Projet - UE Apprentissage statistique II

Dylan BERGER - Nicolas DECOOPMAN - Richard SELVARADJOU

Introduction

L'augmentation des résistances bactériennes aux antibiotiques représente un défi majeur en santé publique. Les méthodes diagnostiques basées sur la culture en laboratoire restent souvent longues et coûteuses. L'émergence de technologies permettant de prédire si une souche bactérienne est résistante ou sensible à un antibiotique à partir de données génomiques et transcriptomiques ouvre des perspectives prometteuses pour améliorer le diagnostic et le traitement. Avec l'accessibilité croissante du séquençage génomique et la réduction continue des coûts, ces approches deviennent de plus en plus envisageables à grande échelle. Ce projet s'inscrit pleinement dans cette problématique et se concentre sur la classification binaire des phénotypes de résistance.

L'intégration de données riches, mêlant variables binaires et quantitatives, permet d'analyser les bactéries sous divers angles. Cependant, cette complexité pose des défis importants en termes de prétraitement des données et de modélisation. L'objectif de cette étude est double : (i) évaluer les performances prédictives des modèles pour chacun des cinq antibiotiques étudiés ; et (ii) identifier les sources d'information les plus pertinentes pour prédire la résistance à chaque antibiotique.

Pour atteindre ces objectifs, nous avons adopté une approche rigoureuse combinant exploration des données, sélection de modèles, optimisation des hyperparamètres et validation croisée. Cette démarche vise à maximiser la robustesse et la pertinence des résultats obtenus.

Méthodes

Structure des données

L'étude porte sur 414 bactéries et s'appuie sur des données relatives aux phénotypes de sensibilité (données binaires sensible/résistant) à cinq antibiotiques (`pheno`) : la tobramycine (TOB), un aminoside ; la ceftazidime (CAZ), une céphalosporine de 3 génération (3G) ; la ciprofloxacine (CIP), une fluoroquinolone ; le méropenème (MER), un carbapénème ; et la colistine (COL), une polymyxine. Cette matrice contient 204 valeurs manquantes. La répartition pour chaque antibiotique de chaque classe est résumée dans la figure 1.

Le jeu de données étudié repose sur une approche intégrative combinant trois sources d'informations complémentaires : 1) la présence ou l'absence de **gènes de résistance** (matrice `X_gpa` de 16005 gènes), souvent associés directement à des mécanismes de résistance ; 2) la présence ou l'absence de **mutations ponctuelles** (SNPs) dans certains gènes (matrice `X_snps` de 72236 SNPs), susceptibles d'altérer leurs

fonctions ; 3) les niveaux d'**expression génique** (matrice `X_genexp` de 6026 gènes), reflétant l'état cellulaire des bactéries. Ces matrices ne contiennent pas de données manquantes.

