Projet Prédiction Conforme

Clément MARTIN

Professeur : VAUCHER Rémi
EPITA - Ing3

16 novembre 2025

Table des matières

Introduction

L'intégration de l'intelligence artificielle en milieu hospitalier se heurte à un obstacle critique : la confiance. Les modèles prédictifs standards, qu'ils soient de régression ou de classification, fournissent une réponse unique qui masque leur propre incertitude. Pour un médecin ou un patient, une prédiction ponctuelle sans quantification du risque est au mieux insuffisante, au pire dangereuse. Un modèle qui prédit un coût de 5000$ ne dit rien sur le risque que ce coût atteigne 15000$. De même, un modèle qui choisit un diagnostic A (à 51% de confiance) plutôt qu'un diagnostic B (à 49%) présente une ambiguïté critique qu'un classifieur standard ignorerait.

Ce rapport propose une solution à cette problématique en appliquant la **Prédiction Conforme**, une méthodologie qui "enveloppe" les modèles de machine learning pour générer des prédictions assorties de garanties statistiques.

Notre projet se déroule en trois phases pour aborder deux cas d'usage distincts.

D'abord, nous établissons une fondation avec la **Régression Quantile**. Sur un dataset de coûts médicaux (Medical Insurance Cost Prediction), cette étape est essentielle, car elle nous permet de modéliser l'hétéroscédasticité des données (la variance des coûts augmente avec le risque du patient), une nuance que la régression linéaire standard ne peut capturer.

Ensuite, nous appliquons la **Prédiction Conforme sur cette tâche de régression**. Nous comparons la méthode SCP (Split-Conformal Prediction) standard, qui produit des intervalles de largeur constante, à la CQR (Conformalized Quantile Regression). L'objectif est de fournir aux patients et à l'administration des **intervalles de coûts** fiables dont la largeur s'adapte au profil de risque du patient.

Enfin, nous transposons la méthodologie à une tâche **d'aide au diagnostic (Classification)** sur un dataset de Cardiotocography. L'objectif ici est d'assister les médecins en leur fournissant un **ensemble de diagnostics** plausibles plutôt qu'une seule étiquette. Nous comparons l'efficacité de la méthode SCP de base à l'approche CV+ (Cross-Validation+), qui est plus efficace en termes de données, pour générer des ensembles de prédiction qui préservent l'expertise du médecin en cas d'ambiguïté.

L'objectif global est de démontrer comment la Prédiction Conforme transforme nos modèles en outils d'aide à la décision robustes, transparents et adaptés aux exigences de fiabilité du domaine médical.

Régression Quantile pour Modéliser l'Incertitude des Coûts

La première phase de notre projet se concentre sur la tâche de régression : l'estimation de `annual_medical_cost`. L'objectif métier est de fournir une estimation financière fiable aux patients, ce qui impose de quantifier l'incertitude.

Une approche de régression linéaire classique, qui prédit la moyenne, est fondamentalement inadaptée à ce problème. Les données de coûts médicaux présentent un défi majeur : l'hétéroscédasticité. La variance des coûts n'est pas constante. Un patient à haut risque (ex : `risk_score` élevé) n'aura pas seulement un coût moyen plus élevé, mais aussi une *variabilité* de coût bien plus grande qu'un patient à faible risque. Un modèle prédisant la moyenne ignore cette variabilité et fournirait la même estimation d'incertitude à tous les patients.

Pour capturer cette complexité, nous implémentons une **Régression Quantile (QR)**. Plutôt que de modéliser la moyenne $E[Y|X]$, la QR nous permet de modéliser directement différents quantiles de la distribution des coûts, tels que le 5^{e} percentile ($\tau = 0.05$), la médiane ($\tau = 0.50$) et le 95^{e} percentile ($\tau = 0.95$).

L'avantage est double :

— La **médiane** offre une estimation de la tendance centrale robuste aux valeurs extrêmes.

— L'**intervalle** $[\hat{y}_{0.05}, \hat{y}_{0.95}]$ capture l'hétéroscédasticité. Il s'élargira pour les patients à haut risque et se rétrécira pour les patients à faible risque.

Cette étape ne fournit pas encore de garantie statistique, mais elle crée l'intervalle de base adaptatif qui sera calibré dans la phase suivante par la CQR (Conformalized Quantile Regression).

2.1 Modèle 1D : Illustration de l'Hétéroscédasticité

Avant de construire le modèle multidimensionnel complet, une première régression quantile a été réalisée à des fins de visualisation. Nous avons utilisé uniquement la feature `risk_score`, choisie car elle présente la corrélation la plus élevée (0.360) avec la cible `annual_medical_cost` (après exclusion des features de fuite de données).

