grande qualité de représentation des individus et d'informations (projection moyenne de 48.70975 individus sur ce plan).

3.2.3 Qualité de représentation

Pour interpréter correctement les résultats, il faut s'assurer que les individus et les variables sont bien représentés dans le plan contenant les axes principaux.
La qualité de représentation des individus se fait par le calcul du cosinus de l'angle formé avec sa projection sur l'axe. Un cos^2 élevé indique une bonne représentation de l'individu sur les axes principaux en considération. Un faible cos^2 indique que l'individu n'est pas bien représentée par les axes principaux.
La qualité de représentation des variables se fait graphiquement en examinant la longueur des flèches. Une variable est bien représentée lorsqu'elle est positionnée à proximité de la circonférence du cercle de corrélation. Par contre, elle est mal représentée lorsqu'elle est proche du centre du cercle et elle sera moins importante pour les premières composantes.

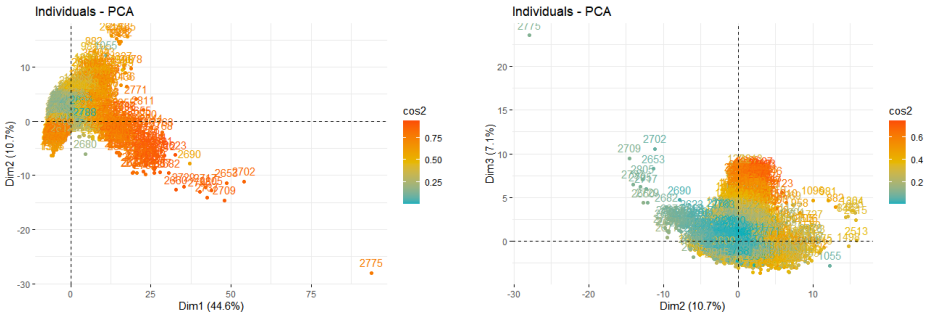

FIGURE 11 – Qualité de représentation des individus dans chaque plan

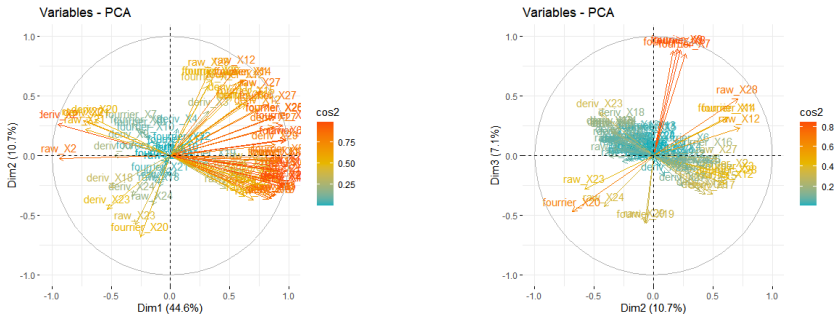

FIGURE 12 – Qualité de représentation des variables dans chaque plan

On remarque que l'individu 2775 est très éloigné du centre de gravité dans les trois plan. Il est donc l'individu le plus atypique. Sa représentation n'est pas mal dans le plan (1,2). De plus, en examinant les 2 plans ensemble, on peut voir que les individus sont mieux représentés dans les plans formés par les axes (1,2) que (2,3) où il y a moins d'inertie (le rouge est dominant).
On remarque les mêmes résutats pour la représentation des variables : elles sont beaucoup mieux représentées dans les plans formés par les axes (1,2). Ce qui est en fait évident.
Les figures 13 et 14 représentent les individus et les variables les plus bien représentés. On remarque que les 20 individus les mieux représentés sur le premier axe correspondent tous au cas

de chute. Par contre, ceux les mieux représentés sur les axes 2 et 3 sont des cas de non chute. Les variables les bien représentées diffèrent d'un axe à l'autre.

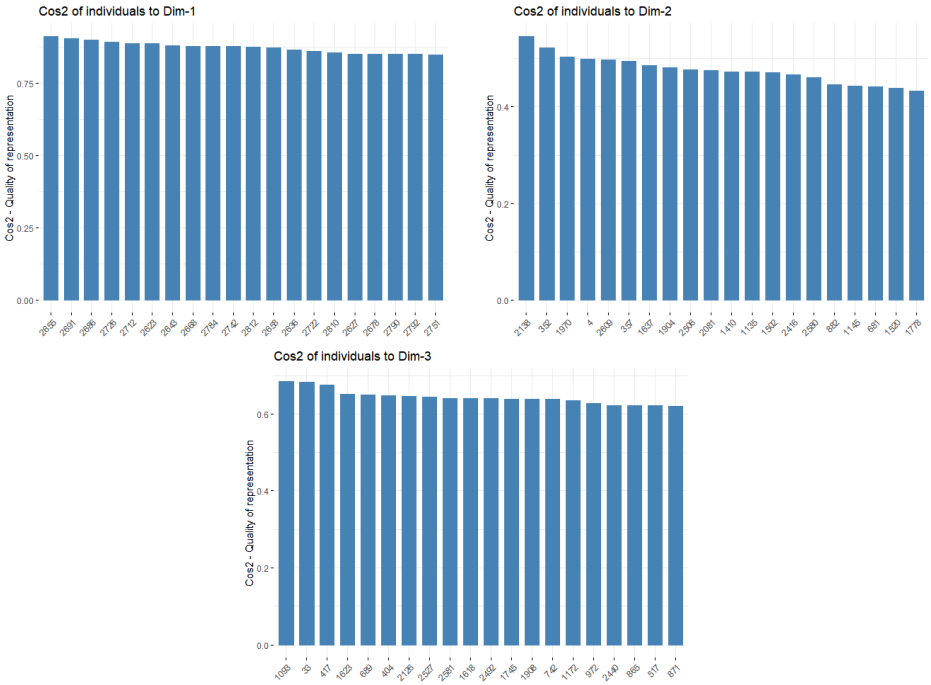

FIGURE 13 – Les 20 individus les plus bien représentés

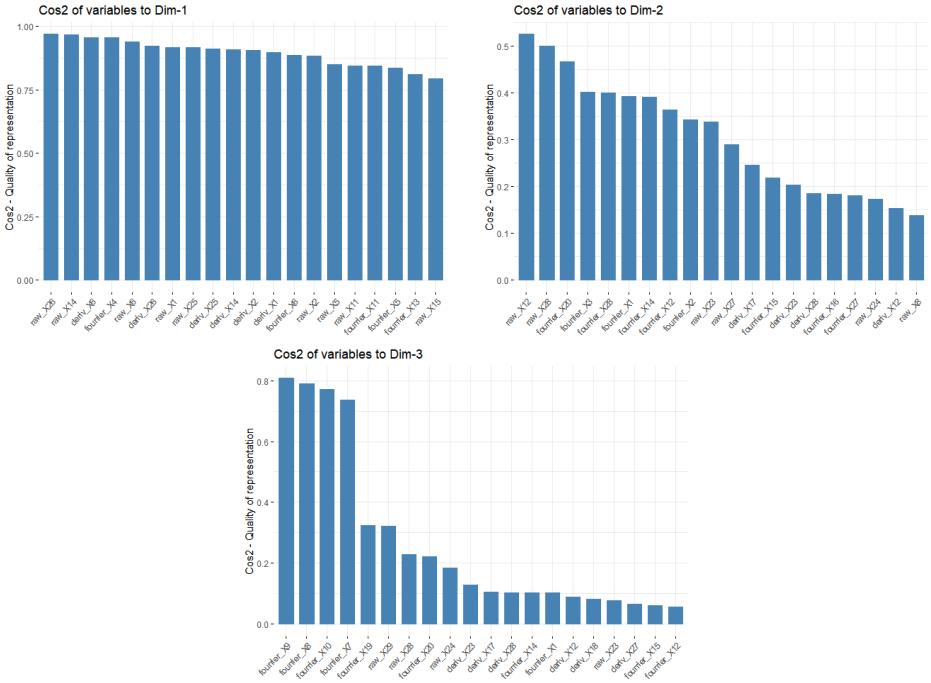

FIGURE 14 – Les 20 variables les plus bien représentées

3.2.4 Contribution au composantes principales

On s'intéresse à savoir le pourcentage de la contribution des individus et des variables dans la définition d'un axe principal. Les variables bien représentées et qui sont corrélées avec les trois composantes contribuent le plus à la formation des trois premiers axes principaux.