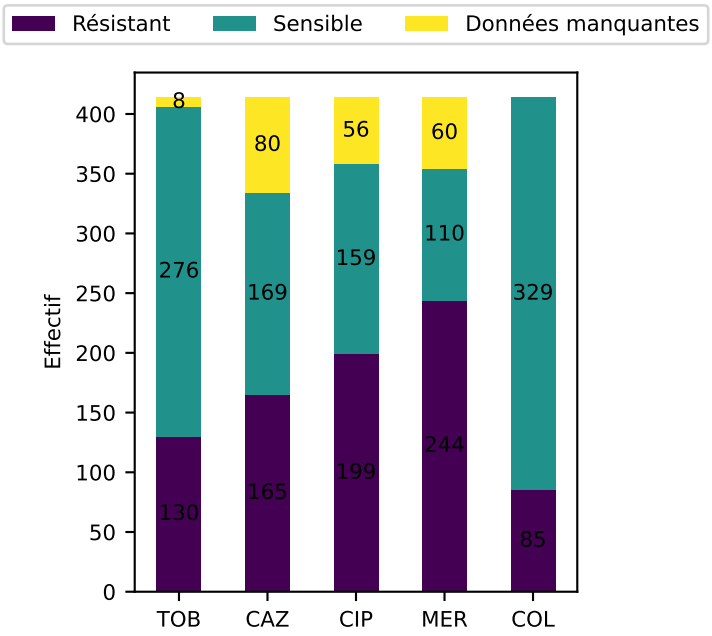

Figure 1: Répartition des classes par antibiotique

Données d'entrée

Une pré-analyse exploratoire des matrices `X_gpa`, `X_snps` et `X_genexp` confirme un problème de grande dimension, avec de potentielles variables redondantes ou inutiles. Pour évaluer leur pertinence, nous utilisons le test du Chi2 pour `X_gpa` et `X_snps` et une ANOVA pour `X_genexp`, en ajustant les p-valeurs via la méthode de Benjamini-Hochberg (FDR) pour limiter les faux positifs à $\alpha = .05$. Ce seuil favorise l'identification de caractéristiques potentiellement informatives tout en maîtrisant le risque d'erreur. Les caractéristiques significatives pour chaque antibiotique et matrice sont répertoriées avec les indices des bactéries valides (sans données manquantes). Les résultats, illustrés dans la figure 3 (modèle 1), montrent que l'importance des différents types de données varie considérablement selon l'antibiotique, reflétant potentiellement des mécanismes biologiques spécifiques de résistance ou de sensibilité. Cette observation souligne la nécessité d'une analyse multi-échelle et adaptative pour chaque antibiotique.

Cependant, pour valider ces observations et tester si la réduction dimensionnelle entraîne une perte d'information, nous utilisons également toutes les colonnes des matrices `X_gpa`, `X_snps` et `X_genexp` lors de l'entraînement des modèles. Cette approche comparative permet de déterminer si les modèles atteignent des performances similaires et d'évaluer l'impact réel des caractéristiques jugées non significatives lors de l'analyse exploratoire. Les matrices ont ainsi été utilisées individuellement pour examiner leur contribution relative à la prédiction des phénotypes de résistance.

Pour les modèles qui ne sélectionnent pas directement des colonnes, mais attribuent un score d'importance à chaque variable (comme Random Forest et XGBoost), nous avons fixé un seuil stricte-

ment positif pour l'importance des colonnes. Les caractéristiques avec un score d'importance égal à 0 ont ainsi été éliminées, ce qui permet de concentrer l'analyse sur les variables jugées contributives par ces modèles. Cette approche garantit une comparaison cohérente avec les méthodes basées sur une sélection explicite des variables.

Comparer les caractéristiques sélectionnées par chaque modèle révèle les variables biologiques ou génétiques liées à la résistance. L'accord entre modèles sur certaines variables renforce leur pertinence dans les mécanismes de résistance bactérienne. Les plages des indices sélectionnés par les modèles, ou hot spots génétiques, ont été étudiées afin de détecter les zones d'intérêt communes dans `X_gpa`, `X_snps` et `X_genexp`.

Prédiction de la résistance

Les données d'entrée ont été divisées en deux ensembles : 80 % des données ont été utilisées pour l'entraînement et l'ajustement des modèles, et 20 % pour l'évaluation finale. Une standardisation a été appliquée aux données d'expression génétique `X_genexp` pour harmoniser les échelles et optimiser la convergence des modèles. Une validation croisée à 5 plis a été appliquée sur l'ensemble d'entraînement afin de sélectionner les meilleurs hyperparamètres pour chaque classe de modèle. Pour optimiser ces hyperparamètres, nous avons utilisé une recherche par grille (`GridSearch`) sur un espace prédéfini de valeurs possibles, adaptée à chaque modèle et à chaque antibiotique. Cette approche systématique garantit que les configurations optimales des modèles sont identifiées pour maximiser leurs performances.

Modèles utilisés

L'utilisation d'une diversité de modèles permet de maximiser les chances d'identifier l'approche la plus performante pour prédire la résistance bactérienne, tout en exploitant les forces spécifiques de chaque méthode. Les modèles, du package `scikit-learn`, ont été sélectionné pour leurs capacités à gérer des données multidimensionnelles et leurs performances reconnues en classification.

