16-27 juin 2025 2nde 3

# Rapport de stage de seconde

Tristan KERMORVANT

Responsable de stage : Marie GENERALI

# Table des matières

# 1 Présentation de l'entreprise

## 1.1 Fiche synthétique

**Date de fondation :** 1967

**Type :** Établissement public à caractère scientifique et technologique

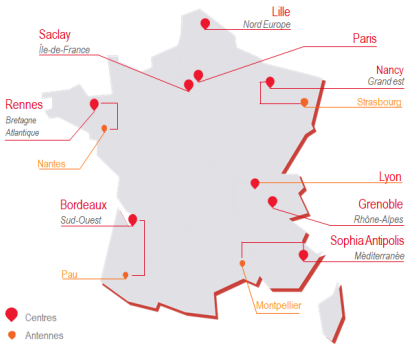
**Effectif total :** 4 800+ dont scientifiques chercheurs (2024)

**Président :** Bruno Sportisse

**Budget :** 307 millions (2024)

**Affiliation :** Ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche

**Centres de l'Inria :**




## 1.2 Qu'est-ce que l'INRIA ?

Fondé en 1967, Inria, ou Institut national de recherche en informatique et en automatique, est un établissement public à caractère scientifique et technologique. Initialement créé sous le nom d'IRIA, il est conçu pour piloter l'axe recherche du « plan Calcul », une initiative stratégique visant à renforcer l'indépendance technologique de la France en informatique face à la domination américaine dans ce domaine.

Dans les années 1970, il développe notamment le projet *Cyclades*, un réseau de communication pionnier, rivalisant avec Arpanet aux États-Unis, qui a inspiré la création d'Internet. En 1980, l'IRIA devient l'I*N*RIA, qui traduit l'ancrage désormais national de L'Institut.

L'objectif initial de l'INRIA et de « Pourvoir la France en technologies de pointe et développer son autonomie et sa souveraineté ».

Aujourd'hui, l'institut se présente comme l'*institut national de recherche dans les sciences et technologies du numérique*. Ses missions sont dans les domaines de l'informatique, de l'automatique et des mathématiques appliquées de promouvoir la recherche, l'innovation et la formation.

> « Porter la révolution du numérique par les mathématiques et l'informatique. » [1]

L'Inria possède de nombreux centres de recherche répartis sur le territoire français. Son siège est situé à Rocquencourt, à proximité de Versailles, et j'ai effectué mon stage au centre de Saclay, au coeur de la technopole Paris-Saclay.

FIGURE 1 – Le bâtiment Turing à Palaiseau, sur le campus de l'École polytechnique et son intérieur.

Les centres sont eux-mêmes divisés en équipes de recherche, dont celle dans laquelle j'ai effectué mon stage, SODA.

## 1.3 Présentation de l'équipe SODA

L'équipe SODA mène des recherches à l'intersection de l'apprentissage automatique, des bases de données et des sciences sociales quantitatives.

> « Nous contribuons des outils d'apprentissage statistique pour répondre à des questions de science des données, typiquement sur des données relationnelles. Nos applications principales sont la santé et l'éducation. » [2]

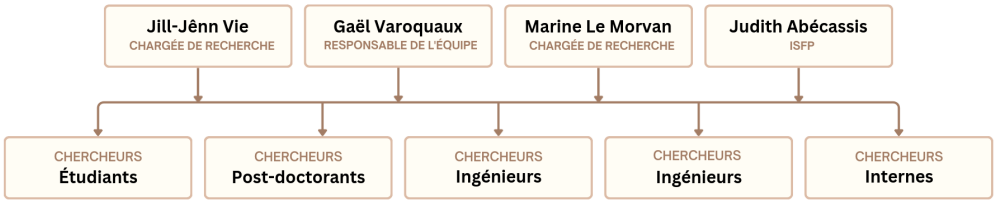



FIGURE 2 – Organisation de l'équipe SODA.

Ils maintiennent notamment le logiciel Scikit-learn, la bibliothèque la plus utilisée pour l'apprentissage automatique en Python.


FIGURE 3 – Logo de l'équipe SODA et du logiciel Scikit-learn.

