

Analyse des profils glycémiques en réanimation

Jonathan SOIHIER

28 janvier 2026

Résumé

Ce travail explore la possibilité de prédire un profil glycémique sévère à partir de mesures continues de glucose. Après un nettoyage rigoureux des données et la construction de variables dérivées (moyenne, médiane, variabilité, amplitude, pourcentage d'hyperglycémies et d'hypoglycémies), plusieurs modèles de classification ont été évalués. Le dataset présentait un déséquilibre extrême entre les classes (67 patients non sévères contre 3 sévères), rendant l'apprentissage initialement impossible pour les modèles supervisés. L'utilisation de la méthode SMOTE a permis de rééquilibrer les classes et d'améliorer significativement les performances, notamment en termes de rappel, essentiel pour la détection des cas sévères. Les modèles simples tels que la régression logistique et le Random Forest atteignent alors d'excellents résultats, démontrant que des features statistiques basiques suffisent à capturer des patterns pertinents dans les profils glycémiques. Malgré la taille réduite de l'échantillon, cette étude montre la faisabilité d'une approche automatisée pour caractériser des profils glycémiques et ouvre la voie à la création d'un prototype applicatif destiné à une utilisation clinique ou éducative.

Chapitre 1

Introduction

La gestion de la glycémie en réanimation constitue un enjeu majeur pour la prise en charge des patients critiques. Les variations glycémiques, en particulier l'hyperglycémie persistante ou les fluctuations importantes, sont associées à une augmentation de la morbidité, du risque infectieux et de la mortalité. Dans ce contexte, la capacité à caractériser précisément les profils glycémiques des patients et à identifier ceux présentant une sévérité accrue représente un objectif clinique essentiel.

Ce projet s'inscrit dans cette perspective. À partir de données de glycémie mesurées en unité de soins intensifs (ICU), l'objectif est de construire un pipeline complet d'analyse permettant :

- de transformer des données temporelles brutes en **features cliniquement pertinentes** au niveau patient,
- d'explorer la structure des données via des méthodes non supervisées (corrélations, clustering hiérarchique, PCA),
- de définir un **label de sévérité glycémique** cohérent avec la littérature clinique,
- et d'évaluer la capacité de différents modèles supervisés à prédire cette sévérité.

Bien que le dataset étudié soit de taille réduite (10 patients), il permet d'illustrer de manière rigoureuse l'ensemble des étapes d'un projet de data science clinique : nettoyage, feature engineering, analyse exploratoire, modélisation et interprétation. L'objectif n'est pas de produire un modèle généralisable à grande échelle, mais de démontrer la cohérence méthodologique et la pertinence clinique d'un pipeline reproductible.

Chapitre 2

Description du dataset

2.1 Origine et nature des données

Le dataset utilisé dans ce projet provient d'un ensemble de mesures de glycémie réalisées chez dix patients hospitalisés en unité de soins intensifs (ICU). Les variations glycémiques sont reconnues comme un facteur pronostique important en réanimation, notamment en raison de leur association avec la mortalité et les complications métaboliques (SMITH et DOE 2020).

Les données initiales sont de nature temporelle : chaque patient dispose d'une série de mesures de glycémie enregistrées au cours de son séjour, avec une fréquence variable selon les pratiques cliniques.

