Classification de la Santé Fœtale

Timofey Abramov, Yazan El Mahmoud, Selima Khessairi

14/08/2025

Résumé

Ce document constitue le rapport de projet de SY09, portant sur la classification de l'état de santé foetale, à partir du jeu de données Fetal Health Classification. Ce rapport aborde l'exploration du jeu de données, puis la prédiction de l'état de santé d'un foetus grâce à des méthodes de classifications non supervisées et supervisées.

Introduction

La surveillance du bien-être fœtal est un enjeu majeur de santé publique. Selon l'ONU, près de 300 000 femmes décèdent chaque année de complications liées à la grossesse, principalement dans les pays à ressources limitées. La cardiotocographie (CTG), méthode largement utilisée pour monitorer la santé fœtale, permet d'évaluer le rythme cardiaque fœtal et les contractions utérines.

Cependant, l'interprétation des tracés CTG reste complexe et subjective. Une analyse plus systématique de ces données pourrait améliorer la détection précoce des situations à risque.

Pour cette raison, nous chercherons lors de cette étude à utiliser ces données afin de prédire efficacement la santé foetale.

Présentation du jeu de données Le jeu de données Fetal Health Classification, rassemble 2 126 enregistrements de cardiotocographies (CTG) avec leurs caractéristiques cliniques associées. Il s'agit de données cliniques réelles collectées en milieu hospitalier. Il se compose de 21 variables descriptives numériques et d'une variable cible qualitative classant la santé fœtale en trois catégories : normal, suspect et pathologique.

Les variables descriptives peuvent être regroupées en trois catégories principales :

- Paramètres du rythme cardiaque fœtal : valeur basale (en bpm), variabilité à court et long terme, nombre et amplitude des accélérations/décélérations
- Mesures utérines : nombre et durée des contractions, valeur maximale des contractions
- Autres indicateurs cliniques : mouvements fœtaux, activité utérine anormale, suspicion de souffrance fœtale

1 Analyse Exploratoire des Données

1.1 Visualisation de la distribution de la variable cible

Dans cette section, nous analysons la distribution de la variable cible, c'est-à-dire l'état de santé fœtale.

Tout d'abord, nous remarquerons que les données sont normalisées et complètes (sans valeurs manquantes). Ce qui simplifie les étapes de traitement.

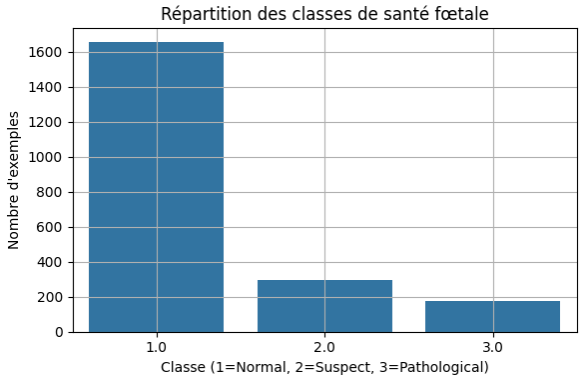

FIGURE 1 – Distribution de la variable cible

On observe un déséquilibre marqué entre les classes, avec une surreprésentation des cas normaux : `77.84%` des cas sont catégorisés comme "normaux", `13.87%` des cas sont catégorisés comme "suspects" et enfin `8.27%` des cas sont "pathologiques". Cette caractéristique devra être prise en compte dans l'analyse.

1.2 Corréaltions entre variables

Les données étant toutes quantitatives (mesures de CTG) mis à part la variable cible, il est possible d'étudier leurs corrélations via une matrice de corrélation.

L'analyse de la matrice de corrélation révèle plusieurs relations intéressantes entre les variables issues de la cardiotocographie. Certaines, comme `histogram_mean`, `histogram_median` et `histogram_mode`, présentent des corrélations très fortes entre elles (supérieures à 0.9) et avec la variable `baseline_value`, ce qui indique une forte redondance : toutes ces mesures sont liées à la fréquence et au rythme cardiaque foetaux. Concernant la variable cible `fetal_health`, aucune caractéristique ne montre de corrélation très élevée (ce qui est courant dans les données médicales multifactorielle), mais certaines se distinguent tout de même : `prolongued_decelerations` (0.48), `percentage_of_time_with_abnormal_long_term_variability` (0.43), et `abnormal_short_term_variability` (0.47) semblent les plus informatives pour la prédiction de l'état de santé fœtale.
Ces résultats soulignent l'intérêt de réduire les redondances via des techniques telles que la sélection de variables ou l'analyse en composantes principales (ACP), tout en orientant les efforts de modélisation sur les variables les plus corrélées à la cible. Les variables moins corrélées, comme `fetal_movement` ou `light_decelerations`, pourraient s'avérer moins discriminantes mais méritent néanmoins d'être testées dans un modèle prédictif.