# 2 Mon travail de stage

## 2.1 Intelligence artificielle et apprentissage automatique

L'intelligence artificielle est une discipline scientifique qui vise à faire réaliser par des machines des tâches mettant en œuvre des fonctions cognitives humaines : perception, language, raisonnement, plannification, mouvement. Elle est à la frontière entre l'informatique, la statistique, les mathématiques, les sciences cognitives et la philosophie.

L'apprentissage automatique ou machine learning est un domaine de l'intelligence artificielle. C'est une technique permettant aux algorithmes d'intelligence artificielle de s'entraîner et d'apprendre grâce à des exemples, sans avoir été préalablement programmés spécifiquement à cet effet. L'entraînement de tels algorithmes nécessite une grande quantité de données sur lesquelles entraîner les modèles.

Il existe plusieurs types d'apprentissage automatique, dont l'un des plus commun, l'apprentissage supervisé, qui consiste à entraîner un modèle sur un ensemble de données étiquetées et de le tester sur un ensemble de données disjoint afin d'évaluer ses performances ou mesurer les biais qu'il peut contenir.

## 2.2 Problématiques des biais et de l'équité

### D'où viennent les biais dans les modèles d'IA ?

Les systèmes d'IA basés sur l'apprentissage automatique apprennent à partir de jeux de données existants, reflétant des pratiques humaines. Lorsque ces données contiennent des biais sociétaux (inégalités, discriminations ou stéréotypes), les algorithmes peuvent involontairement les reproduire, car ils sont programmés à imiter ce qu'ils voient. Bien que le traitement des données soit mathématiquement neutre, les préjugés présents dans les données d'entraînement se retrouvent ainsi répliqués dans les résultats du modèle.

### Comment se manifestent-ils ?

Les biais algorithmiques se concrétisent par des prédictions discriminantes ou des représentations stéréotypées. Dans un algorithmes d'évaluation de candidatures, par exemple, un biais pourrait se traduire par une discrimination à l'embauche envers certains groupes sociaux ou ethniques.

Un algorithme de recommandation de contenu pourrait renforcer des stéréotypes en suggérant des contenus basés sur des préférences biaisées.

Ou encore, une IA de génération d'images pourrait produire des représentations stéréotypées de certaines professions, en associant par exemple les femmes à des rôles de secrétaires ou d'infirmières, tandis que les hommes seraient associés à des rôles de PDG ou de scientifiques.


FIGURE 4 – L'AI générative *Midjourney* générant des images de PDG et de secrétaires. © Havas Paris / Presse

### Pourquoi sont-ils problématiques ?

Les biais algorithmiques sont problématiques car ils renforcent les inégalités sociales et perpétuent les stéréotypes. Ils peuvent conduire à des décisions injustes,

Aux États-Unis, la police utilise un algorithme de prédiction de récidive criminelle. Cette IA cible deux fois plus les accusés noirs que les accusés blancs. [3]

Dans la plupart des algorithmes, les prédictions ne sont pas accompagnées d'explications, ne permettant pas de comprendre leur fonctionnement qui est souvent perçu comme une boîte noire. Les biais sont donc implicites et difficiles à identifier, ce qui rend notre emprise sur eux très limitée.

## 2.3 Méthodes d'atténuation des biais

**Qu'est ce que la fairness ?**

L'équité (fairness) en machine learning consiste à éviter que les modèles apprennent ou reproduisent des biais envers certains groupes (par exemple selon le sexe, la race, l'âge, etc.). Il existe trois grandes approches pour rendre un modèle plus juste.

1. Pré-traitement (*pre-processing*) : On agit avant l'apprentissage, en modifiant les données d'entraînement.
   — Supprimer les variables sensibles (sexe, ethnie, etc.)
   — Rééquilibrer la base de données entre les groupes
   — Modifier ou corriger les étiquettes (labels)
   **Avantages :** simple, compatible avec tous les modèles
   **Limites :** ne garantit pas l'équité finale parfaite
2. Entraînement équitable (*in-processing*) : On agit pendant l'apprentissage, en modifiant le fonctionnement du moldèle.
   — Ajouter une pénalité de biais dans la fonction d'entraînement
   — Contraindre le modèle à respecter certaines propriétés d'équité
   **Avantages :** flexible et puissant
   **Limites :** nécessite des algorithmes spécifiques, peut être complexe à implémenter
3. Ajustement des décisions (*post-processing*) : On agit après l'apprentissage, en corrigeant les prédictions du modèle.
   — Ajuster le seuil de décision pour chaque groupe
   — Recalibrer les probabilités de sortie
   — Forcer une égalité de taux entre les groupes
   **Avantages :** simple à mettre en place, sans réentraîner le modèle
   **Limites :** peut être arbitraire ou perçu comme injuste

Ma tâche durant ce stage a été d'expérimenter avec les differentes méthodes d'atténuation des biais en comparant également les performances et l'équité de différents modèles d'intelligence artificielle.