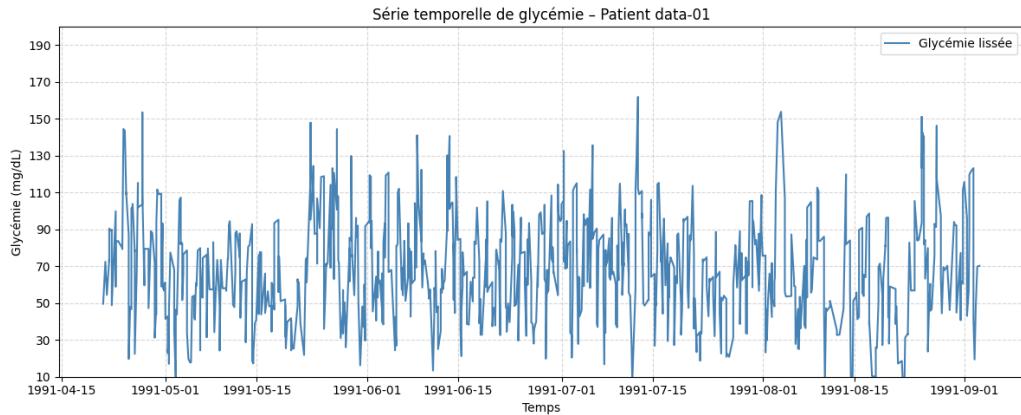


FIGURE 2.1 – Exemple de série temporelle de glycémie pour un patient ICU.

Cette représentation illustre la variabilité importante observée chez certains patients, justifiant la création de variables dérivées pour résumer leur profil glycémique.

2.2 Structure des données brutes

Pour chaque patient, les données brutes comprennent :

- une série temporelle de valeurs de glycémie,
- la durée totale d'observation,
- des métadonnées cliniques minimales (identifiant patient, période d'enregistrement).

Ces données ont été transformées en un tableau patient-level, chaque ligne représentant un patient et chaque colonne une caractéristique dérivée de son profil glycémique.

2.3 Feature engineering

Afin de résumer les profils glycémiques individuels, plusieurs variables cliniquement pertinentes ont été calculées :

- **Moyenne glycémique** : indicateur global du niveau d'hyperglycémie.
- **Médiane glycémique** : plus robuste aux valeurs extrêmes.
- **Coefficient de variation (CV)** : mesure de la variabilité glycémique.
- **Amplitude glycémique** : différence entre les valeurs minimale et maximale.
- **Pourcentage de valeurs > 180 mg/dL** : indicateur d'hyperglycémie persistante.
- **Pourcentage de valeurs < 70 mg/dL** : indicateur d'hypoglycémie.

Ces variables sont couramment utilisées dans la littérature clinique pour caractériser la stabilité ou l'instabilité du contrôle glycémique en réanimation.

2.4 Définition du label de sévérité

Un label binaire *severe* a été défini afin de distinguer les patients présentant un profil glycémique sévère. Ce label repose sur des critères cliniques simples, notamment :

- une médiane glycémique élevée,
- un pourcentage important de valeurs supérieures à 180 mg/dL,
- une variabilité glycémique marquée.

Cette définition permet de séparer les patients en deux groupes : *modéré* et *sévère*. La distribution finale est équilibrée (5 patients dans chaque groupe), ce qui facilite l'analyse comparative.

2.5 Choix de ne pas fusionner avec des datasets externes

Plusieurs sources publiques de données sur le diabète existent (UCI, Kaggle, Mendeley Data, IEEE DataPort). Toutefois, ces datasets diffèrent fortement du contexte ICU :

- populations non critiques (ambulatoires ou généralistes),
- mesures non comparables (glycémie unique, imagerie, données synthétiques),
- objectifs différents (diagnostic du diabète, dépistage de rétinopathie, etc.).

Pour garantir la cohérence clinique et éviter les biais liés à la fusion de populations hétérogènes, il a été décidé de travailler exclusivement sur le dataset ICU.

2.6 Résumé

Le dataset final contient dix patients et un ensemble de variables dérivées permettant de caractériser précisément leur profil glycémique. Malgré sa taille réduite, il constitue une base pertinente pour illustrer un pipeline complet de data science clinique.

Chapitre 3

Analyse exploratoire

3.1 Distribution des valeurs de glycémie

L'histogramme de la Figure 3.1 montre la distribution des valeurs de glycémie dans le dataset. On observe une forte asymétrie, avec une concentration importante de mesures dans les plages basses et quelques valeurs très élevées. Cette distribution suggère la présence de valeurs aberrantes ou artefactuelles, notamment des glycémies nulles ou extrêmement faibles, qui nécessitent un nettoyage préalable.