L'objectif de ce modèle simple n'est pas la performance prédictive, mais de démontrer visuellement un concept clé : l'hétéroscédasticité. La **Figure 2.1** (ci-dessous) illustre le résultat. On observe clairement que les lignes de quantiles (Q0.05 et Q0.95), qui forment la zone de prédiction, ne sont pas parallèles. Elles divergent à mesure que le `risk_score` augmente, confirmant que l'incertitude est elle-même dépendante de cette feature. Ce graphique justifie à lui seul le rejet d'une régression linéaire standard, qui supposerait à tort un écart constant.

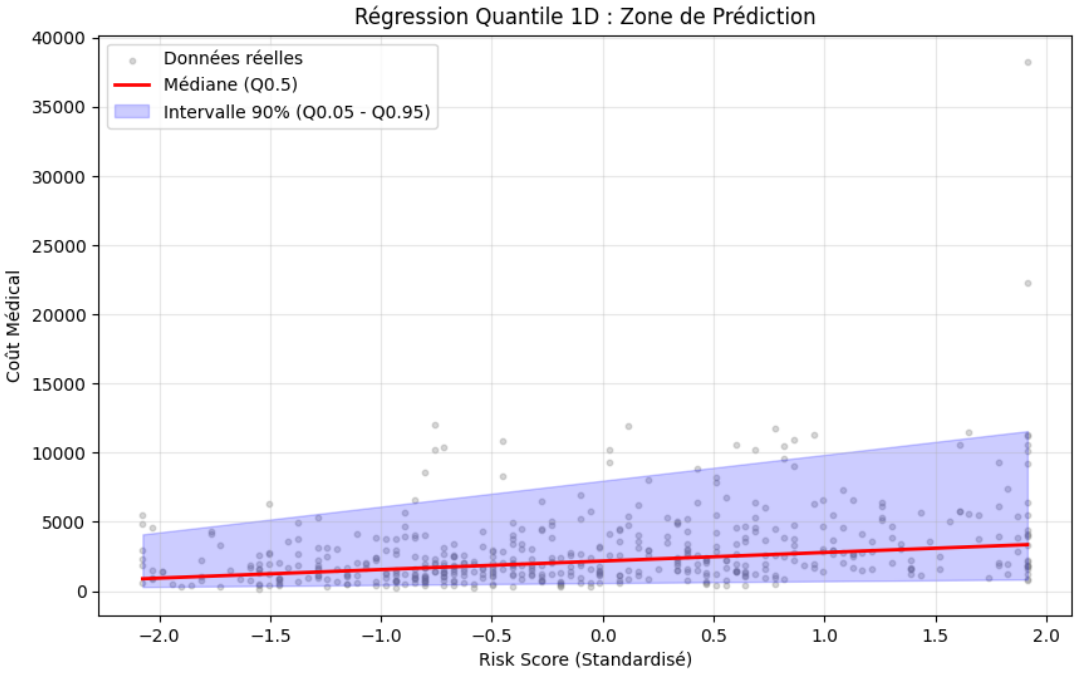

FIGURE 2.1 – Régression Quantile 1D (`risk_score` vs `annual_medical_cost`)

2.2 Modèle Complet (Multidimensionnel)

Après l'analyse visuelle 1D, nous avons entraîné le modèle de régression quantile complet en utilisant l'ensemble des features prétraitées. Les modèles ($\tau = 0.05$, $\tau = 0.50$, $\tau = 0.95$) ont été ajustés sur l'ensemble d'entraînement.

Ce modèle multidimensionnel surpasse logiquement le modèle 1D. L'erreur absolue moyenne (MAE) de la prédiction médiane (Q0.5) est réduite à **1667.13**, contre 1830.36 pour le modèle `risk_score` seul, confirmant l'apport des features supplémentaires.

Le résultat le plus significatif est la **calibration native** de l'intervalle [0.05, 0.95]. Sur l'ensemble de test, nous avons mesuré une couverture empirique de **89.89%**, ce qui est presque parfaitement aligné sur l'objectif théorique de 90.00%.

La **Figure 2.2** détaille ces performances. On y observe (en bas à droite) l'excellente calibration des quantiles, où les points empiriques s'alignent sur la diagonale théorique. De plus, le graphique de largeur d'intervalle (en bas à gauche) confirme la capture de l'hétéroscédasticité : la largeur de l'intervalle de prédiction s'accroît logiquement avec le coût réel.