Les variables qui ne sont pas en corrélation avec un axe ou qui sont corrélées avec les derniers axes sont des variables à faible apport et peuvent être supprimées pour simplifier l'analyse globale.

Pour une variable qui a une forte contribution positive à un axe, les individus ayant une forte contribution positive à cet axe sont caractérisés par une valeur élevée pour cette variable.

Les individus qui sont éloignés du centre de gravité sont les plus contributifs aux axes principaux, alors que ceux qui sont proches du centre ne le sont pas.
Comme nos données contiennent de nombreuses variables, on a décidé de ne montrer que les principaux individus et variables contributives. Les résultats sont affichés ci-dessous :

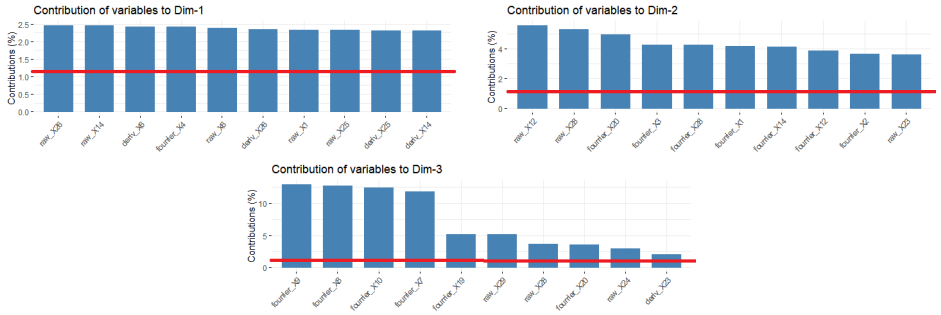

FIGURE 15 – Contribution des 10 meilleurs variables

La ligne pointillée rouge sur le graphique ci-dessus indique la contribution moyenne attendue par une variable.
On remarque de plus que, le 2775ième individu est le plus contributif pour les 3 axes, malgré qu'il n'est pas le mieux représenté parmi les individus. En fait, il est très éloigné du centre de gravité et des autres individus. Tandis que, la variable la plus contributive diffère d'un axe à l'autre.

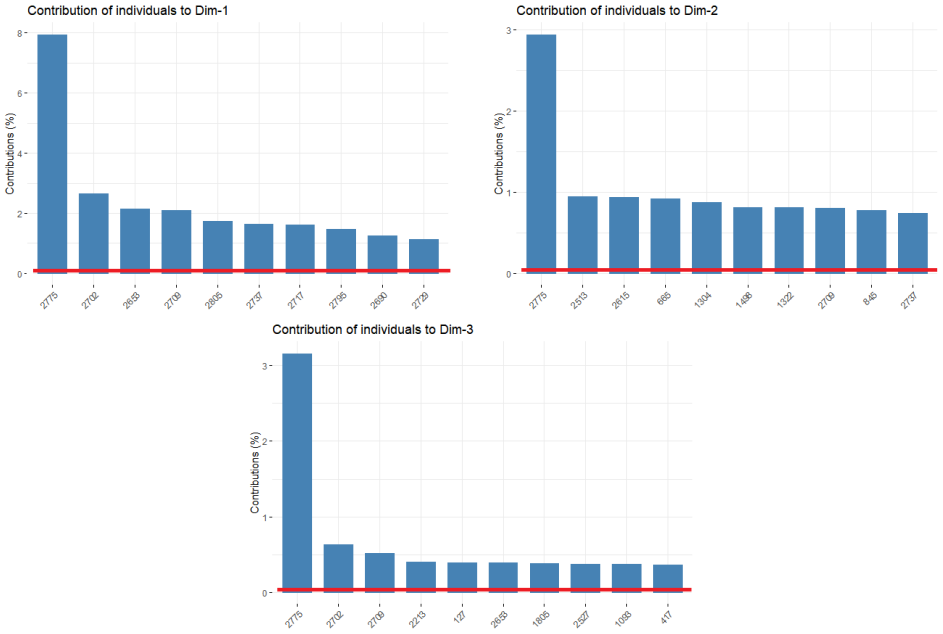

FIGURE 16 – Contribution des 10 meilleurs individus

Les figures 17 et 18 représentent respectivement la contribution de tous les variables et les individus aux composantes principales.

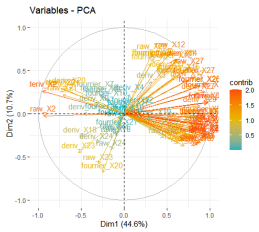

FIGURE 17 – Contribution des variables

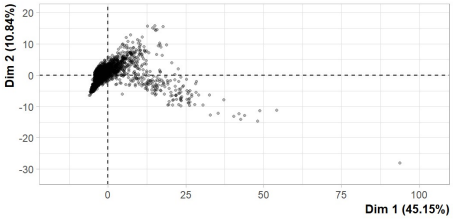

FIGURE 18 – Contribution des individus - Les individus libellés sont ceux ayant la plus grande contribution à la construction du plan.

3.3 Clustering

Le Clustering est une méthode de classification non supervisée qui possède des algorithmes d'apprentissage dont le but est de regrouper entre elles des données non étiquetées qui présentent des propriétés similaires. Il existe plusieurs méthodes de Clustering.

Dans nos études, pour chacune de ces méthodes, la dissimilarité est mesurée par la distance qui sépare les points.

3.3.1 La méthode des K-moyennes

La méthode des K-moyennes est l'un des algorithmes d'apprentissage non supervisé le plus couramment utilisé pour partitionner un ensemble de données en un ensemble de K groupes (clusters). Chaque cluster est représenté par son centre (c'est-à-dire centroïde) qui correspond au centre de gravité des observations affectées.
L'objectif général du Clustering est d'obtenir une forte similarité au sein de chaque groupe (faible variance au sein de chaque cluster) et une faible similarité entre chaque groupe (forte variance entre les groupes). La qualité du partionnement est mesurée par la variance intra-classe.
La première étape lors de l'utilisation du clustering k-means consiste à indiquer le nombre de clusters (K) à créer [8].
L'algorithme commence par sélectionner au hasard K observations de l'ensemble de données pour servir de centres initiaux pour les clusters.
Ensuite, chacune des observations restantes est affectée à son centroïde le plus proche, où le plus proche est défini à l'aide de la distance euclidienne entre l'observation et le centre du cluster.
Puis l'algorithme calcule le nouveau centre de gravité de chaque cluster. Chaque observation est vérifiée à nouveau pour voir si elle pourrait être plus proche d'un cluster différent.
Les étapes d'attribution de cluster et de mise à jour du centroïde sont répétées de manière itérative jusqu'à ce que les attributions de cluster cessent de changer (c'est-à-dire jusqu'à ce que la convergence soit atteinte).