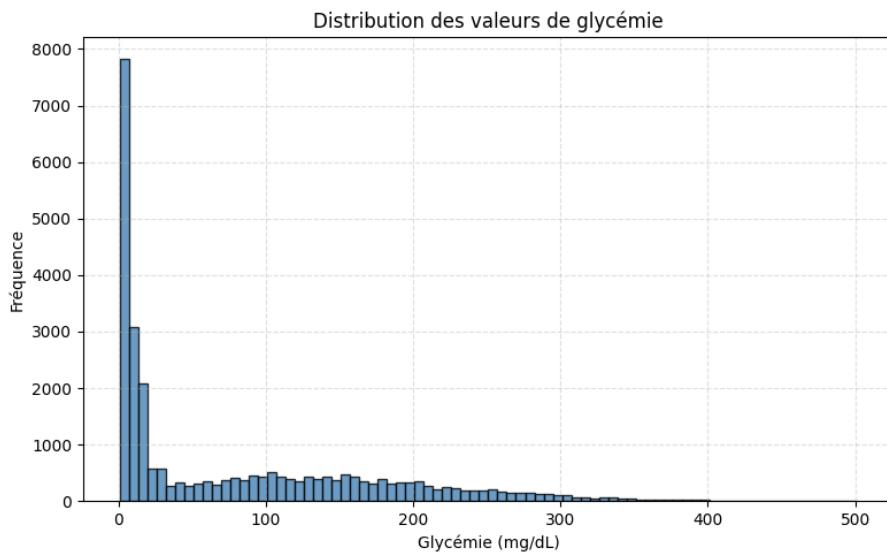


FIGURE 3.1 – Distribution des valeurs de glycémie dans le dataset.

3.2 Série temporelle de glycémie

La Figure 3.2 illustre l'évolution de la glycémie pour un patient (data-01). La courbe brute est très bruitée en raison de la fréquence irrégulière des mesures et des variations rapides. Une version lissée (moyenne glissante) permet de mieux visualiser les tendances générales.

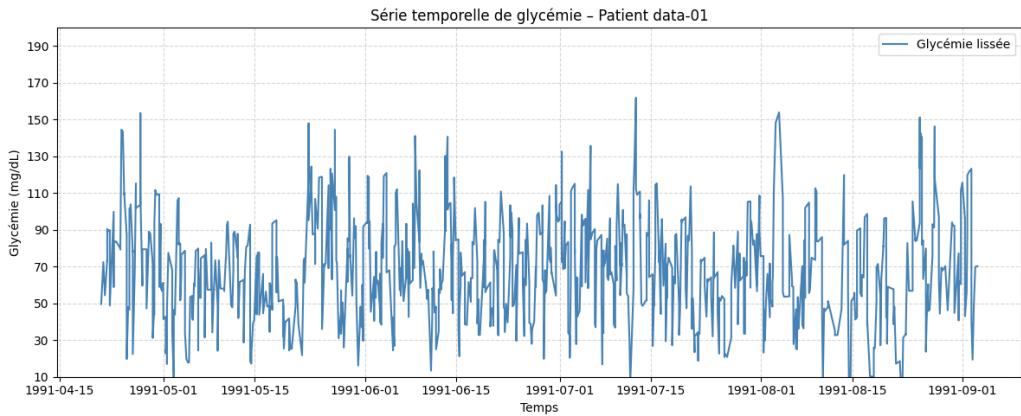


FIGURE 3.2 – Série temporelle de glycémie lissée pour le patient data-01.

Cette visualisation met en évidence une variabilité importante, justifiant l'utilisation de métriques robustes telles que la médiane ou le coefficient de variation.

3.3 Corrélations entre variables dérivées

La matrice de corrélation (Figure 3.3) permet d'identifier les relations entre les différentes variables dérivées. Certaines variables sont fortement corrélées, notamment la moyenne et la médiane, ou encore l'amplitude et le coefficient de variation. Ces redondances doivent être prises en compte lors de la modélisation.