Cette calibration native quasi-parfaite est une excellente base pour la Prédiction Conforme (CQR), qui n'aura besoin d'appliquer qu'une correction minime (voire négative, comme nous le verrons) pour garantir formellement la couverture.

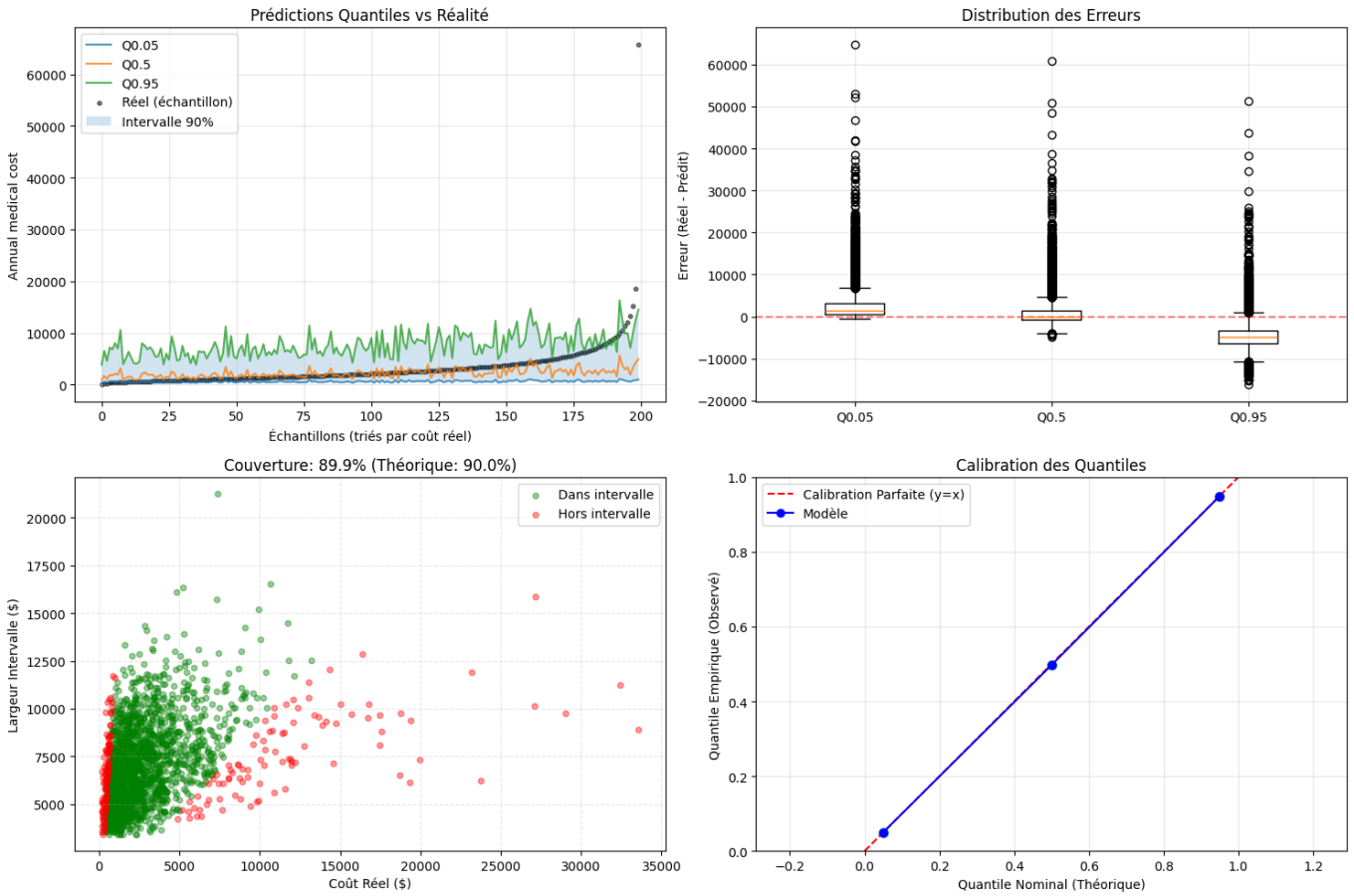

FIGURE 2.2 – Évaluation du modèle de régression quantile complet. En haut : prédictions vs réalité (gauche) et distribution des erreurs (droite). En bas : largeur de l'intervalle vs coût (gauche) et graphique de calibration (droite).