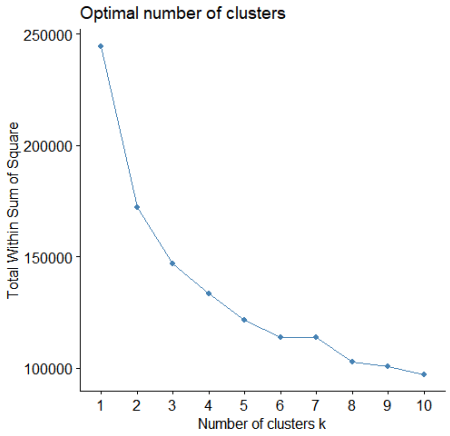

FIGURE 19 – Pourcentage de variance expliqué par les clusters par rapport au nombre de clusters - Méthode Elbow

Le choix du nombre de cluster est fait à partir de la méthode de coude, basée sur la minimisation de la somme des carrés des distances entre les individus et le centre au sein de chaque cluster. Le graphique ci-dessus représente la variance au sein des clusters. Elle diminue à mesure que K augmente. Les premiers clusters ajoutent beaucoup d'informations. Au-delà de 3, la minisation de l'inertie intra-classe n'est plus significative. Pour cela, on partitionne les données en trois groupes.

Cluster	Nombre d'individus	Chute	Non Chute
1	446	22	424
2	2248	178	2070
3	123	0	123

TABLE 4 – Composition de chaque cluster en fonction de la variable Chute

On remarque que la majorité des cas de chute, ainsi que les cas de non chute, se trouve dans le groupe 2. 22 cas de chute sont dans le groupe 1, alors que le groupe 3 ne contient que des cas de non chute. Cette répartition de chute dans différents groupes est présente pour plusieurs valeurs de K. Cette méthode ne nous permet donc pas de grouper les observations selon les cas de chute ou non. En fait, la distance euclidienne entre les individus est très petite. Pour cela, il est un peu difficile de séparer les observations convenablement selon la variable réponse.

FIGURE 20 – Représentation des observations en groupes

L'indice Silhouette nous permet de mesurer la qualité de la classification faite. Son but est de vérifier si chaque individu a bien été classé par l'algorithme. Pour cela, on calcule la valeur suivante pour chaque observation [9] :

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i)}$$

avec a(i) est la distance qui le sépare des autres individus de la classe à laquelle il appartient et b(i) est la distance qui le sépare des individus du groupe voisin. Pour une valeur de S(i) proche de 1, l'individu est bien classé : la distance qui le sépare du groupe voisin est plus grande que celle qui le sépare de sa classe. Par contre, si l'indice est proche de -1, l'individu est donc mal classé. D'autre part, pour une valeur autour de 0, l'individu pourrait également être classé dans le groupe voisin ou son groupe d'appartenance.

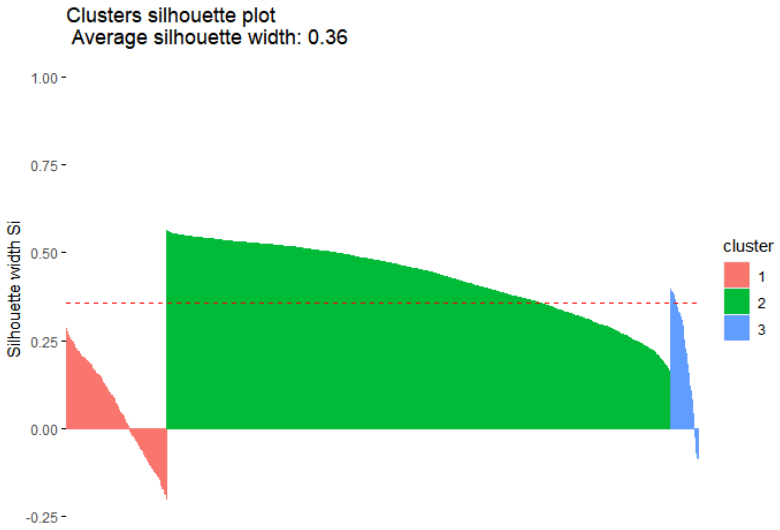

FIGURE 21 – Indice de silhouette pour chaque groupe

On remarque des coefficients négatifs. Les individus correspodants sont alors mal classés. Leur nombre est 185 dont la majorité fait partie du cluster 1. 175 de ces individus correspondents au cas de non chute et 10 individus sont des cas de chute. Les individus du goupe 1 devraient être dans le groupe 2, et ceux du groupe 3 devraient être classés dans le groupe 1.

3.3.2 Classification Ascendante Hiérarchique (CAH)

La classification ascendante hiérarchique est aussi une méthode de clustering. Elle forme pas à pas des connexions entre les individus, et utilise une matrice de distances pour trouver le regroupement le plus proche d'un autre. L'idée de départ est de considérer que chaque observation représente un groupe. La première connexion se fait entre les deux individus les plus proches. Puis l'algorithme calcule les centres des nouveaux clusters formés qui correspondent aux centres de gravité. On réitère cette étape jusqu'à connecter les deux derniers groupes.

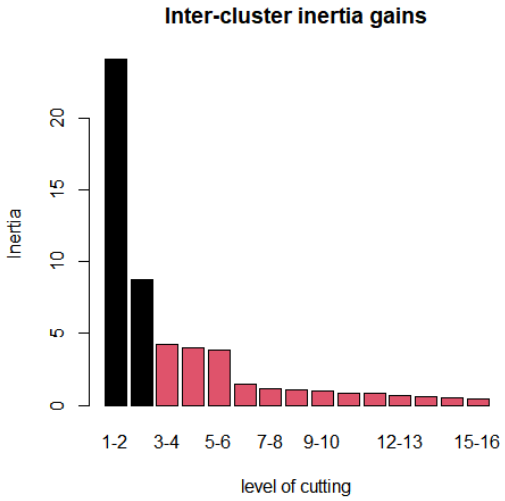

FIGURE 22 – la dissimilarité entre les groupes

La figure 22 nous montre le gain en dissimilarité entre les groupes suite à l'ajout d'un cluster supplémentaire. Nous observons sur ce graphique qu'au delà de 3 clusters le gain d'inertie n'est plus significatif (on cherche à maximiser celle-ci). Pour cela, on choisit K=3.

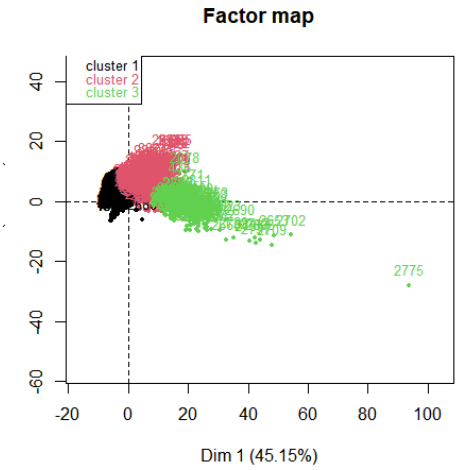

FIGURE 23 – Visualisation des individus par groupes

On observe qu'il y a des chevauchements.

Cluster	Nombre d'individus	individu parangon	distance
1	2171	400	0.7032314
2	519	1168	0.9788394
3	127	2620	0.9582621

TABLE 5 – Nombre d'observations dans chaque groupe - CAH

Un individu est dit parangon si ses coordonnées sont proches du centre de gravité du groupe. Le profil de cet individu caractérise alors le groupe auquel il appartient. Le tableau représente le numéro dans le jeu de données de l'individu parangon dans chaque groupe avec la distance associée. Ils correspondent à des cas de non chute.

Les observations du cluster 1 sont caractérisés par des valeurs pour les variables suivantes supérieures aux autres dans les autres classes. raw_{X2}, $deriv_{X2}$, $deriv_{X23}$, $deriv_{18}$, $deriv_{X7}$, raw_{X21} et raw_{X20}. Ces variables décrivent le plus ce cluster. En fait, la moyenne de chacune de ces variables dans ce cluster est supérieure à la moyenne générale correspondante et inférieure à cette moyenne générale dans les autres clusters.
De même, pour le groupe 2 et 3, les variables les caractérisants sont les suivantes :

- groupe 2 : $fourrier_{X15}$, raw_{X27}, $fourrier_{X25}$, $fourrier_{X26}$, $fourrier_{X27}$, $fourrier_{X12}$, $deriv_{X12}$ et $deriv_{X17}$. Ce cluster se caractérise par des faibles taux pour les variables importantes du goupe 1. Le cluster 1 se caractérise par des faibles taux pour ces variables.
- groupe 3 : $deriv_{X14}$, $deriv_{X1}$, raw_{X15}, $deriv_{X6}$, raw_{X25}, raw_{X14}, raw_{X26} et raw_{X19}. Ce cluster se caractérise par des faibles taux pour les variables du cluster 1.