## 2.4 Experimentation

### 2.4.1 Données

Le jeu de donnée est celui d'étudiants en droit, contenant de nombreuses informations comme l'éthnie, le sexe, le score au test d'entrée en école de

drotit, la moyenne des études universitaires, etc. sur plus de 20 000 étudiants. Il est utilisé pour prédire si un étudiant réussira ou non son examen de fin de première année. Un extrait du jeu de données est présenté sur la figure 5.

| idx | race | sex | LSAT | UGPA | region_first | ZFYA | sander_index | first_pf |
|---|---|---|---|---|---|---|---|---|
| 0 | White | 1 | 39.0 | 3.1 | GL | -0.98 | 0.782738 | 1.0 |
| 1 | White | 1 | 36.0 | 3.0 | GL | 0.09 | 0.735714 | 1.0 |
| 2 | White | 2 | 30.0 | 3.1 | MS | -0.35 | 0.670238 | 1.0 |
| 5 | Hispanic | 2 | 39.0 | 2.2 | NE | 0.58 | 0.697024 | 1.0 |
| 6 | White | 1 | 37.0 | 3.4 | GL | -1.26 | 0.786310 | 1.0 |
| 7 | White | 1 | 30.5 | 3.6 | GL | 0.30 | 0.724107 | 1.0 |
| 8 | White | 2 | 36.0 | 3.6 | GL | -0.10 | 0.792857 | 1.0 |
| 9 | White | 2 | 37.0 | 2.7 | NE | -0.12 | 0.719643 | 0.0 |
| 13 | White | 1 | 37.0 | 2.6 | GL | 1.53 | 0.710119 | 1.0 |
| 14 | White | 2 | 31.0 | 3.6 | GL | 0.34 | 0.730357 | 1.0 |

FIGURE 5 – Extrait du jeu de données d'étudiants en droit.

La première étape consiste à visualiser le jeu de données pour tenter d'identifier des biais qu'il pourrait contenir. Dès lors, on remarque sur la figure 6 que les étudiants blancs ont une moyenne de réussite grandement supérieure à celle des étudiants noirs ou amérindiens, de même pour les notes.

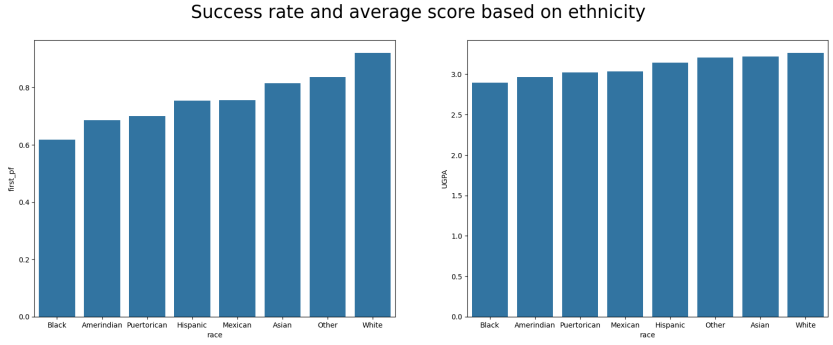



FIGURE 6 – La moyenne de réussite des étudiant et leur des notes en fonction de leur ethnie.

De plus, on remarque que les étuidiants blancs sont sur-représentés dans le jeu de données, constituant la quasi totalité des étudiants, à hauteur de 83,9%. Figure 7. Cela constitue donc un biais que l'on pourra exploiter.