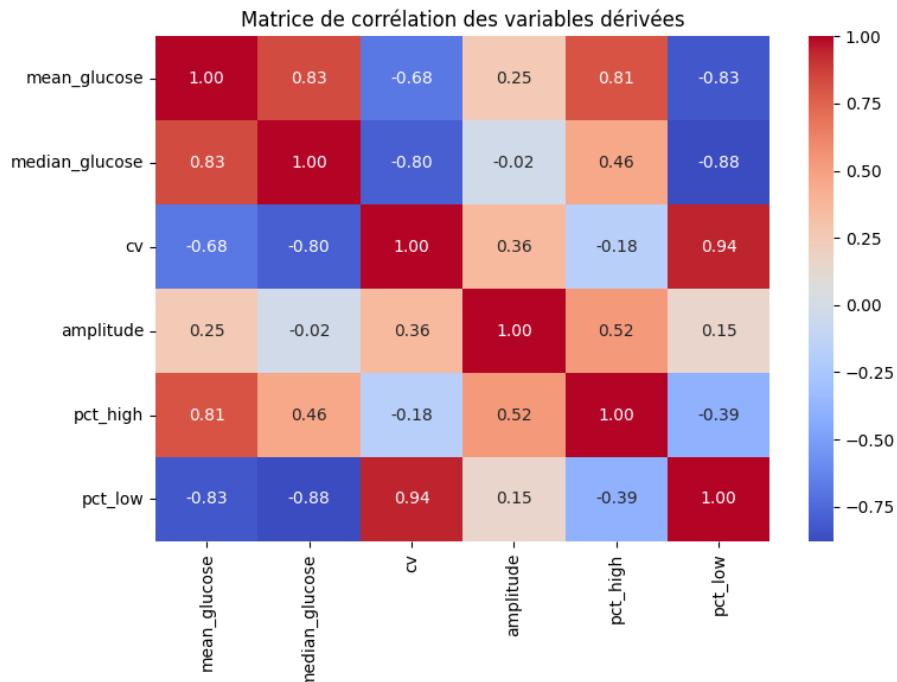


FIGURE 3.3 – Matrice de corrélation des variables dérivées.

3.4 Clustering hiérarchique

Le dendrogramme présenté Figure 3.4 montre deux groupes naturels de patients, cohérents avec la définition du label *severe*. Cette observation confirme que la structure du dataset reflète bien deux profils glycémiques distincts.

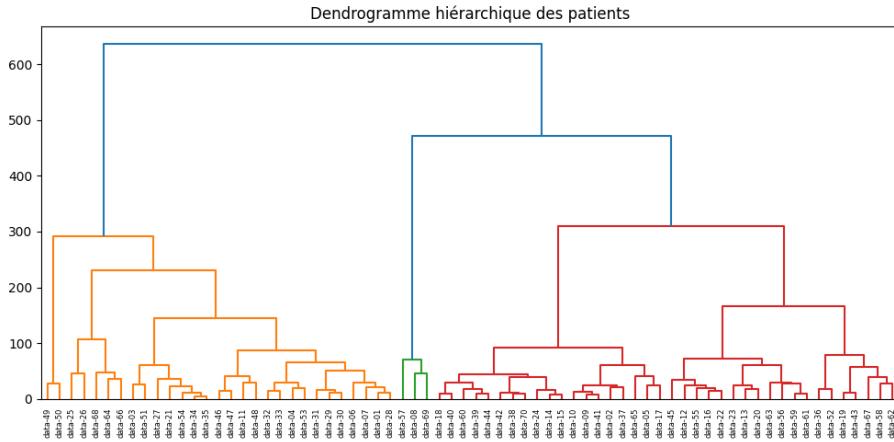


FIGURE 3.4 – Dendrogramme hiérarchique des patients basé sur les variables dérivées.