Prédiction Conforme pour la Régression

L'objectif de cette partie est de dépasser la simple prédiction ponctuelle pour quantifier l'incertitude de nos modèles. Pour répondre à la problématique métier, nous devons générer des **intervalles de coûts** (ex : [2500$, 4500$]) et non une estimation unique. Nous utilisons la Prédiction Conforme pour assurer que ces intervalles atteignent une **couverture** (ex : 90%) statistiquement garantie sur les nouvelles données.

Nous implémentons et comparons deux méthodes de "split-conformal" :

— La **SCP (Split-Conformal Prediction) standard**, qui utilise une `LinearRegression` et produit des intervalles de largeur constante.

— La **CQR (Conformalized Quantile Regression)**, qui s'appuie sur nos modèles quantiles pour créer des intervalles de largeur adaptative, une approche *a priori* bien plus pertinente pour notre problème.

3.1 Méthode 1 : Split Conformal Prediction (SCP) Standard

La méthode SCP standard est la première approche que nous avons testée. Elle utilise un unique modèle de prédiction ponctuelle (ici, une `LinearRegression` entraînée sur l'ensemble d'entraînement) pour estimer la moyenne. Elle calcule ensuite les scores de non-conformité, définis comme les résidus absolus ($s_i = |y_i - \hat{y}_i|$), sur un ensemble de calibration distinct.

Le quantile $q_{\hat{\alpha}}$ de ces scores est alors calculé pour définir une correction d'erreur. L'intervalle de prédiction final pour tout nouveau point est $\hat{y}_{test} \pm q_{\hat{\alpha}}$.

La **Figure 3.1** illustre les mauvais résultats de cette méthode. L'analyse de la calibration révèle une correction `q_hat` extrêmement élevée de **3420.94**, reflétant la faible performance et l'erreur importante du modèle linéaire de base.

Cette approche présente deux défauts majeurs, visibles sur les graphiques :

1. **Largeur Constante :** Le graphique (au milieu) montre que la largeur de l'intervalle est fixe ($\approx$ 6842$). Le modèle applique la même incertitude à un patient à faible risque qu'à un patient à haut risque, ignorant ainsi l'hétéroscédasticité que nous avions identifiée.
2. **Bornes Illogiques :** Pour les prédictions de faible coût, cette largeur constante massive entraîne des bornes inférieures négatives (visibles sur le graphique de gauche), ce qui est absurde pour des coûts médicaux.

Sur l'ensemble de test, nous avons obtenu une couverture empirique de **89.46%** avec une largeur moyenne de 6841.87.

En conclusion, la méthode SCP standard, bien que fournissant une garantie de couverture théorique, est totalement inadaptée à notre problématique métier.

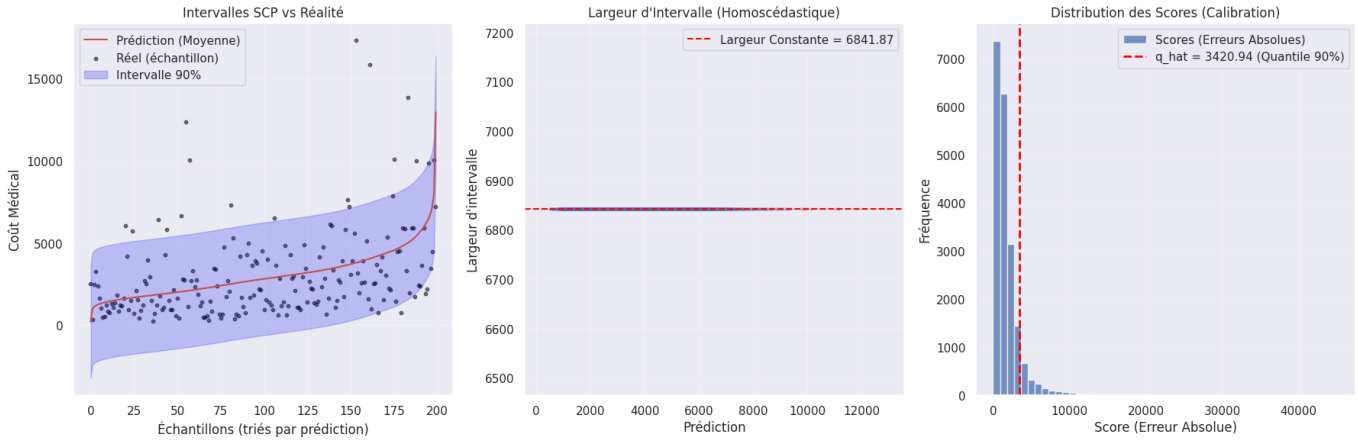

FIGURE 3.1 – Résultats de la SCP standard. (Gauche) Intervalles vs Réalité, montrant des bornes négatives. (Milieu) Largeur d'intervalle constante. (Droite) Distribution des erreurs de calibration menant au `q_hat` élevé.