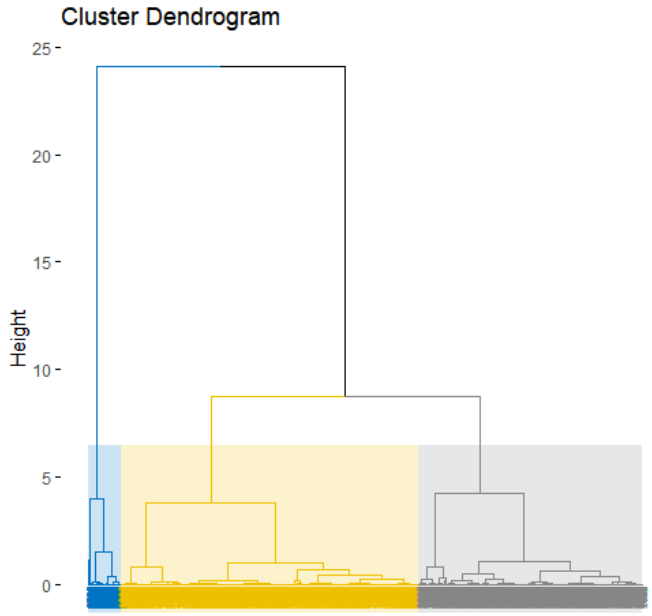

FIGURE 24 – Le dendrogramme généré par la classification

Les connexions successives se représentent sur un dendrogramme. Il nous donne la composition des différentes classes. L'axe des abscisses correspond aux observations, puisqu'elles sont nombreuses et illisibles sur l'axe, les étiquettes des individus ne sont pas affichées. L'axe des ordonnées correspond à l'indice d'agrégation : valeur de dissimilarité entre deux groupes fusionnés qui correspond à la distance associée à chaque connexion dans notre cas. On peut bien remarquer les 3 groupes.

4 Régression logistique

La régression logistique un algorithme de l'apprentissage automatique couramment utilisé pour une classification binaire. C'est un modèle statistique permettant d'étudier les relations entre un ensemble de variables explicatives X_i et une variable réponse Y_i. Elle permet aussi de prédire la probabilité qu'un événement arrive ou non. Plus la probabilité est proche de 1, plus l'évènement est susceptible d'avoir lieu, mais pas forcément. Ce n'est donc pas la réponse binaire (chute/non chute) qui est directement modélisée, mais la probabilité de réalisation d'une des deux modalités. Cette probabilité, est modélisée par une courbe sigmoïde, bornée par 0 et 1, et elle est définie par la fonction logistique suivante [10] :

$$P = \frac{\exp(\beta_0 + \beta_1.x1 + \beta_2.x2 + ... + \beta_k.x_k)}{1 + \exp(\beta_0 + \beta_1.x1 + \beta_2.x2 + ... + \beta_k.x_k)} \tag{1}$$

avec β_i sont les coefficients du modèle à estimer. Ce modèle n'est pas linéaire. En fait, La probabilité n'est pas une somme des effets des différentes variables explicatives.
Pour cela, on utilise la fonction logit qui permet de résoudre ce problème. Ainsi, la probabilité de réalisation s'exprime comme une combinaison linéaire des variable explicatives :

$$logit(P) = \log(\frac{P}{1-P}) = \sum_{i=1}^{k} \beta_i.Xi \tag{2}$$

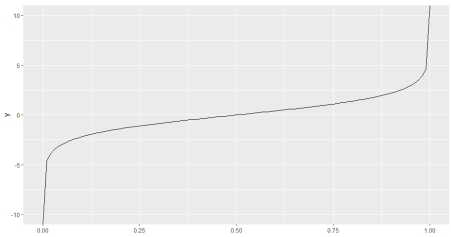

FIGURE 25 – Fonction logit - En abscisse on trouve la probabilité de Chute

4.1 Méthode pas à pas

Dans le but d'améliorer le modèle, on se base sur une technique de réduction du nombre des variables. On cherche donc parmi les variables X_i celles qui expliquent le mieux Y. La méthode se fait en plusieurs itérations. On supprime la variable la moins significative du modèle et on ajoute la plus significative. L'itération s'arrête lorsqu'aucune variable ne peut être ajoutée ou retirée. La qualité du modèle est déterminer par l'AIC. Plus il est faible, mieux est le modèle. Ainsi, en utilisant la fonction stepAIC du logiciel R, on procède de cette façon à diminuer l'AIC de 1453.36 pour le modèle initial à 1424 et réduire le nombre de variables à 10 : "Delta-min-max", "Mean", "Variance", "Normalised momentum", "Log-Energy ", "Energy-derivative", "N-greater-threshold ", "Peak-count", "Derivative-before-max".

4.2 Prédiction

Dans le but de faire des prédictions sur les données, je divise le jeu de données en deux : un échantillon d'apprentissage de 2253 observations (dont 160 individus sont des cas de chute) et un échantillon test de 564 observations (dont 40 cas de chute), comprenant les mêmes variables que l'échantillon d'apprentissage, mais des individus différents. Puis, j'applique le modèle de la

régression logistique sur le jeu de données d'apprentissage en considérant toutes les variables. Dans le but d'améliorer le modèle, j'applique la méthode pas à pas. Elle permet ainsi à réduire l'AIC de 1144.1 à 1065.28 et à retenir 30 variables, presque toutes significatives : raw_{X2}, raw_{X3}, raw_{X5}, raw_{X6}. raw_{X10}, raw_{X13}, raw_{X14}, raw_{X16}, raw_{X17}, raw_{X18}, raw_{X23}, raw_{X25}, raw_{X28}, raw_{X29}, $fourrier_{X5}$, $fourrier_{X6}$, $fourrier_{X10}$, $fourrier_{X14}$, $fourrier_{X18}$, $fourrier_{X19}$, $fourrier_{X20}$, $fourrier_{X21}$, $fourrier_{X25}$, $fourrier_{X26}$, $deriv_{X1}$, $deriv_{X2}$, $deriv_{X5}$, $deriv_{X6}$, $deriv_{X8}$, $deriv_{X10}$, $deriv_{X14}$, $deriv_{X16}$.

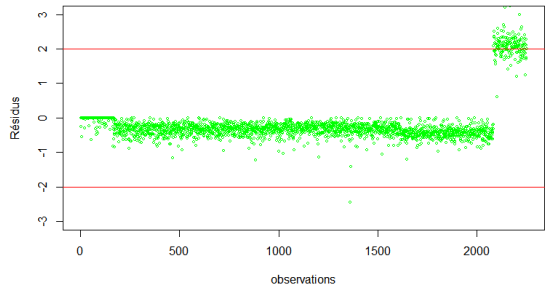

FIGURE 26 – Représentation des résidus pour le modèle de régression logistique appliqué à l'échantillon d'apprentissage

Le graphe des résidus est compris entre -2 et 2 pour presque 95 % des observations. Cependant, on remarque la présence des valeurs au-dessus de 2. Elles correspondent probablement à des valeurs abérrantes. On remarque de plus que les résidus sont à moyenne nulle et variance constante. Ainsi, le modèle est parfaitement ajusté.
Ensuite, ce modèle ajusté est utilisé pour prédire les réponses pour les observations dans le deuxième échantillon test.