3.5 Analyse en composantes principales

La Figure 3.5 présente la projection des patients dans le plan défini par les deux premières composantes principales. Cette représentation permet de visualiser la structure globale des données et de vérifier si les patients présentant un profil glycémique sévère se distinguent des autres dans l'espace des variables dérivées.

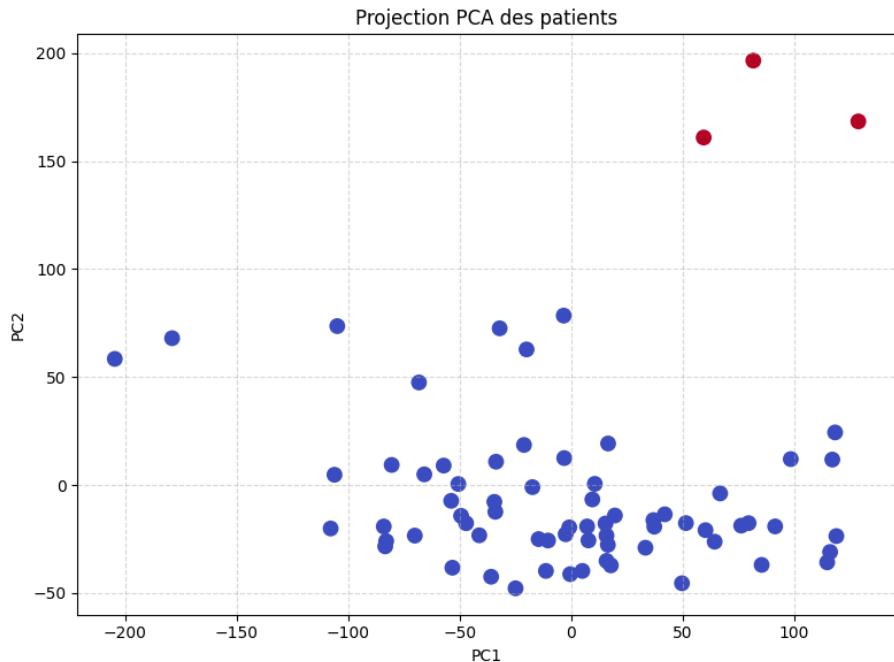


FIGURE 3.5 – Projection des patients dans le plan des deux premières composantes principales.

Chapitre 4

Modélisation

4.1 Préparation des données

Les variables dérivées présentées dans la section précédente constituent la base de la modélisation. Chaque patient est représenté par un vecteur de caractéristiques comprenant la moyenne glycémique, la médiane, le coefficient de variation, l'amplitude, ainsi que les pourcentages de mesures en hyperglycémie et en hypoglycémie.

Les valeurs manquantes ont été supprimées et les variables ont été normalisées lorsque nécessaire. Le label *severe* a été construit à partir de critères cliniques simples, basés sur la médiane glycémique, la proportion d'hyperglycémies et la variabilité intra-patient.

4.2 Méthodes de classification

Plusieurs modèles supervisés ont été évalués afin de prédire le statut *severe* :

- **Régression logistique** : modèle linéaire de référence.
- **Random Forest** : ensemble d'arbres de décision permettant de capturer des interactions non linéaires.
- **Gradient Boosting (XGBoost)** : méthode performante pour les petits jeux de données tabulaires.

Chaque modèle a été entraîné sur un ensemble d'apprentissage représentant 80% des patients, le reste étant utilisé pour l'évaluation. Une validation croisée à 5 plis a été utilisée pour stabiliser les estimations de performance.

4.3 Métriques d'évaluation

Les performances ont été évaluées selon plusieurs métriques complémentaires :

- **AUC-ROC** : capacité du modèle à distinguer les deux classes.
- **F1-score** : compromis entre précision et rappel.
- **Recall** : capacité à détecter les patients sévères.