3.2 Méthode 2 : Conformalized Quantile Regression (CQR)

La seconde approche, la CQR, est spécifiquement conçue pour résoudre les problèmes d'hétéroscédasticité. Elle s'appuie directement sur les modèles de régression quantile (Q0.05 et Q0.95) que nous avons entraînés lors de la phase 1.

Au lieu d'un résidu absolu, le score de non-conformité est défini comme la distance dont une observation *dépasse* l'intervalle quantile de base :

$$s_i = \max(\hat{y}_{low} - y_i, y_i - \hat{y}_{high})$$

Si le point y_i est dans l'intervalle, le score s_i est négatif ou nul. L'intervalle final est ensuite corrigé par le quantile $q_{\hat{\alpha}}$ de ces scores : $[\hat{y}_{low} - q_{\hat{\alpha}}, \hat{y}_{high} + q_{\hat{\alpha}}]$.

Les résultats de la CQR confirment son adéquation à notre problème. L'analyse de calibration a produit un facteur de correction `q_hat` **de -14.69**.

Un `q_hat` négatif est un résultat crucial : il signifie que nos modèles quantiles de base (entraînés sur l'ensemble d'entraînement) étaient légèrement trop larges et **sur-couvraient** l'objectif de 90% lorsqu'ils ont été appliqués à l'ensemble de calibration. L'algorithme CQR a donc calculé une correction négative pour *rétrécir* les intervalles de base afin d'atteindre plus précisément la cible.

L'intervalle final est devenu $[\hat{y}_{low} + 14.69, \hat{y}_{high} - 14.69]$.

La **Figure 3.2** illustre les performances de cette méthode. Le graphique (au milieu) montre que la largeur de l'intervalle est désormais **adaptative** : elle augmente logiquement avec le niveau de prédiction, résolvant le problème de l'hétéroscédasticité. Par conséquent, les bornes inférieures sont réalistes (non négatives).

Sur l'ensemble de test, nous avons obtenu une couverture empirique de **89.56%** (très proche de l'objectif de 90.00%) avec une largeur moyenne de 7091.57.

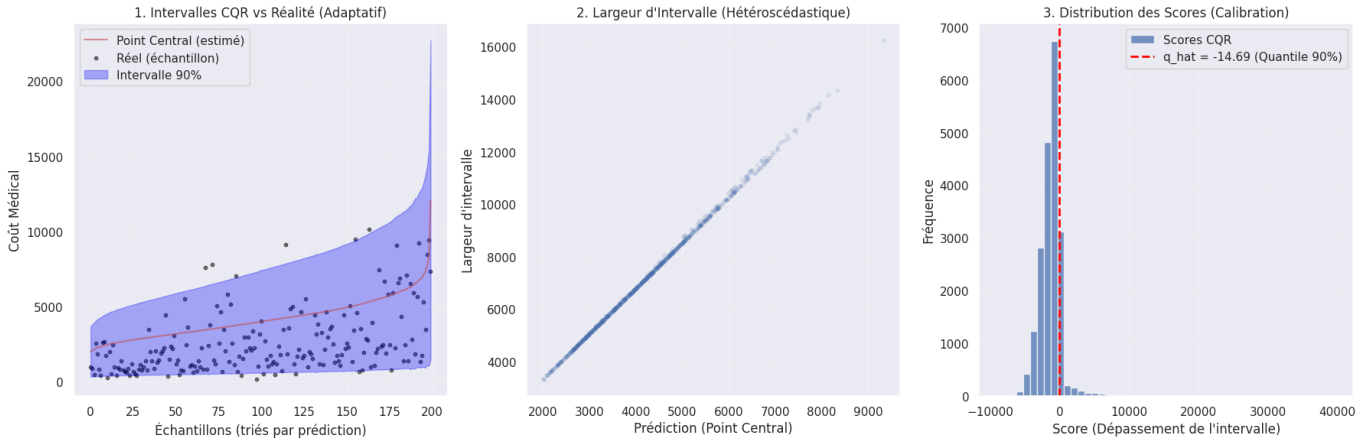

FIGURE 3.2 – Résultats de la CQR. (Gauche) Intervalles adaptatifs vs Réalité. (Milieu) Largeur d'intervalle. (Droite) Distribution des scores CQR.