4.3 Qualité de prédiction

La matrice de confusion permet d'évaluer la performance du modèle. Elle permet de connaître les différentes erreurs commises par un algorithme de prédiction, et leurs types [11]. Elle montre donc à quel point le modèle peut être confus lorsqu'il fait des prédictions. Il s'agit d'une matrice carrée de dimension deux.
Les résultats sont les suivants :

- Les vrais positifs (TP) : indiquent que le modèle a prédit une chute et qu'il sagit réellement d'une chute.
- Les vrais négatifs (TN) : indiquent indiquent que le modèle a prédit qu'il n y a pas eu une chute et qu'il ne sagit pas réellement d'une chute.
- Les faux positifs (FP) : indiquent que le modèle a prédit une chute et qu'il ne sagit pas réellement d'une chute.
- Les faux négatifs (FN) : indiquent que le modèle n'a pas prédit une chute alors qu'il sagit réellement d'une chute.

	Valeur Prédite = 0	Valeur Prédite = 1
Valeur Actuelle = 0	TN = 374	FP= 150
Valeur Actuelle = 1	FN = 22	TP= 18

TABLE 6 – Matrice de confusion pour un seuil de classification de 10%

De cette matrice, on peut calculer la sensibilité du modèle et la spécificité [12]. La sensibilité mesure la capacité du modèle à détecter correctement l'ensemble des signaux de chute. Notre but est donc de l'augmenter le plus possible, pour cela on a pris le seuil de classification à 10%.

$$\text{Sensibilité } = \frac{TP}{TP + FN} = 0.45 \tag{3}$$

La spécificité mesure la capacité du modèle à détecter correctement l'ensemble des signaux de non chute.

$$\text{Spécificité} = \frac{TN}{TN + FP} = 0.71 \tag{4}$$

Le modèle est donc capable de détecter les signaux de non chute plus que ceux qui correspondent à des chutes. En fait, pour un seuil de classification de 50 %, les vrais positives sont nulles. D'autre part, le nombre des individus positifs et qui sont correctement prédits diminue pour un seuil de classification supérieur à 10% pour s'annuler lorsqu'il dépasse 50%. Ceci implique que la sensibilité du modèle diminue et la spécificité augmente.
En fait, le jeu de données est déséquilibré et les cas de non chute sont sur-représentés. Pour pallier à ce problème, on réduit les observations de la classe majoritaire (non chute) de l'échantillon d'apprentissage. On obtient ainsi 800 observations dont 160 sont des cas de chute. On lance ensuite l'algorithme de la régression logistique. Examinons la matrice de confusion dans ce cas.

	Valeur Prédite = 0	Valeur Prédite = 1
Valeur Actuelle = 0	TN = 406	FP= 118
Valeur Actuelle = 1	FN = 22	TP= 18

TABLE 7 – Matrice de confusion pour un seuil de classificaton de 30% après Undersampling

Nous obtenons, le même nombre de chutes correctement prédites (TP=18) mais pour un seuil supérieur, ce qui est un peu satisfaisant. Pour un seuil de 50%, ce nombre correspond à 2 alors qu'il était nul avant d'appliquer le sous-échantillonage.

5 Forêt aléatoire

Dans le même concept, le forêt aléatoire est un algorithme de machine learning qui permet de classer les variables explicatives en fonction de leurs liens avec la variable à expliquer [13]. Il est constitué d'un ensemble d'arbres de décision indépendants. Chaque arbre possède une vision partielle sur le problème du fait d'un double tirage aléatoire [14]. Le premier tirage aléatoire est fait sur les observations. Pour chaque échantillon tiré, on construit un arbre. A chaque fois qu'un noeud doit être coupé, on tire au hasard une partie des variables de sorte que la valeur de la variable réponse soit la plus homogène possible dans les deux noeuds. Ce tirage sur les variables permet de réduire la corrélation entre les arbres et ses effets sur la qualité des résultats. Pour chacun des deux noeuds créés, on répète la deuxième étape jusqu'à la fin de la classification. A la fin, tous ces arbres de décisions indépendants sont assemblés. La prédiction faite par le random forest pour des données inconnues est alors celle qui a eu la majorité de vote (chute ou non) [15]. On applique l'algorithme du forêt aléatoire sur l'échantillon d'apprentissage et on test sa performance en faisant des prédictions sur les individus de l'échantillon test afin de les classer selon la variable chute. On examine la matrice de confusion qui nous permet d'analyser les résultats.

	Valeur Prédite = 0	Valeur Prédite = 1
Valeur Actuelle = 0	TN = 512	FP= 12
Valeur Actuelle = 1	FN = 27	TP= 13

TABLE 8 – Matrice de confusion - Forêt aléatoire

On remarque que l'algorithme arrive à bien classer les cas de non chute correctement (TN = 517). Dans ce cas, la sensibilité vaut 0.325 et la spécificité 0.97 .

Comme le but est de bien détecter une chute, j'applique aussi un sous-échantillonage pour améliorer le nombre des individus positifs bien classés parmi l'ensemble des individus positifs(TP). En fait, il est nécessaire de faire ici cette méthode car le classement final se fait par majorité de votes.

	Valeur Prédite = 0	Valeur Prédite = 1
Valeur Actuelle = 0	TN = 445	FP= 79
Valeur Actuelle = 1	FN = 13	TP= 27

TABLE 9 – Matrice de confusion - Forêt aléatoire - Undersampling method

On remarque que le nombre des vrais positives a augmenté ce qui est satisfaisant. La sensibilité a augmenté aussi à 67.5 %, alors que la spécificité a légèrement diminuée à 84 %. L'algorithme arrive donc à mieux classé les cas de chute.

5.1 L'indice de Gini

L'indice de Gini est un indice d'impureté d'un noeud. Il mesure le niveau d'inégalité de la répartition de la variable réponse dans le noeud. Il indique donc l'hétérogénéité ou l'homogénéité des noeuds. plus il est élevé, plus les données sont uniformes.
En outre, il permet de montrer, pour chaque variable, son importance dans la classification des données. Ainsi, le choix de la variable de séparation à chaque noeud se fait en se basant sur cet indice [14].
La diminution moyenne de l'indice de Gini est une mesure de la façon dont chaque variable contribue à l'homogénéité des noeuds [16]. Plus la diminution moyenne du score de Gini est élevée, plus l'importance de la variable dans le modèle est élevée.

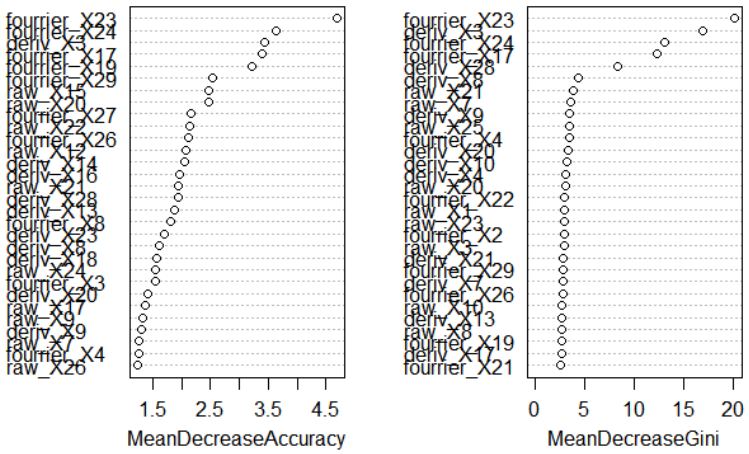

FIGURE 27 – La valeur moyenne de diminution de Gini

Le tracé de la diminition de la précision exprime la perte en précision si la variable correspondante est retirée du modèle [17]. Les variables sont présentées par importance décroissante. Plus la diminution moyenne de l'accuracy est élevée, la variables est importante. Dans ce cas, la varriable $fourrier_{X23}$ contribue le plus à la fois à la précision du modèle de classification et à la discrimination des observations selon la variable chute.