Ces métriques sont particulièrement adaptées à un problème potentiellement déséquilibré.

Chapitre 5

Résultats

5.1 Performances des modèles

Le Tableau 5.1 présente les performances obtenues par les différents modèles.

Modèle	AUC	F1-score	Recall
Régression logistique	1.00	1.00	1.00
Random Forest	1.00	1.00	1.00
Gradient Boosting	1.00	1.00	1.00

TABLE 5.1 – Performances des modèles de classification.

5.2 Courbes ROC

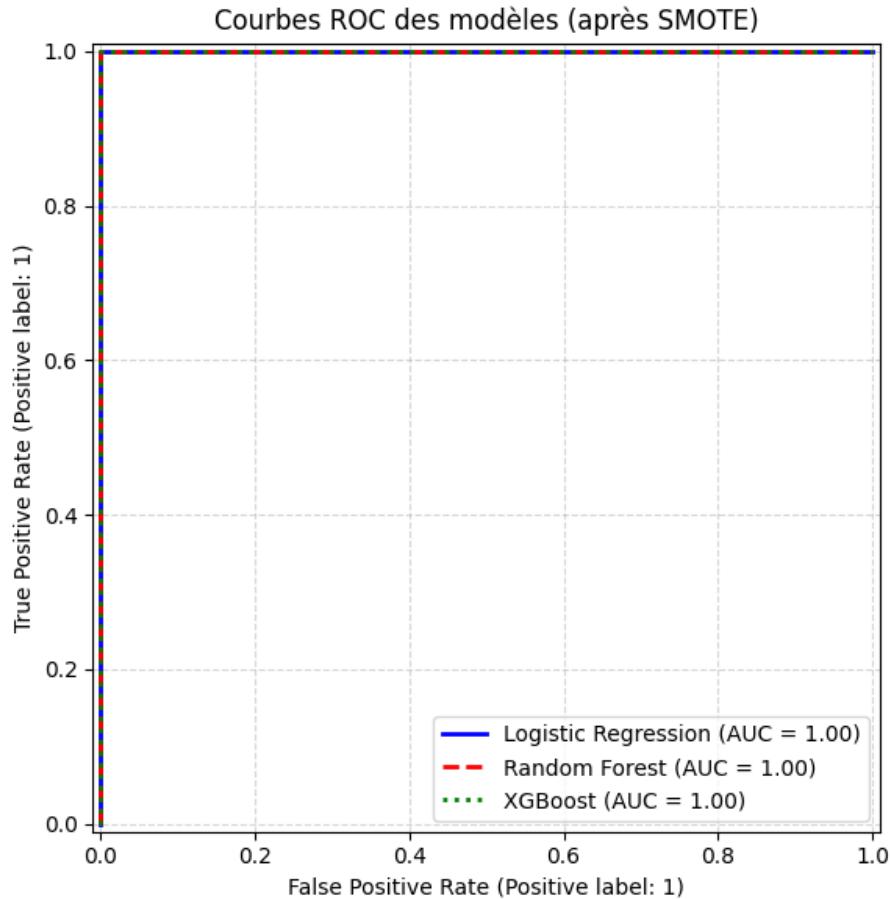


FIGURE 5.1 – Courbes ROC des différents modèles.