1. **LGR FDR** (*Logistic Regression with False Discovery Rate*)
   Cette méthode réduit la dimensionnalité en sélectionnant uniquement les variables pertinentes via le contrôle du False Discovery Rate (FDR) par la méthode de Benjamini-Hochberg. Cela garantit que seules les caractéristiques les plus informatives sont utilisées, tout en maîtrisant le risque de faux positifs dans un contexte de grande dimension.
2. **LGR L1** (*Logistic Regression with Lasso Regularization*)
   La régularisation L1 (Lasso) permet de sélectionner directement les variables significatives en attribuant des coefficients nuls aux caractéristiques non pertinentes. Cette méthode favorise la parcimonie et simplifie l'interprétation en réduisant la complexité du modèle.
3. **Random Forest et XGBoost**
   Ces deux modèles, basés sur des arbres de décision, offrent une mesure d'importance des caractéristiques. Cette capacité permet d'identifier les variables les plus influentes sur les prédictions. Random Forest est robuste et adapté à la détection d'interactions non linéaires entre les variables. XGBoost est plus performant dans des contextes complexes et combine une précision élevée et une gestion efficace des interactions dans des ensembles de données volumineux.

4. **SVM** (*Support Vector Machines*)
   Le SVM est particulièrement performant pour capturer les relations complexes et les marges entre les classes, notamment en présence de données de grande dimension. Cette méthode est utile pour évaluer la structure sans dépendre d'une réduction de la dimensionnalité.

5. **MLP** (*Multilayer Perceptron*)
   Les réseaux de neurones profonds (MLP) sont capables de capturer des relations non linéaires complexes qui peuvent échapper aux autres modèles. Bien qu'ils soient moins directement interprétables, leur flexibilité et leur puissance d'apprentissage en font un outil précieux dans la détection de patterns complexes. Dans ces modèles on choisit de ne garder que les 1000 colonnes qui contribuent afin de garder une visualisation correcte. Sinon presque la totalité des colonnes étaient selectionnées par le modèle.

Evaluation des modèles

En comparant les performances des différents modèles, nous identifions celui qui est le plus fiable pour prédire la résistance bactérienne. La performance des modèles a été mesurée en utilisant le recall de `scikit-learn` (Recall $= \frac{\text{Vrais positifs}}{\text{Vrais positifs+Faux négatifs}}$), pour que les résultats soient facilement comparables entre antibiotiques au vu des classes déséquilibrées (figure 1).

Résultats

Génotype

Dans la figure 2, les deux premiers graphiques (a et b) mettent en évidence une variabilité génétique parmi les bactéries, probablement due à des différences dans leurs génomes ou leurs environnements. La majorité des bactéries ont entre 3500 et 4500 gènes et 14000 SNPs. Le troisième graphique montre que les données d'expression génique sont équilibrées et bien prétraitées pour une analyse statistique. Les bactéries ayant un nombre exceptionnellement faible de gènes ou de SNPs méritent une attention particulière, car elles pourraient représenter des cas biologiques spécifiques ou des artefacts des données.

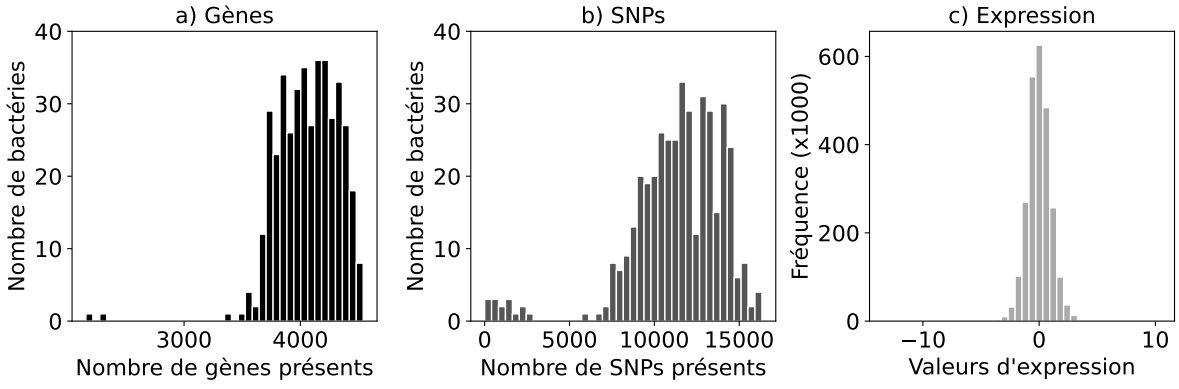

Figure 2: Analyse des distributions des données bactériologiques

Recall (fig. 3)