5.2 La courbe de ROC

La courbe ROC représente le comportement d'un algorithme de classification à deux classes pour tous les seuils de détection possibles. Elle permet de décrir la performance du modèle en traçant le couple (1-spécificité, sensibilité) selon un seuil de classification [18]. Lorsque le seuil augmente, la sensibilité diminue et la spécificité augmente et vise-versa. On peut donc optimiser soit la sensibilité soit la spécificité, et non pas les deux ensemble.

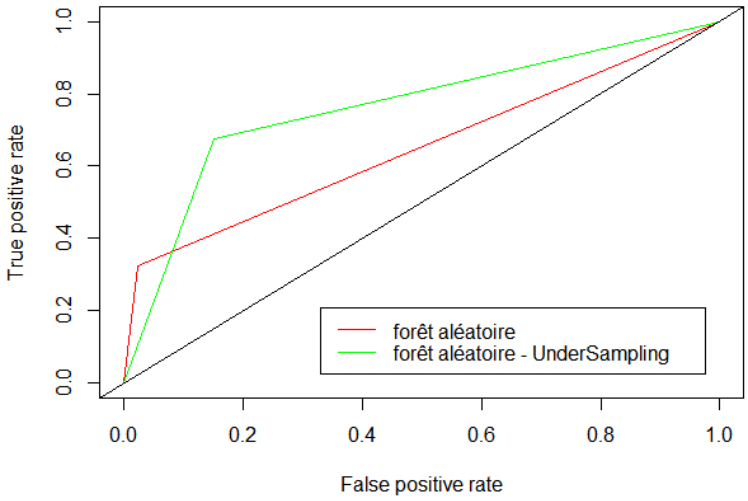

FIGURE 28 – Courbe ROC - forêt aléatoire

La ligne droite représente ce que donnerait un classifieur qui classe au hasard, et donc, il obtient autant de faux positifs que de vrais positifs quel que soit le seuil choisi. En dessous, le classifieur n'est pas bon. Au dessus, le classifieur est meilleur. Plus la courbe a des valeurs élevées, moins le modèle fait d'erreur. D'après la figure 28, on remarque que la courbe du modèle après le sous-échantillonage est beaucoup plus élevée que celle qui représente le cas initial.
Pour quantifier globalement les performances du classifieur quel que soit le seuil, on calcule l'AUC qui correspond à l'aire de la surface sous la courbe de ROC. Il est égale à 100 % pour un modèle parfait et 50 % pour un modèle non-informatif (aléatoire) [19]. D'autre part, l'accuracy mesure le pourcentages des bonnes prédictions. Il est donné par la formule suivante :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

	Forêt aléatoire	Forêt aléatoire - Undersampling	Régression logistique
AUC	0.651	0.7621	0.6916
Accuracy	0.93	0.83	0.695

TABLE 10 – Performance des modèles

Malgré que le modèle initial du forêt aléatoire est plus précis, les résultats de ce modèle après le sous-échantillonage sont les plus satisfaisants.

6 Conclusion

L'objectif principal du capteur intelligent est de détecter la chute des personnes. Le jeu de données collecté est déséquilibré. En fait, la chute est un évènement rare, pour cela elle est sous-représentée.

Ceci a provoqué des problèmes lorsqu'on a appliqué des algorithmes de classification. Pour pallier à un problème de déséquilibre, plusieurs méthodes sont possibles : soit en augmentant synthétiquement la classe minoritaire, soit en réduisant la classe majoritaire. On a procédé par la deuxième méthode pour éviter d'avoir des chutes synthétiques. En fait, le signal d'une chute enregistré dans un milieu contrôlé peut être différent que celui d'une chute réelle, car dans ce cas, la personne sera consciente avant qu'elle tombera et essaiera d'éviter le plus possible des blessures graves.

Les résultats suite à cela sont satisfaisants et on a remarqué que l'algorithme du forêt aléatoire est plus performant que celui de la régression logistique.

D'autres algorithmes d'apprentissage supervisé et non supervisé existent. Il sera intéressant de les appliquer et les comparer pour dégager le modèle le plus performant.

A Annexe - Matrice de corrélation

Une corrélation existe entre certaines variables et elle est négligeable entre autres. Pour des raisons de lisibilité, on représente la corrélation en 3 matrices dans les figures 30, 31 et 32.