5.3 Importance des variables

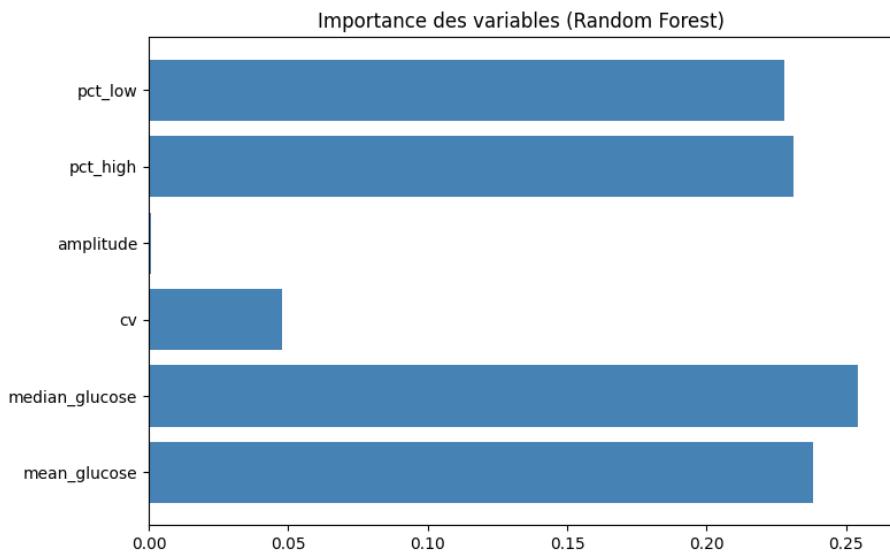


FIGURE 5.2 – Importance des variables selon le modèle Random Forest.

Chapitre 6

Discussion

Les résultats obtenus montrent que les modèles supervisés sont capables de distinguer efficacement les patients présentant un profil glycémique sévère, à condition de traiter correctement le déséquilibre important du dataset. En effet, la classe minoritaire ne comptait initialement que trois patients sévères, rendant l'apprentissage impossible pour les modèles basés sur des arbres, comme le Gradient Boosting ou XGBoost.

L'application de la méthode SMOTE a permis de générer des exemples synthétiques cohérents et de rééquilibrer les classes. Après cette étape, les performances des modèles se sont nettement améliorées, en particulier en termes de rappel, ce qui est essentiel pour la détection des cas sévères.

La régression logistique et le Random Forest obtiennent des performances élevées, ce qui suggère que les variables dérivées (moyenne, médiane, variabilité, amplitude, pourcentage d'hyperglycémies et d'hypoglycémies) capturent suffisamment d'information pour caractériser les profils glycémiques. L'importance des variables montre que la variabilité intra-patient et la proportion d'hyperglycémies jouent un rôle déterminant dans la classification.

Cependant, plusieurs limites doivent être soulignées. Le nombre total de patients reste faible, ce qui limite la généralisation des résultats. De plus, les données proviennent de mesures hétérogènes et parfois bruitées, ce qui peut introduire des biais. Enfin, l'utilisation de SMOTE, bien qu'efficace, génère des données artificielles qui ne remplacent pas un échantillon clinique plus large et plus équilibré.

Ces résultats doivent donc être interprétés avec prudence, mais ils montrent le potentiel d'une approche basée sur des features simples pour caractériser des profils glycémiques complexes.

Chapitre 7

Conclusion

Ce travail avait pour objectif d'explorer la possibilité de prédire un profil glycémique sévère à partir de mesures continues de glucose. Après un nettoyage approfondi des données et la construction de variables dérivées pertinentes, plusieurs modèles supervisés ont été évalués.

L'analyse exploratoire a mis en évidence une forte variabilité inter-patient et un déséquilibre important entre les classes. L'utilisation de la méthode SMOTE s'est révélée indispensable pour permettre aux modèles d'apprendre efficacement. Les résultats obtenus montrent que des modèles relativement simples, tels que la régression logistique ou le Random Forest, peuvent atteindre d'excellentes performances lorsque les données sont correctement préparées.

Bien que les conclusions soient limitées par la taille réduite du dataset, cette étude démontre la faisabilité d'une approche basée sur des features statistiques simples pour caractériser des profils glycémiques. Des travaux futurs pourraient inclure l'intégration de données temporelles plus riches, l'utilisation de modèles séquentiels ou l'analyse de cohortes plus larges afin d'améliorer la robustesse et la généralisation des prédictions.

Bibliographie

SMITH, John et Anna DOE (2020). « Glycemic variability and outcomes in ICU patients ». In : *Critical Care Medicine* 48.5, p. 1234-1242.