1.3 Analyse en Composantes Principales

Afin de réduire cette redondance entre variables, nous avons appliqué une ACP. Cette dernière nous a permis de visualiser les données dans un espace à 2 dimensions tout en conservant une part significative de l'information. Les deux premières composantes principales expliquent à elles seules environ 45,55% de la variance totale.

La projection dans le premier plan factoriel montre alors une certaine séparation des classes, surtout pour les cas pathologiques et normaux, mais surtout un chevauchement entre les classes suspectes et les autres, ce qui confirme la complexité du problème.

2 Approches non supervisées

Afin d'évaluer la performance de différents modèles de classification sur cette tâche, nous avons tout d'abord testé une approche non supervisée à l'aide de l'algorithme k-means et la CAH, dans l'espoir de voir si les différentes classes (normal, suspect, pathologique) pouvaient émerger spontanément à partir des données.

2.1 K-means

L'algorithme K-means a été appliqué sur les données brutes, avec `n_clusters` = 3, déterminé par la méthode du coude, mais également par cohérence au jeu de données (nous avons 3 classes pour la variable). Les résultats se sont révélés insatisfaisants, avec des clusters qui ne correspondaient pas bien aux vraies classes et un indice de rand ajusté très faible (0.045).

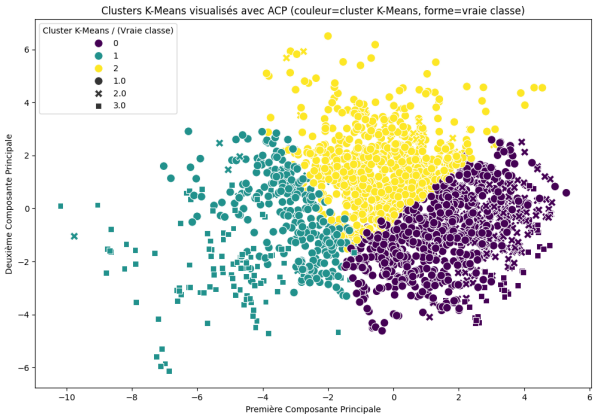

FIGURE 2 – Comparaisons entre les classes issues des K-means et les vraies classes du jeu de données.

Ceci peut être dû à plusieurs raisons, notamment aux hypothèses fortes faites par cet algorithme. En effet, l'algorithme k-means suppose que les données forment des clusters sphériques, de taille similaire, séparés par des frontières linéaires. Or, dans notre cas, les classes ("normal", "suspect", "pathologique") se chevauchent fortement, notamment entre "normal" et "suspect", comme l'a révélé l'ACP. De plus, k-means minimise uniquement la variance intra-cluster sans tenir compte de la signification médicale des classes, qu'il ne connaît pas. Il est aussi peu adapté à des données redondantes, corrélées ou distribuées de manière non linéaire, car il repose sur la distance euclidienne. Enfin, le déséquilibre entre les classes favorise la création de clusters dominés par la classe majoritaire ("normal"), au détriment des cas plus rares mais importants comme les "pathologiques".

Afin de faire face au problème des variables redondantes, nous avons tenté d'effectuer les K-means sur les 5 premiers axes principaux d'inerties issus de l'ACP (expliquant 67% de l'inertie) ainsi que sur les 5 variables les plus corrélées à la variable cible. Bien que nous obtenons de meilleurs résultats par rapport au K-means sur données brutes, les résultats ne sont toujours pas satifaisants (pour K-means sur ACP : ARI = 0.201 et pour les 5 variables les plus corrélées : ARI = 0.550). De même, l'utilisation de la distance Mahalanobis à la place de la distance euclidienne, dans le but de tenir compte de la dispersion des points qui varie selon la classe considérée (les classes ne sont pas spéhriques ni toutes pareils), n'a pas été très concluante (indice de rand ajusté de 0.189).