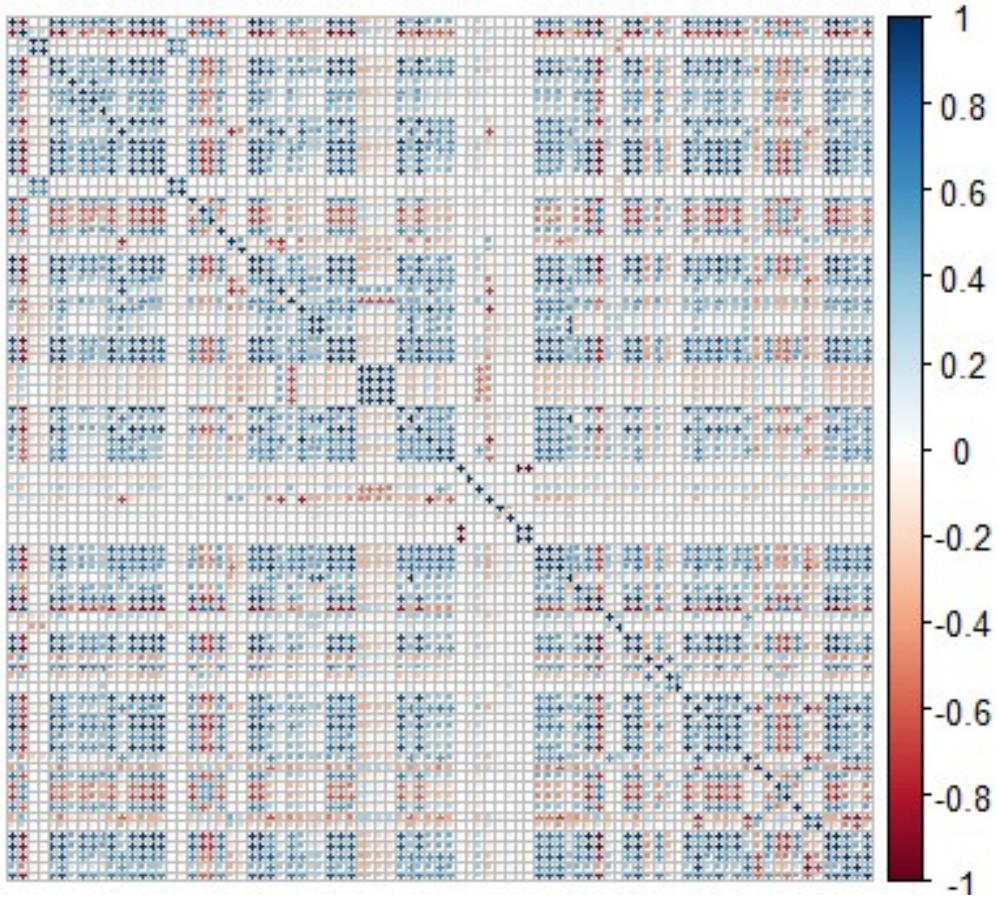

FIGURE 29 – Matrice de corrélation des variables

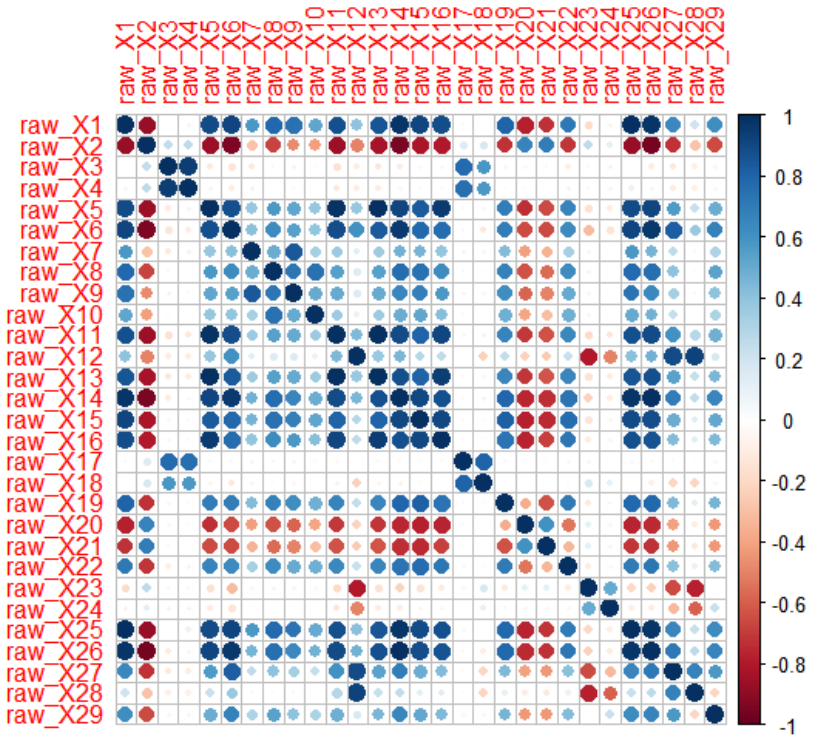

FIGURE 30 – Matrice de corrélation des variables

La couleur bleu foncé indique une forte corrélation positive entre les variables. La couleur rouge foncé indique une forte corrélation négative entre les variables et la couleur blanc indique une indépendance des variables.
On remarque que chacune des variables suivantes : raw_{X1}, raw_{X2}, raw_{X5}, raw_{X6}, raw_{X11}, raw_{X13}, raw_{X14},raw_{X15}, raw_{X16}, raw_{X19}, raw_{X20}, raw_{X21}, raw_{X22} ,raw_{X25} , raw_{X26} sont fortement corrélées avec les autres. On peut donc les enlever pour améliorer les algorithmes de classification.

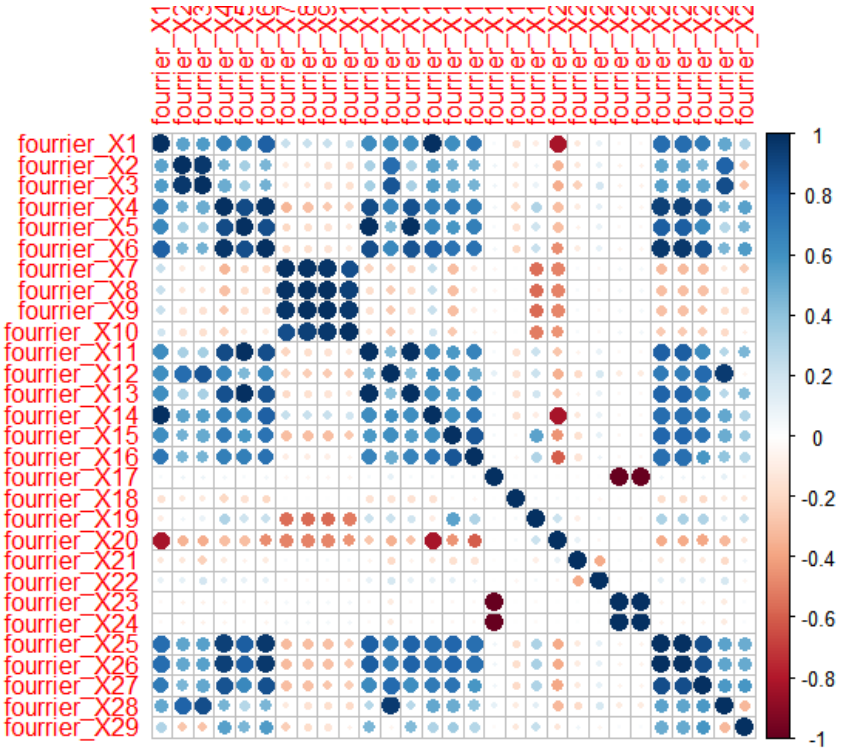

FIGURE 31 – Matrice de corrélation des variables

Chacune des variables suivantes : $fourrier_{X8}$, $fourrier_{X9}$,$fourrier_{X10}$, $fourrier_{X3}$, $fourrier_{X6}$, $fourrier_{X5}$, $fourrier_{X25}$, $fourrier_{X26}$ sont fortement corrélées avec les autres. On peut donc les enlever pour la même raison qu'avant.

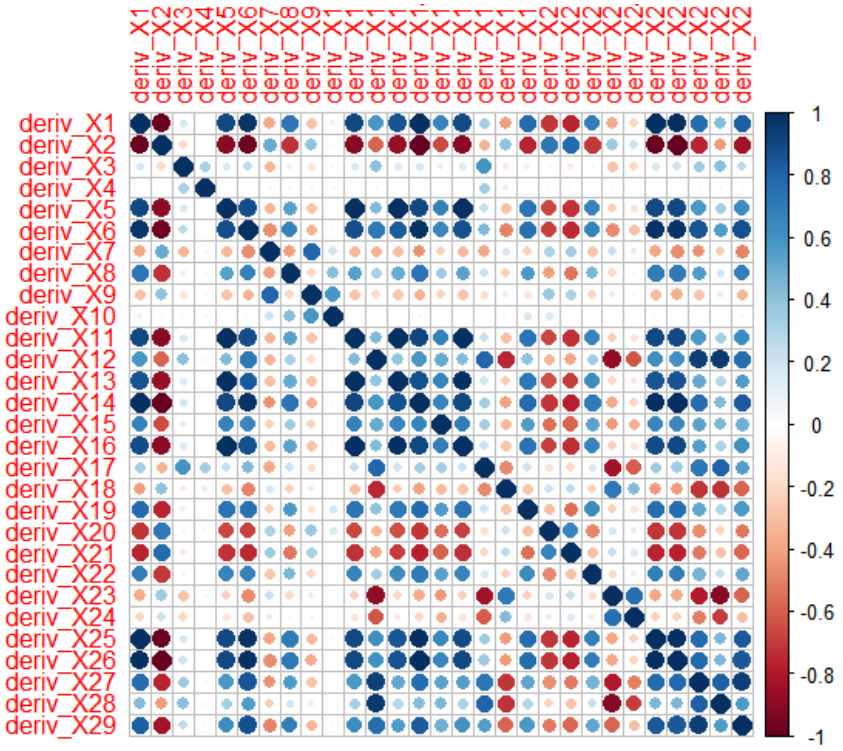

FIGURE 32 – Matrice de corrélation des variables

De même pour les variables suivantes : $deriv_{X1}$, $deriv_{X2}$, $deriv_{X5}$, $deriv_{X6}$, $deriv_{X11}$, $deriv_{X13}$, $deriv_{X14}$, $deriv_{X16}$, $deriv_{X25}$, $deriv_{X26}$, $deriv_{X21}$, $deriv_{X22}$, $deriv_{X29}$

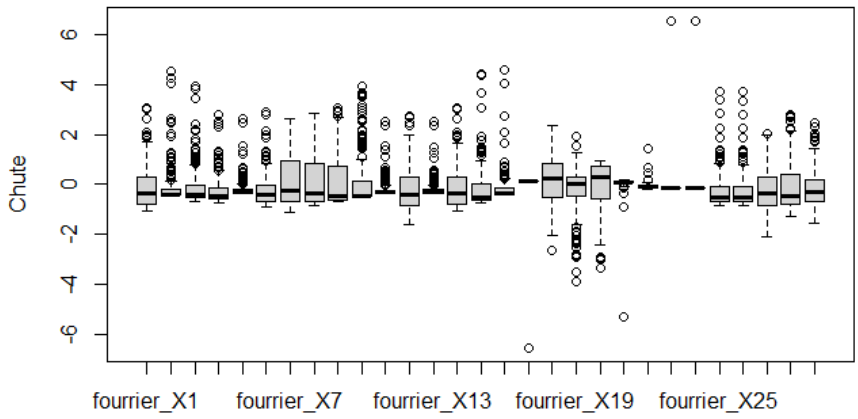

FIGURE 33 – Boîtes à moustache des variables centrées réduites définies sur la transformée de fourrier du signal