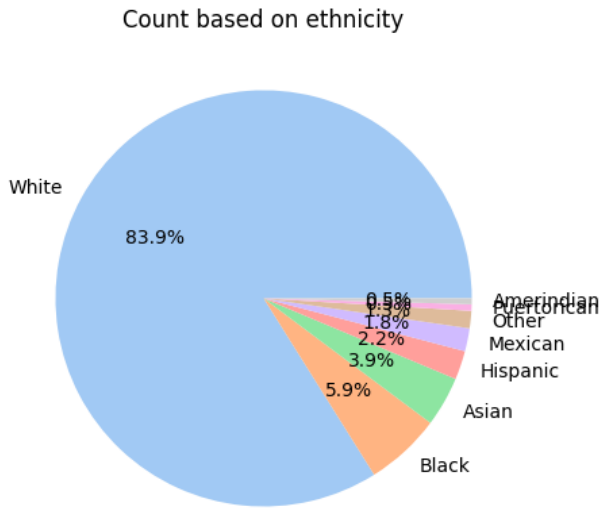



FIGURE 7 – Répartition des étudiants en fonction de leur ethnie.

Néanmoins, on ne remarque pas de réelle différence entre les étudiants masculins et féminins, ni dans la moyenne de réussite, ni dans les notes. On peut donc supposer que le sexe n'est pas un facteur de biais dans ce jeu de données. Figure 8

Par la suite, on utilisera donc le biais lié à l'éthnie que l'on tentera de réduire grâce aux différentes méthodes d'atténuation des biais.

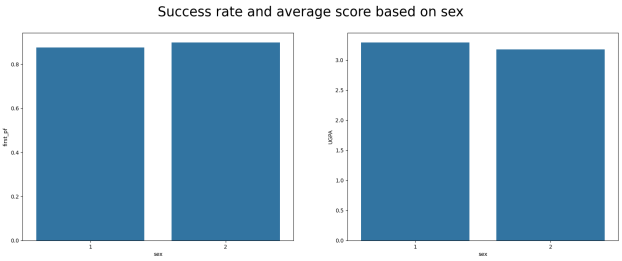



FIGURE 8 – Répartition des étudiants en fonction de leur sexe.

### 2.4.2 Modèles

J'ai testé les performances de trois algorithmes d'intelligence artificielle différents, mais par simplicité, je n'en presenterai ici que deux.

#### La régression logistique

La régression logistique est un algorithme statistique simple qui utilise une fonction affine ansi qu'une fonction sigmoïde pour prédire la probabilité d'appartenance à une classe. Il est souvent utilisé pour des problèmes de classification binaire.

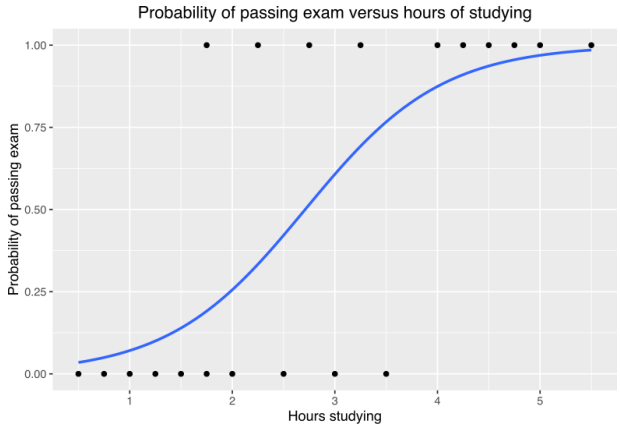



FIGURE 9 – Courbe d'une régression logistique à une dimension. © Wikipedia

#### Les réseaux de neurones

Les réseaux de neurones sont des modèles informatiques inspirés du fonctionnement du cerveau humain. Ils sont constitués de couches de neurones interconnectés, où chaque neurone est une unité de traitement simple. Ils possèdent trois types de couches : la couche d'entrée, les couches cachées et la couche de sortie. La couche d'entrée comporte toutes les informations que l'on donne au modèle, cela peut être les données d'une image ou ici les informations sur les étudiants par exemple. Les neurones des couches cachées vont alors effectuer des opérations sur les valeurs de la couche d'entrée, qui

vont permettre de prendre une décision en sortie. À chaque itération d'entrainement, l'algorithme d'apprentissage va modifier les paramètres de chaque neurone (qu'on appelle poids) pour se rapprocher de la sortie voulue.