2.2 Classification ascendante hiérarchique (CAH)

En algorithme non supervisé, nous avons également essayé d'effectuer une CAH avec le critère de Ward, qui n'a pas non plus très bien fonctionné : en effet, le score de silhouette issu de cette dernière a été très faible (0,084), ce qui signifie que le clustering est quasiment fait aléatoirement. L'indice de Rand ajusté est également très faible (0.147). Ceci se voit également en comparant les clusters créés avec nos vraies classes issues de la variable cible. Ceci peut être dû à plusieurs raisons : tout d'abord, les clusters sont imbriqués et non hiérarchiques. Les trois classes de `fetal_health` ne sont pas nécessairement bien séparées dans l'espace des variables à cause des chevauchements entre classes ce qui nuit aux regroupements successifs de la CAH. De plus, la CAH ne gère pas bien les données bruitées ou non sphériques et ne s'adapte pas bien aux clusters de forme complexe ou allongée. Enfin, le nombre de dimensions est de nouveau problématique. En haute dimension (en l'occurence 21), la notion de distance euclidienne devient peu discriminante. Cela peut faire que les distances se ressemblent, donc les regroupements sont peu significatifs. Afin de contourner le problème de fléau des dimensions, nous allons essayer d'appliquer la CAH sur les composantes principales issues de l'ACP, notamment les 5 premières, permettant également de supprimer la redondance issues de fortes corrélations entre certaines variables. Toutefois, même après ACP, si les classes naturelles ne sont pas séparables linéairement, la CAH va continuer à faire des regroupements peu cohérents avec les vraies classes (`fetal_health`). La CAH sur les 5 premières composantes principales n'améliore pas significativement les résultats : le score de silhouette est de 0.189 et l'indice de Rand ajusté est de 0.116.

3 Approches supervisées

Etant donné que nos données contiennent d'ores et déjà la variable cible (données labellisées) et que les algorithmes non supervisés n'ont pas été très concluants, nous nous intéresserons par la suite à des algorithmes supervisés. L'ensemble de ces méthodes ont été effectuées sur des données stratifiées, et divisées en données d'apprentissage (80%) et en données de test (20%).

3.1 K Plus Proches Voisins

Nous nous intéressons donc d'abord aux KPPV. Afin de déterminer le nombre optimal de voisins à choisir, une validation croisée avec 10 folds et une stratification a été effectuée. Ainsi, nous nous assurons que chaque sous-ensemble de données conserve la même distribution de classes. Le nombre optimal de voisin obtenu est k=4. Les résultats de l'application des KPPV sont présentés dans le tableau 1. Nous remarquons en effet de bons résultats pour les classes *normal* et *pathologique* tandis que les mesures pour la classe *suspects* sont plus bas. Ceci pourrait être dû à l'ambiguïté de la classe suspects, dont les mesures médicales peuvent être très similaires à celles des cas pathologiques ainsi qu'à celles des cas sains.

TABLE 1 – Tableau résumant les mesures liées au KPPV effectué sur les trois classes

Classe	Précision	Rappel	F-1 Score
Normal	0.95	0.96	0.96
Suspect	0.69	0.69	0.69
Pathologique	0.84	0.77	0.81

Afin de limiter le problème de représentation de la classe *suspect*, nous avons testé de regrouper cette dernière avec la classe pathologique (car il est toujours moins grave de diagnostiquer une maladie chez un individu sain que de ne pas le diagnostiquer alors qu'il est véritablement malade). Nous avons désormais deux classes : la classe *normal* et la classe *anormal*. Nous remarquons des résultats similaires au KPPV sur deux classes qui permet lui de distinguer les individus malades des individus suspects (tableau 2) notamment pour la classe *anormal*. Il est ainsi plus intéressant de garder les trois classes.

TABLE 2 – Tableau résumant les mesures liées au KPPV effectué sur deux classes

Classe	Précision	Rappel	F-1 Score
Normal	0.94	0.96	0.95
Anormal	0.86	0.77	0.81

3.2 Analyse discriminante

Nous avons ensuite exploré l'Analyse discriminante en testant l'Analyse Discriminante Quadratique (ADQ), l'Analyse Discriminante linéaire (ADL) et le classifieur naïf de Bayes (NB). Nous avons ainsi effectué l'hypothèse de normalité des données. Les données ont été standardisées et stratifiées pour garantir une distribution équilibrée lors de l'entraînement et du test.