Les performances des modèles, évaluées à travers le recall (fig. 3), révèlent des disparités notables selon les antibiotiques et les approches algorithmiques. **TOB** se distingue avec un recall moyen de **0,92**, démontrant une capacité quasi-exemplaire des modèles à identifier les résistances. **CIP** affiche également de bonnes performances (**0,77**), traduisant une détection fiable des résistances pour cet antibiotique. En revanche, **CAZ** et **MER** présentent un recall moyen (respectivement **0,70** et **0,69**), tandis que **COL**, avec seulement **0,63**, est l'antibiotique le plus complexe à prédire, soulignant des limites communes aux modèles sur cette molécule. Cependant, les modèles **SVM** et/ou **MLP** font parfois considérablement chuter la moyenne des antibiotiques ayant un mauvais recall.

Au niveau des modèles, les résultats confirment la supériorité des approches linéaires et boostées. Les modèles **LGR L1** (**0,85**) et et **XGBoost** (**0,81**) se démarquent par leur robustesse et leur efficacité globale, suivis par **LGR FDR** à **0,79**. En revanche, **Random Forest** (**0,70**) et **MLP** (**0,65**) montrent des performances intermédiaires, suggérant une capacité limitée à capturer certains motifs complexes. Enfin, **SVM**, avec un recall moyen de **0,62**, se positionne comme le modèle le moins performant, reflétant une inadéquation à ces données.

Ces résultats mettent en évidence deux points majeurs : d'une part, l'identification des résistances reste un défi particulier pour certains antibiotiques comme **COL**, nécessitant des améliorations sur l'ingénierie des données ou des ajustements spécifiques des modèles. D'autre part, les performances exceptionnelles de **LGR** et **XGBoost** montrent leur rôle central dans ce type d'analyse, tandis que **SVM** semble inadapté. Ces observations appellent à une optimisation ciblée, notamment pour les antibiotiques aux faibles rappels, afin de renforcer la fiabilité globale du processus de prédiction.

Sélection des caractéristiques importantes (fig. 3)

GPA

Les colonnes de la matrice GPA (**16 005 au total**) sont peu exploitées par les modèles, avec des sélections globalement faibles, allant de **0 % à 2,1 %** selon l'antibiotique et le modèle. **COL** affiche le pourcentage le plus élevé avec **2,1 %** des colonnes sélectionnées par **SVM**, tandis que les autres antibiotiques montrent des sélections plus modestes (ex. : **1,8 %** pour **TOB** avec **Random Forest**). **Random Forest** exploite cette matrice de manière plus intensive pour tous les antibiotiques, suivi de **SVM.** En revanche, les **LGR** que ce soit **FDR** ou **L1** ne selectionnent aucune voir quasiement aucune colonne (autour de **0 %**). Cette matrice semble apporter des informations limitées pour la prédiction des résistances, son exploitation étant fortement dépendante des modèles non linéaires.

SNPs

Avec **72 236 colonnes**, la matrice SNPs est davantage exploitée que **GPA**, bien qu'elle reste sous-utilisée par certains modèles. Les sélections révèlent des résultats identiques à **GPA** avec des pourcentages allant jusqu'à **1,5 %** des colonnes sélectionnées par **Random Forest**. **Random Forest** se démarque avec une sélection significative par rapport aux autres modèles pour tous les antibiotiques, traduisant sa capacité à capturer les interactions complexes des données SNPs. **MLP** affiche une sélection équilibrée mais modérée (ex. : **1,0 %** pour **CIP** et **MER**), tandis que **LGR** (**FDR** et **L1**) et **SVM** n'exploitent aucune ou quasiement aucune colonne, indiquant une contribution restreinte des SNPs dans ces modèles.