Plus il y a de couches cachées et de neurones, plus le modèle est complexe et long à entraîner, mais plus il peut réaliser des tâches complexes. Ici, j'utilise un réseau à une seule couche, dit linéraire.

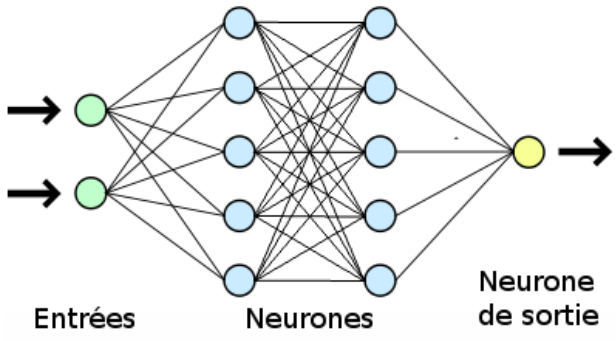
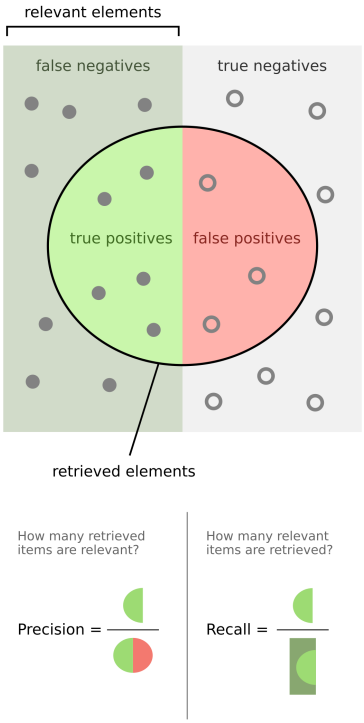



FIGURE 10 – Schéma d'un réseau de neurones.




Voici la performance du réseau de neurones. La précision et le rappel sont deux métriques de performance utilisées pour évaluer les modèles d'apprentissage automatique. La précision mesure la proportion de prédictions correctes parmi toutes les prédictions faites, tandis que le rappel mesure la proportion de positives parmis celles qui auraient dû être prédites.

Sur le *classification report* Table 1, on remarque une faible performance du modèle sur la prédiction *0*, qui correspond à l'échec de l'examen de fin de première année, mais on constate que la classe *1* est bien plus représentée que la classe *0* (5853 contre 685), ce qui explique la grande performace du modèle sur cette classe.

### 2.4.3 Méthode de réduction des biais

#### Pré-traitement : *Correlation Remover*

La méthode de pré-traitement *Correlation Remover* consiste à supprimer la corrélation entre les variables sensibles (ici l'ethnie) et les autres variables du jeu de données. Elle a été réalisée sur le modèle de régression logisitique.

Pour quantifier l'équité, on utilise la métrique *demographic parity*, qui

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0.0 | 0.57 | 0.31 | 0.40 | 685 |
| 1.0 | 0.92 | 0.97 | 0.95 | 5853 |
| **Accuracy** | | | 0.90 | 6538 |
| **Macro avg** | 0.75 | 0.64 | 0.67 | 6538 |
| **Weighted avg** | 0.89 | 0.90 | 0.89 | 6538 |

TABLE 1 – Classification report pour le réseau de neurones.

assure que la proportion d'appartenace à chaque groupe est indépendante de la variable sensible. Il existe deux mesures, comprises entre 0 et 1, qui quantifient cette métrique : la *demographic parity difference* et la *demographic parity ratio*. La première doit tendre vers 0 et la seconde vers 1 pour que l'équité soit respectée. Plus d'informations sur `https://fairlearn.org/main/user_guide/assessment/common_fairness_metrics.html`

On observe sur le tableau 2 que l'équité est entièrement respectée, neanmoins, on remarque que sur la figure 11 que la performance du modèle est devenue nulle, car en supprimant la corrélation entre les variables sensibles et les autres variables, on supprime également une information sans doute nécessaire à la prédiction de la variable cible.

| | Avant | Après |
|---|---|---|
| Demographic parity difference | 0.35 | 0.00 |
| Demographic parity ratio | 0.65 | 1.00 |

TABLE 2 – Évolution des métriques d'équité (*demographic parity*) avant et après la méthode *Correlation Remover*.