3.2.1 Classifieur Naive Bayes Gaussien

Le modèle GNB atteint une précision globale de 81 %. Il est particulièrement performant pour détecter les cas *Normaux*, mais a des difficultés avec les cas *Suspects* et *Pathologiques*.

TABLE 3 – Métriques de performance par classe

Classe	Précision	Rappel	F-1 Score
Normal	0.97	0.85	0.91
Suspect	0.45	0.80	0.57
Pathologique	0.53	0.46	0.49

Ces résultats montrent les limites du modèle face au chevauchement entre les classes, notamment entre les cas à risque.

Cette approche permet également de prendre en compte les coûts liés aux erreurs de classification, ce qui est crucial dans un contexte médical, où certaines erreurs peuvent avoir de lourdes conséquences. La matrice des coûts utilisée est représentée dans le tableau 4).

TABLE 4 – Matrice de coût utilisée

	Prédit Normal	Prédit Suspect	Prédit Pathologique
Vrai Normal	0	0.5	2
Vrai Suspect	1	0	3
Vrai Pathologique	5	4	0

Cette approche fait baisser légèrement la précision globale à 80.3%, mais oriente le modèle vers des décisions plus prudentes. Le modèle accepte de classer davantage de cas "Normaux" comme "Suspects" (46 contre 42 précédemment) afin de réduire le risque de sous-estimer un cas sérieux. Le rappel de la classe "Suspect" s'améliore légèrement (de 0.80 à 0.81), tandis que les performances sur la classe "Pathologique" restent stables (rappel 0.46, précision 0.53).

La nouvelle matrice de confusion montre une stratégie différente : le modèle préfère surestimer le risque (faux positifs) plutôt que de passer à côté d'un cas grave. Cela se traduit par une réduction des erreurs critiques, ce qui est bien plus adapté au domaine médical.

Même si l'accuracy globale est un peu plus basse (0.80 face à 0.92), l'approche coût-sensible est mieux adaptée à la réalité clinique. Elle permet de hiérarchiser les erreurs : mieux vaut suspecter à tort une pathologie que de la manquer. Ce compromis entre précision globale et sécurité des patients en fait une stratégie plus réaliste et responsable en médecine.

3.2.2 Analyse discriminante Linéaire

La méthode ADL suppose une distribution normale avec une matrice de covariance partagée entre les classes. Elle offre une précision globale de 85.9 %, la meilleure obtenue jusqu'ici. La classe "Normal" est très bien reconnue, avec 93% de précision et 95% de rappel. Cela signifie que la quasi-totalité des cas sains sont bien identifiés. En revanche, les performances chutent pour les classes à risque : La classe "Suspect" n'atteint que 54% de précision et 58% de rappel, ce qui indique des hésitations fréquentes du modèle. La classe "Pathologique", bien plus critique, est correctement identifiée dans seulement 46% des cas, avec 67% de précision.

TABLE 5 – Métriques de performance par classe

Classe	Précision	Rappel	F-1 Score
Normal	0.93	0.95	0.94
Suspect	0.54	0.58	0.56
Pathologique	0.67	0.46	0.54

3.2.3 Analyse discriminante Quadratique

La QDA permet une matrice de covariance différente par classe, ce qui accroît la flexibilité mais augmente significativement le nombre de paramètres à estimer.

Avec la QDA, l'objectif était d'exploiter la flexibilité accrue pour mieux séparer les classes complexes. Pourtant, ses performances globales sont légèrement en baisse, avec une précision globale de 81.2 %.

La classe "Normal" reste bien identifiée (tableau 6) et le rappel de la classe "Suspect" croît, un net progrès par rapport à la ADL. Toutefois, cette amélioration se fait au détriment de la classe "Pathologique", pour laquelle le rappel decroît fortement, bien que la précision s'améliore.

Table 6 – Métriques de performance par classe

Classe	Précision	Rappel	F-1 Score
Normal	0.95	0.86	0.90
Suspect	0.44	0.81	0.57
Pathologique	0.76	0.37	0.50

En clair, la QDA identifie mieux les cas suspects mais rate encore plus de cas pathologiques, avec 6 cas sur 35 classés à tort comme "Normaux", une erreur encore plus fréquente qu'avec la ADL.