GenExp

La matrice GenExp, avec **6 026 colonnes**, est celle qui affiche les pourcentages de sélection les plus élevés. **TOB** domine largement, avec **46,2 %** des colonnes sélectionnées par **LGR FDR**, suivi par **CIP** (**20,7 %**). **LGR FDR** exploite cette matrice de manière agressive pour **TOB**, **CIP** et **CAZ**, mais réduit considérablement ses sélections pour **COL** (**0,9 %**). Les autres antibiotiques présentent des sélections modérées à élevées, comme **CAZ** (**18,5 %** pour **Random Forest**). **Random Forest** affiche une exploitation robuste et équilibrée, atteignant des pourcentages élevés pour plusieurs antibiotiques (de **9,7%** à **18,5 %**). **XGBoost** et **MLP** utilisent cette matrice de manière modérée (environ **1,3-6,5 %**), tandis que **SVM** reste totalement inactif. Ces résultats soulignent l'importance de GenExp pour la détection des résistances, en particulier pour les modèles linéaires et les approches non linéaires flexibles.

Bilan

Les trois matrices montrent des niveaux d'exploitation très différents. **GenExp** est la plus utilisée, en particulier pour TOB et CIP, tandis que **GPA** et **SNPs** sont faiblement exploités. Random Forest et LGR FDR exploitent de façon importante la matrice **GenExp**, tandis que SVM montre une exploitation quasi inexistante, limitant sa capacité à capturer la complexité des données. Ces observations indiquent que l'adéquation entre le modèle, la matrice explicative et l'antibiotique est essentielle pour optimiser la détection des résistances.