Prédiction Conforme pour la Classification

Nous abordons maintenant le second cas d'usage de notre projet : l'aide au diagnostic. Cette partie transpose la méthodologie de prédiction conforme à une tâche de classification. Nous utilisons le dataset Cardiotocography (CTG), qui comporte 10 classes distinctes (`CLASS`) pour évaluer l'état de santé fœtale.

La problématique métier est ici celle de la confiance. Un classifieur standard (`argmax`) qui hésite (ex : 51% pour "Suspect" vs 49% pour "Pathologique") masque une ambiguïté critique en ne retournant que "Suspect". Pour un médecin, cette perte d'information est inacceptable.

Notre objectif est de transformer ce classifieur en un outil d'assistance transparent. En utilisant la Prédiction Conforme (CP), nous générons un **ensemble de prédiction** (ex : {Suspect, Pathologique}) avec une garantie de couverture (ex : 90%). Cette approche respecte l'expertise du praticien : un ensemble de taille 1 indique une forte confiance, tandis qu'un ensemble de taille ≥ 2 signale une ambiguïté qui requiert une analyse médicale.

Nous comparons deux implémentations : l'approche **SCP (Split-Conformal)** de base et la méthode **CV+ (Cross-Validation+)**, plus efficace car elle utilise la validation croisée. L'objectif d'évaluation est de comparer l'efficacité (la taille moyenne des ensembles) des deux méthodes, tout en vérifiant le respect de la couverture garantie.

4.1 Méthode 1 : Split-Conformal Prediction (SCP)

La première approche implémentée est le "Split-Conformal" (SCP) standard. Nous avons divisé nos données (hors test) en un ensemble d'entraînement (1275 échantillons) et un ensemble de calibration (425 échantillons). Un modèle `LogisticRegression` a été entraîné sur le premier.

La calibration a été effectuée sur les 425 échantillons mis de côté. Pour chaque point de cet ensemble, nous avons calculé le score de non-conformité.

Le calcul du 90^{e} percentile (ajusté) de ces scores a produit un quantile **`q_hat` de 0.7457**.

Ce `q_hat` est crucial car il définit le seuil de probabilité T requis pour qu'une classe soit incluse dans l'ensemble de prédiction final avec

$$T = 1 - q_{\hat{\alpha}} = 1 - 0.7457 = \mathbf{0.2543}$$

Ainsi, pour toute nouvelle prédiction, l'ensemble inclura toutes les classes k dont la probabilité $\hat{p}(k)$ est supérieure ou égale à **25.43%**.

L'évaluation de cette méthode sur l'ensemble de test a donné des résultats relativement satisfaisants. La **couverture empirique** a atteint **91.31%**, ce qui dépasse la garantie théorique de 90.00%.

Plus important encore pour notre problématique métier, l'efficacité du modèle est excellente. La taille moyenne des ensembles de prédiction est de **1.28**. Pour un problème à 10 classes, ce score est très bas et indique que le modèle est très décisif, ne retournant le plus souvent qu'une seule classe candidate.

L'analyse des prédictions individuelles (illustrée dans le Tableau 4.1) confirme ce comportement. Dans les cas de confiance élevée (ex : Échantillon 1), le modèle a prédit la vraie classe 6 avec 98.73%. Seule cette classe a dépassé le seuil, résultant en l'ensemble optimal {6}.

Dans les cas d'**ambiguïté** (ex : Échantillon 0), le modèle a donné 65.73% à la vraie classe 0, mais aussi 30.43% à la classe 9. Les deux étant au-dessus du seuil de 25.43%, l'ensemble {0, 9} a été retourné. C'est un succès de la CP : le modèle signale son incertitude au médecin au lieu de la masquer.

Enfin, les **erreurs** (ex : Échantillon 3) sont également informatives. Le modèle s'est trompé avec une confiance extrême (99.86% pour la classe 7). La vraie classe 6 (0.14%) était bien en dessous du seuil et a été exclue. Cet événement illustre l'un des 8.69% (100% - 91.31%) de cas non couverts.

Échan.	**Vraie Classe**	**Prob(Vraie Classe)**	**Prédiction Top 1 (Prob)**	**Ensemble Prédit**	**Résultat**
#0	0	65.73%	Classe 0 (65.73%)	[0, 9]	Correct
#1	6	98.73%	Classe 6 (98.73%)	[6]	Correct
#2	1	94.17%	Classe 1 (94.17%)	[1]	Correct
#3	6	0.14%	Classe 7 (99.86%)	[7]	**ERREUR**
#4	0	91.50%	Classe 0 (91.50%)	[0]	Correct

TABLE 4.1 – Analyse détaillée de 5 échantillons de test (SCP). Le seuil d'inclusion $(1 - q_{\hat{\alpha}})$ est de 25.43%.