La ADL se montre plus stable et plus adaptée à ce jeu de données. Malgré sa simplicité, elle offre un meilleur équilibre entre performance globale et sécurité. La QDA, plus complexe, échoue à mieux détecter les cas critiques, notamment les pathologies, probablement à cause de chevauchements importants entre classes. Ces résultats rappellent qu'un modèle plus complexe n'est pas toujours synonyme de meilleures performances, surtout quand il s'agit de prendre des décisions sensibles comme en médecine.

3.3 Régression logistique

La Régression Logistique, souvent sous-estimée, est un algorithme robuste et efficace, même dans des contextes multiclasse. Nous avons utilisé ici sa version multinomiale, adaptée à notre tâche de classification en trois catégories.

Les résultats sont excellents, avec une précision globale de 88.5 %.

La classe "Normal" est presque parfaitement reconnue, avec très peu d'erreurs critiques. La classe "Suspecte" montre des performances respectables, tandis que la classe "Pathologique", la plus critique en médecine, est détectée avec une grande précision mais encore un rappel perfectible.

Malgré quelques confusions entre les cas "Suspect" et "Normal", la Régression Logistique se distingue par sa stabilité et sa fiabilité, notamment grâce à une quasi-absence de faux positifs critiques.

Table 7 – Métriques de performance par classe

Classe	Précision	Rappel	F1-score
Normal	0.94	0.95	0.94
Suspect	0.61	0.68	0.64
Pathologique	0.88	0.66	0.75

Elle parvient à classer 314 cas normaux sur 332 correctement, et 23 cas pathologiques sur 35, avec seulement 3 cas pathologiques classés à tort comme "Normaux", ce qui reste acceptable dans un cadre médical.

En résumé, la Régression Logistique combine simplicité et efficacité : elle surpasse tous les modèles précédents sauf KPPV en précision globale et offre une excellente reconnaissance des cas sains, un bon compromis pour la classe "Suspecte", et une grande prudence sur les cas pathologiques.

3.4 Arbre de Décision

L'Arbre de Décision est un modèle très populaire, apprécié pour son interprétabilité et sa capacité à capturer des relations complexes et non linéaires dans les données. Nous l'avons appliqué à notre tâche de classification en trois catégories pour évaluer sa performance face à des modèles plus linéaires.

La classe "Normal" est très bien reconnue, et la classe "Pathologique", la plus critique en médecine, est détectée avec un très bon rappel, supérieur à celui de la régression logistique. La classe "Suspecte", quant à elle, reste la plus difficile à cerner.

Table 8 – Métriques de performance par classe

Classe	Précision	Rappel	F1-score
Normal	0.94	0.95	0.94
Suspect	0.72	0.64	0.68
Pathologique	0.81	0.83	0.82

Malgré une précision globale supérieure, l'Arbre de Décision présente un profil de risque différent. S'il identifie plus de cas pathologiques (rappel de 83 %), il commet également plus d'erreurs critiques.

Il parvient à classer 316 cas normaux sur 332 correctement, et 29 cas pathologiques sur 35. Cependant, 5 cas pathologiques sont classés à tort comme "Normaux", ce qui représente une augmentation du risque par rapport à la régression logistique (3 cas). Cette faiblesse est un

compromis direct de sa plus grande sensibilité aux cas pathologiques.

L'analyse des modèles supervisés a confirmé leur supériorité sur les approches non supervisées. Le K-Nearest Neighbors (0.906), l'Arbre de Décision (0.899) et la Régression Logistique (0.885) ont obtenu les meilleures accuracies globales (voir figure 4).

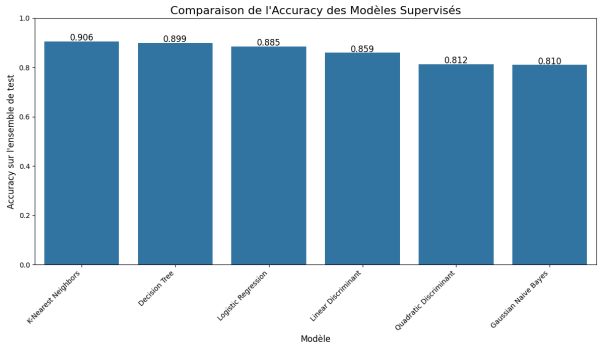

FIGURE 3 – Comparaison de l'Accuracy des Modèles Supervisés.