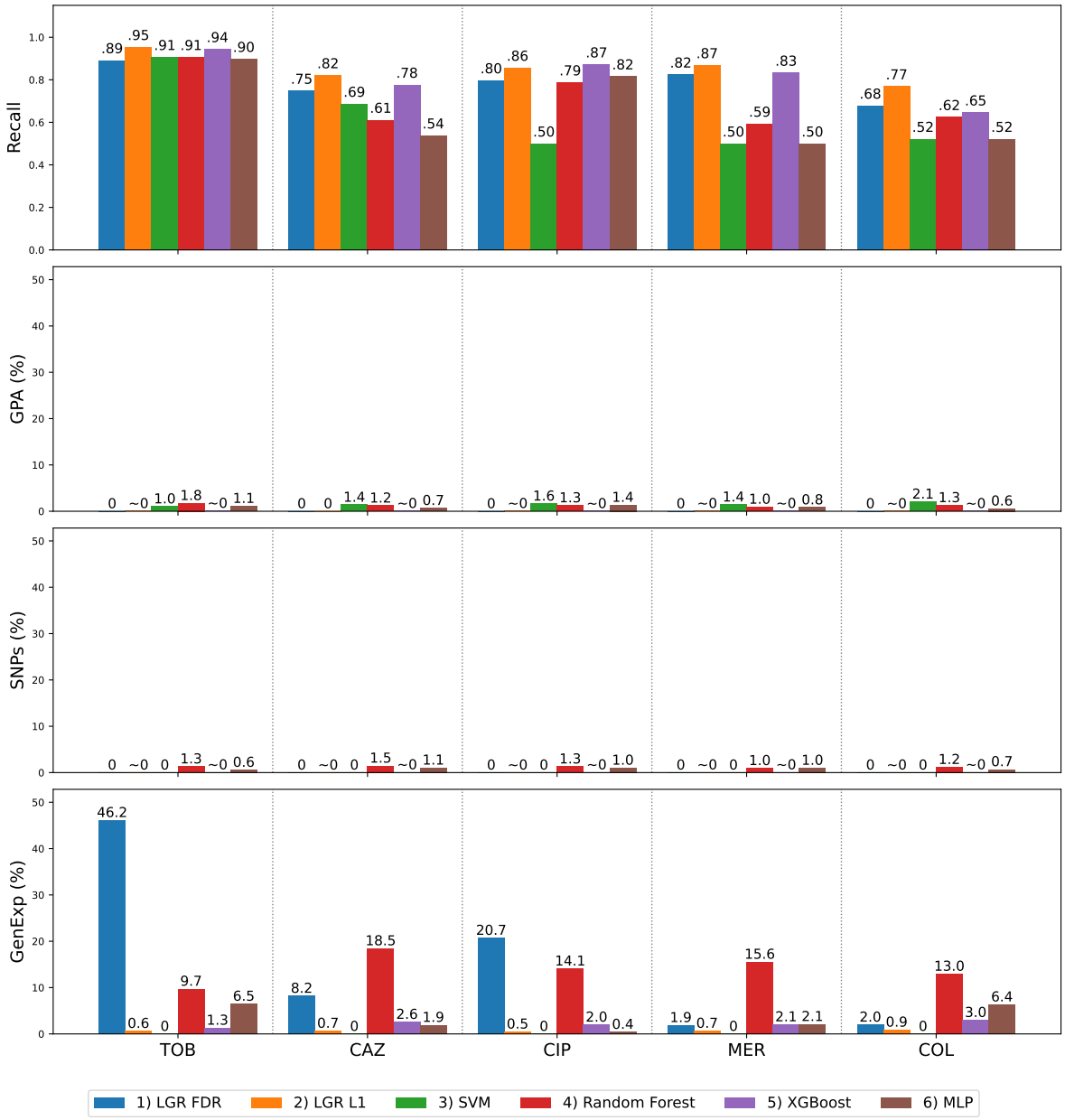

Figure 3: Performances et pourcentages de variables explicatives selectionnées par différents modèles. 1) LGR FDR : régression logistique avec réduction de la dimension via FDR. 2) LGR L1 : régression logistique pénalisée (Lasso). 3) SVM : machines à vecteurs de support. 4) Random Forest : forêt aléatoire (ensemble de décideurs). 5) XGBoost : méthode de boosting par gradient extrême. 6) MLP : modèle de réseau de neurones profond.