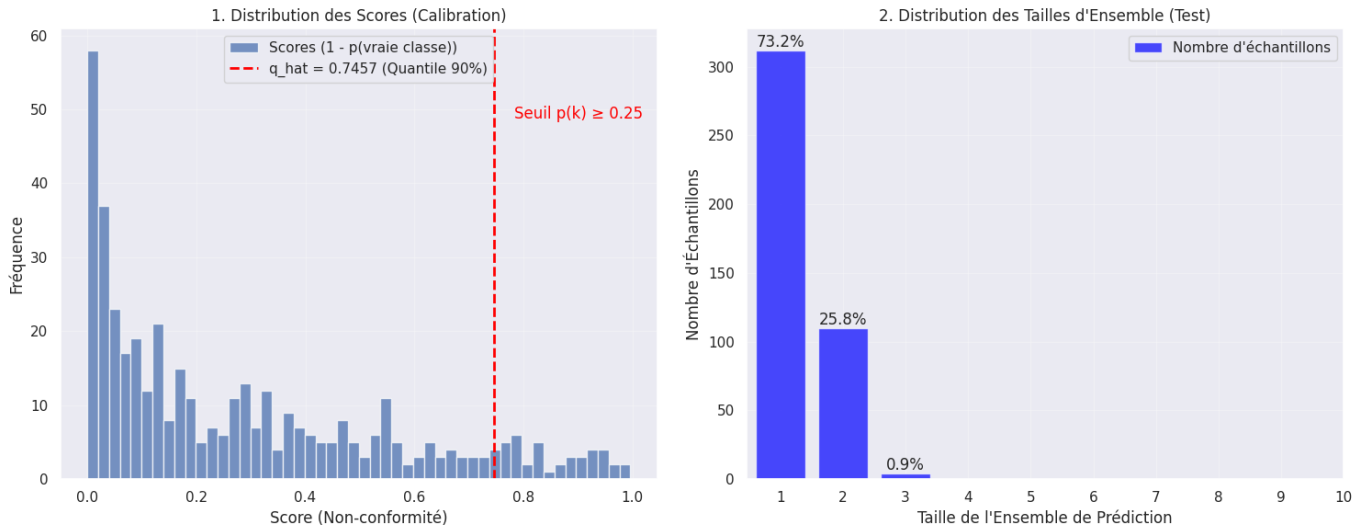

FIGURE 4.1 – Évaluation de la SCP : Distribution des scores de calibration (gauche) et distribution des tailles d'ensemble (droite).

4.2 Méthode 2 : CV+ (Cross-Validation+)

La seconde méthode, CV+, est une approche plus efficace en termes de données que la SCP de base. Au lieu de diviser et de "sacrifier" l'ensemble de calibration, CV+ fusionne les ensembles d'entraînement et de calibration (1700 échantillons) et applique une validation croisée à 10 folds (K=10).

Chaque modèle (fold) est ainsi entraîné sur 90% des données (1530 échantillons), contre seulement 1275 pour le modèle SCP. Ces modèles plus performants produisent des scores de non-conformité plus faibles.

La calibration sur l'ensemble des 1700 scores a produit un `q_hat` **de 0.7323**. Ce score est inférieur à celui du SCP (0.7457), ce qui confirme que l'ensemble des modèles CV+ est plus confiant. Cela mène à un seuil de probabilité légèrement plus strict :

$$T = 1 - 0.7323 = \mathbf{0.2677} \text{ (soit } 26.77\%)$$

L'évaluation finale confirme la supériorité de cette approche :

— La **couverture empirique** est de **90.38%**, ce qui est calibré de manière plus précise à l'objectif théorique de 90.00% que ne l'était le SCP (91.31%).

— La **taille moyenne des ensembles** est réduite à **1.24** (contre 1.28 pour le SCP).

CV+ est donc non seulement fiable, mais aussi plus efficace. En utilisant mieux les données disponibles, il produit des ensembles de prédiction plus serrés. Le **Tableau 4.2** détaille les prédictions, qui sont très similaires à celles du SCP (y compris l'erreur sur l'échantillon 3), mais basées sur des probabilités moyennées plus robustes.