4 Conclusion

En somme, ce projet a démontré que les approches supervisées sont très efficaces pour prédire la santé fœtale, surpassant largement les méthodes non supervisées qui ont échoué face à la complexité des données. Les modèles K-Plus Proches Voisins, Arbre de Décision et Régression Logistique se sont avérés les plus performants, avec des précisions globales autour de 90%. L'analyse révèle cependant un compromis crucial : si l'Arbre de Décision détecte le plus de cas pathologiques, la Régression Logistique se montre plus fiable en minimisant les erreurs de diagnostic les plus graves, ce qui en fait un choix potentiellement plus sûr en contexte clinique.

A Signification médicale des variables

A.1 Variables Numériques

— **baseline_value** – Rythme cardiaque fœtal de base (en bpm), normalement entre 110–160 bpm. La distribution montre une majorité de valeurs dans cette plage (fœtus sains), avec quelques cas <110 bpm. Les fœtus "à risque" présentent une médiane plus élevée et des outliers plus fréquents.

— **accelerations** – Nombre d'accélérations cardiaques par seconde, indicateur de bonne santé. Majoritairement nulles (histogramme), sauf pour les fœtus sains (boxplot).

— **fetal_movement** – Fréquence des mouvements fœtaux par seconde. Distribution asymétrique : peu de mouvements fréquents vs absence majoritaire.

— **uterine_contractions** – Fréquence des contractions utérines par seconde. Pic à 0, puis distribution normale (médiane 0.005). Les fœtus sains ont des contractions plus fréquentes.

— **light_decelerations** – Décélérations légères liées aux contractions. Histogramme décroissant (majorité à 0).

— **severe_decelerations** – Décélérations sévères (potentiellement pathologiques). Absentes dans ce jeu de données (toutes valeurs à 0).

— **prolongued_decelerations** – Décélérations >2 min. Rares (quelques valeurs à 0.001–0.003), exclusivement associées aux fœtus pathologiques (boxplot).

— **abnormal_short_term_variability** – Variabilité anormale à court terme (%). Distribution bimodale (50–75 et 20–45). Corrélée à la pathologie (outliers bas pour les "à risque").

— **mean_value_of_short_term_variability** – Moyenne de la variabilité à court terme. Pic à 1 (faible variabilité dominante). Médiane plus basse pour les pathologiques.

— **percentage_of_time_with_abnormal_long_term_va** – Pourcentage de temps en variabilité longue anormale. Pic à 0, mais séparation claire entre classes (boxplot).

— **mean_value_of_long_term_variability** – Moyenne de la variabilité longue. Pics à 0 et 9.

A.2 Métriques d'Histogramme (dérivées du CTG)

— **histogram_width/min/max** – Largeur, minimum et maximum de l'histogramme des fréquences cardiaques. Largeur majoritairement entre 0–150 bpm, min souvent à 50, max centré à 150 (distribution normale).

— **histogram_number_of_peaks/zeroes** – Nombre de pics/valeurs nulles. Peaks : 1–5 pics dominants. Zeroes : rares (pic à 0).

— **histogram_mode/mean/median/variance/tendency** – Statistiques descriptives de l'histogramme. Distributions similaires pour mode, moyenne et médiane.

Note : Les métriques d'histogramme résument la distribution des fréquences cardiaques issues des cardiotocogrammes (CTG). Elles permettent une analyse diagnostique simplifiée.

B Analyse exploratoire des données

B.1 Matrice de corrélation complète

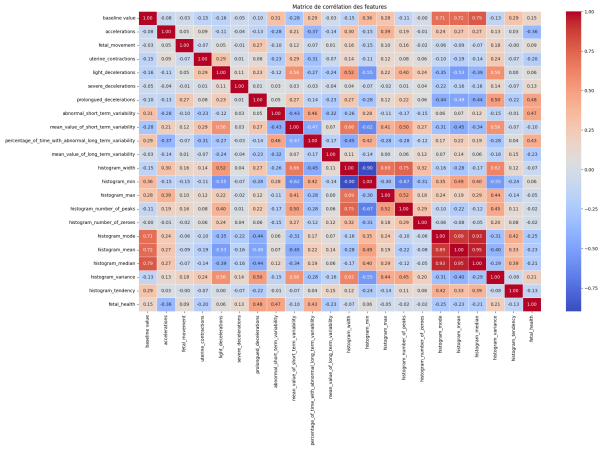

FIGURE 4 – Matrice de corrélation des features