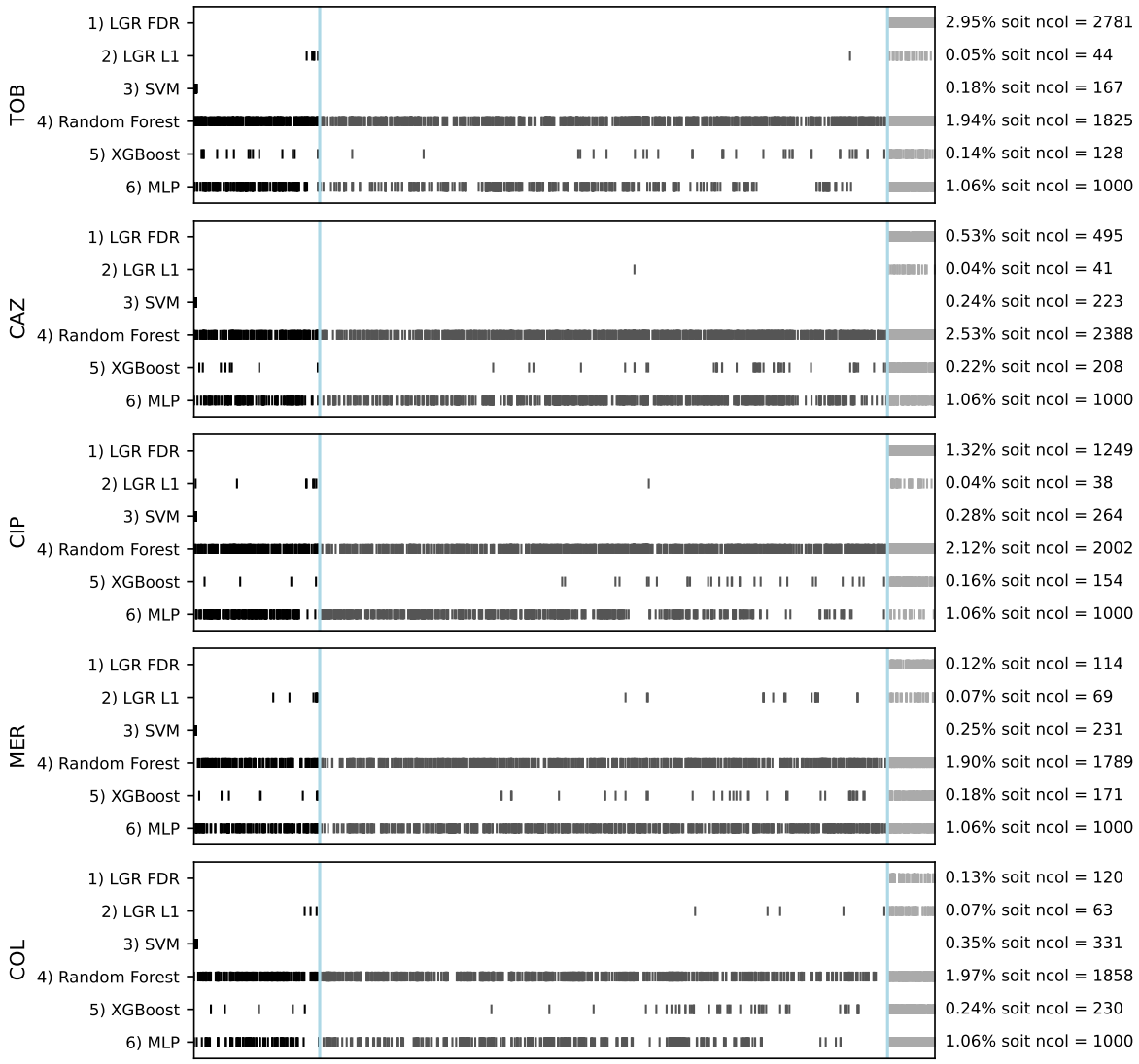

Figure 4: Marqeurs des colonnes sélectionnées des matrices GPA (noir), SNPs (gris foncé), GenExp (gris clair) séparées par un trait bleu clair, avec le pourcentage total de colonnes sélectionnées et l'équivalent chiffré en bout de ligne pour chaque antibiotique et pour les modèles suivant : 1) LGR FDR : régression logistique avec réduction de la dimension via FDR. 2) LGR L1 : régression logistique pénalisée (Lasso). 3) SVM : machines à vecteurs de support. 4) Random Forest : forêt aléatoire (ensemble de décideurs). 5) XGBoost : méthode de boosting par gradient extrême. 6) MLP : modèle de réseau de neurones profond.

Hots spots génétiques (fig. 4)

GPA

Bien que **GPA** soit globalement sous-exploitée, certaines zones spécifiques jouent un rôle clé dans la prédiction des résistances. Les modèles **Random Forest** et **MLP** convergent sur des hot spots partagés pour tous les antibiotiques, tandis que **SVM** identifie un hot spot commun quel que soit l'antibiotique. En revanche, le modèle linéaire (**LGR L1**) et **XGBoost** ne montrent pas de convergence notable, leurs sélections restant dispersées.

SNPs

Les modèles **Random Forest** et **MLP** montrent une convergence notable dans certaines plages de SNPs pour plusieurs antibiotiques. **XGBoost** présente une couverture plus diffuse mais sans convergence stricte entre antibiotiques. Le modèle linéaire **LGR L1** se montre beaucoup moins cohérent, n'identifiant que des zones isolées sans chevauchement clair.

GenExp

Pour **GenExp**, les modèles affichent une convergence remarquable, à l'exception de **SVM**, qui reste inactif. Des hot spots partagés sont identifiés pour tous les antibiotiques, avec une concentration particulièrement marquée pour **CAZ**, qui présente le plus grand nombre de zones communes entre modèles.

Bilan

Random Forest et MLP se distinguent par leur cohérence sur **GPA**, **SNPs** et **GenExp**, identifiant des hot spots communs, tandis que XGBoost et les modèles linéaires restent moins convergents, hormis sur **GenExp**. SVM converge sur **GPA**. Globalement, les modèles non linéaires surpassent les autres en termes de convergence.

Discussion

Les résultats de cette étude montrent que les approches de machine learning jouent un rôle clé dans la prédiction des résistances bactériennes et dans l'identification des caractéristiques biologiques pertinentes. Les performances des modèles varient selon les données utilisées et leur capacité à exploiter la richesse et la complexité de ces données. En particulier, la régression logistique avec sélection des variables via contrôle du **FDR** (False Discovery Rate) se distingue par son efficacité et sa simplicité. En se basant sur des variables préalablement sélectionnées, ce modèle maintient des performances solides tout en limitant les risques liés à la dimensionnalité élevée des données. Sa capacité à produire des résultats interprétables et rapides en fait une option attrayante pour une utilisation pratique, en particulier dans un contexte clinique où la transparence et l'efficacité sont cruciales.