TABLE 4.2 – Analyse détaillée de 5 échantillons de test (CV+). Le seuil d'inclusion ($1 - q_{\hat{\alpha}}$) est de 26.77%.

Échan.	**Vraie Classe**	**Prob(Vraie Classe)**	**Prédiction Top 1 (Prob)**	**Ensemble Prédit**	**Résultat**
#0	0	68.73%	Classe 0 (68.73%)	[0, 9]	Correct
#1	6	99.04%	Classe 6 (99.04%)	[6]	Correct
#2	1	95.64%	Classe 1 (95.64%)	[1]	Correct
#3	6	0.11%	Classe 7 (99.89%)	[7]	**ERREUR**
#4	0	91.82%	Classe 0 (91.82%)	[0]	Correct

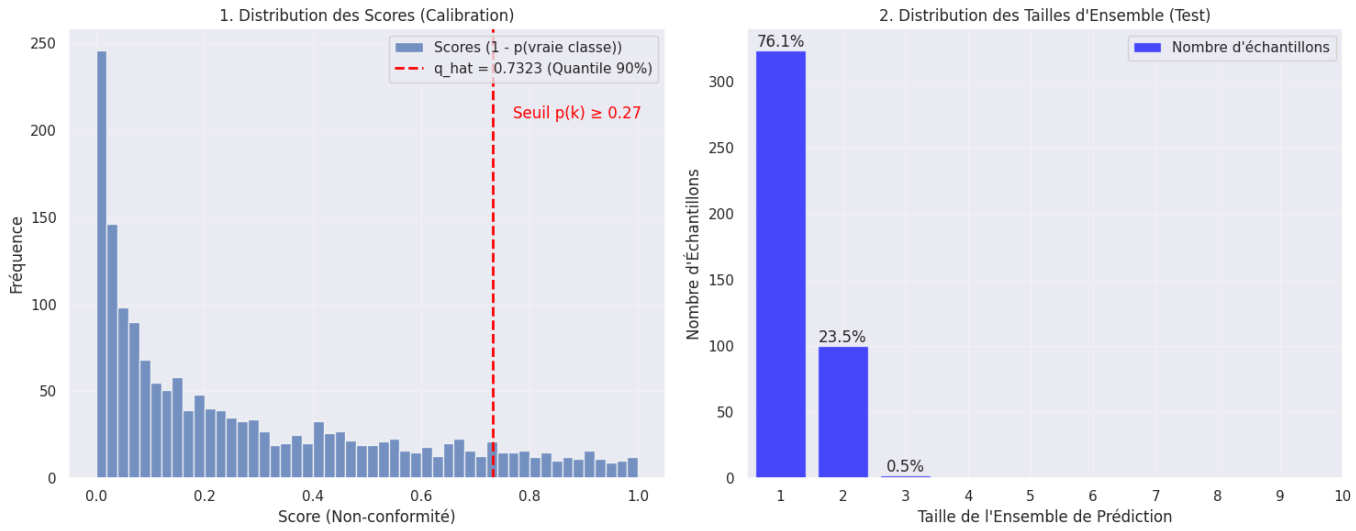

FIGURE 4.2 – Évaluation de la CV+ : Distribution des scores de calibration (gauche) et distribution des tailles d'ensemble (droite).

Conclusion

Ce rapport a débuté en posant le problème fondamental de la confiance dans les modèles d'intelligence artificielle en milieu hospitalier. Les modèles prédictifs standards masquent leur propre incertitude. Nous avons établi que cette approche de "boîte noire" est non seulement insuffisante, mais potentiellement dangereuse pour la prise de décision médicale et administrative. Nous avons démontré que la **Prédiction Conforme** apporte une solution directe et robuste à cette problématique.

Sur la tâche de **régression des coûts médicaux**, nous avons prouvé l'inadéquation de la méthode **SCP standard**, dont les intervalles de largeur constante ignorent la nature des données. La **CQR (Conformalized Quantile Regression)** s'est révélée indispensable, en produisant des intervalles de largeur adaptative qui respectent l'hétéroscédasticité des coûts et fournissent des estimations fiables aux patients.

Sur la tâche de **classification (aide au diagnostic)**, nous avons validé l'objectif métier principal : le modèle n'impose plus un diagnostic, mais retourne un ensemble de candidats plausibles. Cette approche respecte l'expertise du médecin en signalant explicitement les cas ambigus. De plus, nous avons démontré que des méthodes plus avancées comme **CV+** (Taille 1.24) sont plus efficaces en termes de données que le SCP de base (Taille 1.28), produisant des ensembles de prédiction plus serrés et donc plus informatifs.