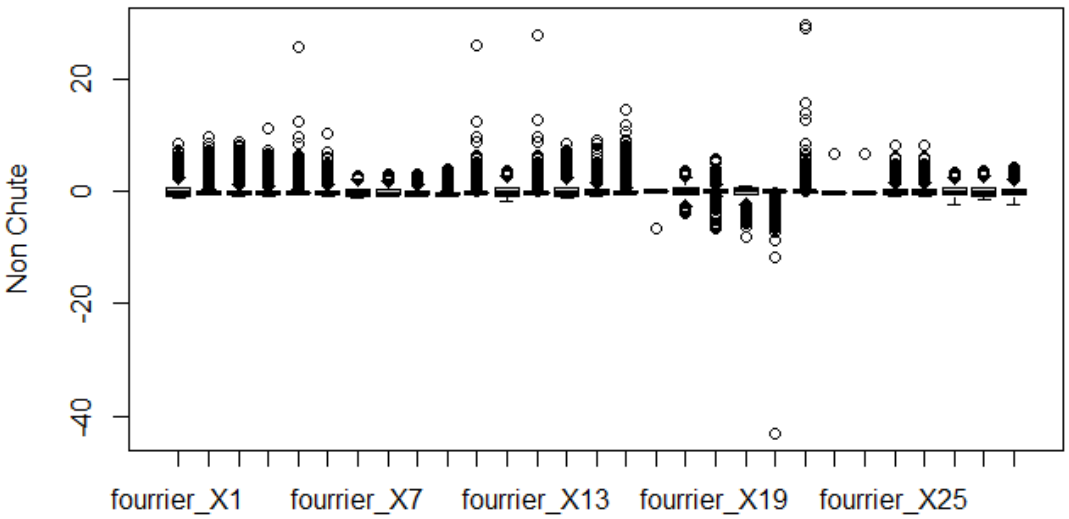

FIGURE 34 – Boîtes à moustache des variables centrées réduites définies sur la transformée de fourrier du signal

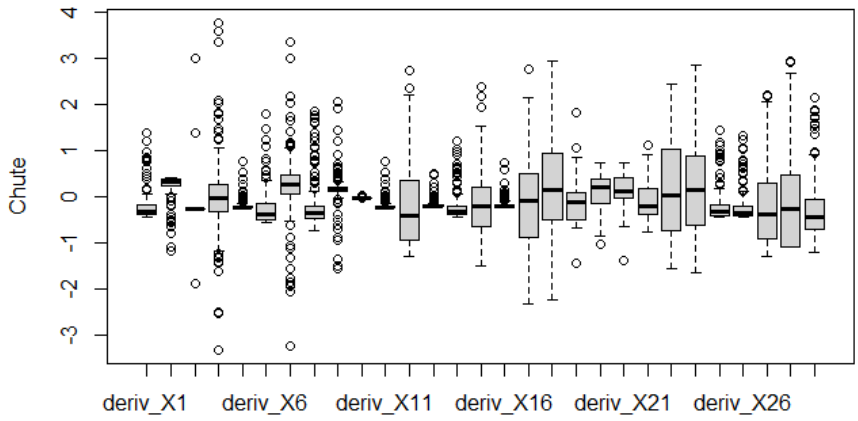

FIGURE 35 – Boîtes à moustache des variables centrées réduites définies sur la dérivée du signal

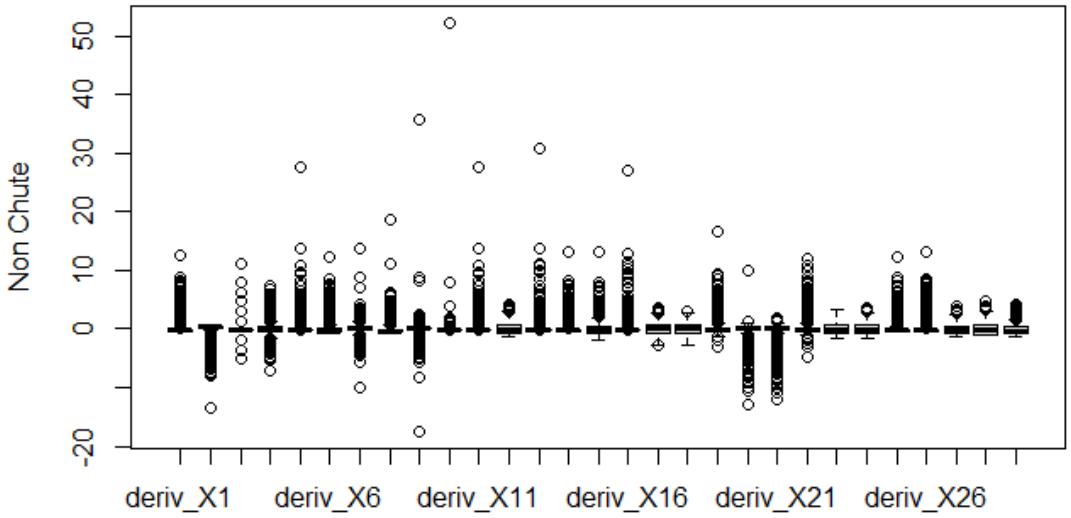

FIGURE 36 – Boîtes à moustache des variables centrées réduites définies sur la dérivée du signal

Références

[1] Ludovic Minvielle. Classification d'événements à partir de capteurs sols - application au suivi de personnes fragiles. September 2020.

[2] WHO. Falls. 26 April 2021.

[3] Brigitte Bourguignon. Prévention des chutes des personnes âgées. page 1, Février 2022.

[4] France Bleu Limousin Nathalie Col. La nouvelle eco - morphée + : un dispositif innovant de détection des chutes inventé par un limougeaud. octobre 2021.

[5] Rédaction SilverEco. Floorinmotion : un sol connecté qui veille sur les seniors. janvier 2016.

[6] Jean-Charles Guézel. Un revêtement de sol détecteur de chutes pour les établissements médicaux. 05 2014.

[7] Kassambara. Articles - principal component methods in r : Practical guide. septembre 2017.

[8] Younes Benzaki. Tout ce que vous voulez savoir sur l'algorithme k-means. avril 2018.

[9] Jack Sparrow. Silhouette algorithme pour déterminer la valeur optimale de k. juillet 2022.

[10] Wikipédia. Régression logistique — wikipédia, l'encyclopédie libre, 2022.

[11] Matrice de confusion : comment la lire et l'interpréter ? décembre 2021.

[12] xlstat. Analyse de sensibilitÉ et spÉcificitÉ. décembre 2022.

[13] Husson et al. Forêts aléatoires. septembre 2008.

[14] Équipe Data Science. Algorithme n°2 - comprendre comment fonctionne un random forest en 5 min. juin 2020.

[15] Antoine Crochet-Damais. Random forest (ou forêt aléatoire) : définition et cas d'usage. mai 2022.

[16] Lvaudor. Classification par forêts aléatoires posted on. juin 2015.

[17] Jose Ignacio Martinez-Taboada, Fernando ; Redondo. Variable importance plot (mean decrease accuracy and mean decrease gini). 2020.

[18] Aymeric Duclert. Courbe roc. mai 2016.

[19] Classification : Roc curve and auc. juillet 2022.