En revanche, certains modèles comme **SVM** et **MLP** affichent des performances globalement faibles dans cette étude. La faible performance de **SVM**, qui se base sur la maximisation des marges pour séparer les classes, peut s'expliquer par une inadéquation avec la nature des données utilisées. **SVM** semble particulièrement limité dans son exploitation des matrices de grande dimension, comme celles

contenant les données transcriptomiques ou les SNPs, où des relations complexes et non linéaires dominent. Pour **MLP**, la performance limitée pourrait également être liée à plusieurs facteurs : un surajustement sur des données complexes, un nombre insuffisant d'échantillons pour l'apprentissage ou une mauvaise optimisation des hyperparamètres. L'hypothèse d'une mauvaise implémentation ou d'un calibrage insuffisant, notamment en ce qui concerne les architectures réseau ou les critères de convergence, mérite d'être explorée pour ces modèles.

D'autre part, les modèles non linéaires comme **Random Forest** et **XGBoost** montrent une capacité accrue à exploiter les relations complexes entre variables. Ils identifient efficacement des hot spots génétiques dans les données SNPs et GenExp, suggérant leur pertinence pour capturer des interactions non linéaires liées aux mécanismes de résistance. Toutefois, cette performance se fait parfois au détriment de la simplicité et de l'interprétabilité, ce qui peut limiter leur intégration dans des analyses visant à éclairer directement les mécanismes biologiques.

Enfin, les différentes sources de données biologiques n'ont pas le même impact sur les performances des modèles. Les données transcriptomiques (GenExp) se démarquent comme étant les plus informatives, particulièrement pour les modèles linéaires et non linéaires, car elles permettent de capturer des variations globales liées à l'état cellulaire des bactéries. Les SNPs offrent une perspective complémentaire mais sont surtout exploitées par des modèles non linéaires comme Random Forest, qui peuvent gérer leur structure complexe. En revanche, les données GPA, bien que pertinentes sur le plan biologique, sont peu utilisées et n'apportent qu'une contribution marginale aux performances globales des modèles.

Ces résultats mettent en lumière la nécessité d'un calibrage minutieux des modèles, en particulier pour ceux comme **SVM** et **MLP**, qui semblent sous-performants dans leur implémentation actuelle. Une exploration plus approfondie des configurations et hyperparamètres de ces modèles pourrait améliorer leur exploitation des données et fournir des résultats plus compétitifs, en particulier dans des environnements où des relations complexes doivent être capturées.

Conclusion

Cette étude met en évidence le potentiel des approches de machine learning pour prédire les résistances bactériennes à partir de données génomiques et transcriptomiques. Les performances des modèles dépendent fortement du type de données utilisées, de leur complexité et de leur adéquation avec les algorithmes choisis. Les données transcriptomiques (GenExp) se révèlent particulièrement informatives, en permettant de capturer des variations biologiques globales, tandis que les SNPs apportent des informations complémentaires, surtout exploitées par des modèles non linéaires comme **Random Forest** et **XGBoost**. En revanche, les données GPA semblent moins contributives dans ce cadre.

Parmi les modèles testés, **LGR** (**L1** et **FDR**) et **XGBoost** se démarquent par leur robustesse et leur capacité à fournir des résultats fiables tout en restant relativement interprétables. Ces deux approches apparaissent comme les plus prometteuses pour une application pratique, notamment en contexte clinique. À l'inverse, des algorithmes comme **SVM** et **MLP** montrent des performances limitées, nécessitant des ajustements supplémentaires pour exploiter pleinement les données.

Ces résultats soulignent l'importance de choisir les modèles en fonction des données disponibles et des objectifs spécifiques. Ils appellent également à une optimisation plus fine des algorithmes moins performants, afin de maximiser leur potentiel. Enfin, l'intégration de ces outils dans des pipelines automatisés pourrait significativement améliorer la rapidité et la précision des diagnostics, ouvrant la voie à des approches personnalisées et plus efficaces pour lutter contre la résistance bactérienne.