### In-processing : *pénalité d'équité*

Cette méthode a été implémentée sur le réseau de neurones car il est très facile d'ajouter une pénalité dans la fonction de perte, c'est à dire durant la boucle d'entraînement de l'algorithme. Elle consiste à calculer la métrique d'équité *demographic parity* à chaque itération d'entraînement et à ajouter une pénalité dans la fonction de perte si cette métrique n'est pas respectée. À force, le modèle apprend à respecter l'équité tout en essayant de maximiser la performance.

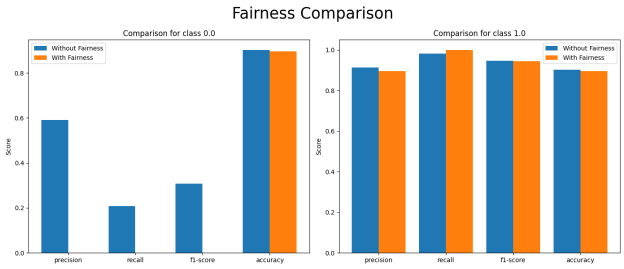



FIGURE 11 – Résultat de la méthode *Correlation Remover* sur le jeu de données.

| | **Avant** | **Après** |
|---|---|---|
| Demographic parity difference | 0.45 | 0.37 |
| Demographic parity ratio | 0.54 | 0.62 |

TABLE 3 – Évolution des métriques d'équité (*demographic parity*) avant et après la pénalité d'équité (in-processing) sur le réseau de neurones.

On remarque une amélioration de l'équité, plus légère, néanmoins la performance du modèle est restée quasi semblable, comme on peut le voir sur la figure 12.

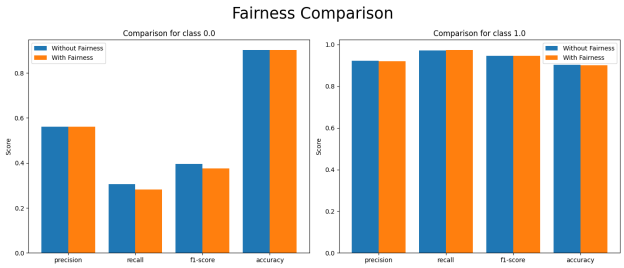



FIGURE 12 – Résultat de la pénalité d'équité sur le réseau de neurones.

### Post-processing : *Threshold Optimization*

Cette méthode consiste à ajuster le seuil de décision du modèle pour chaque groupe afin de respecter l'équité. C'est l'équivalent à faire des "quotas" pour chaque groupe, en ajustant le seuil de décision pour que la proportion de prédictions positives soit égale entre les groupes. Elle a été effectuée

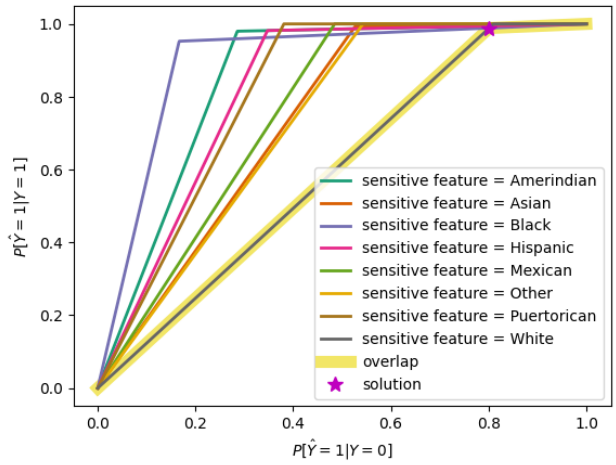



FIGURE 13 – Les seuils de décision pour chaque groupe, où la classe Black est "avantagée"

sur le modèle de régression logistique.

| | **Avant** | **Après** |
|---|---|---|
| Demographic parity difference | 0.35 | 0.07 |
| Demographic parity ratio | 0.65 | 0.93 |

TABLE 4 – Évolution des métriques d'équité (*demographic parity*) avant et après la méthode de *Threshold Optimization*.

On remarque une amélioration significative de l'équité, mais la performance du modèle a donc été également dégradée, cependant moins que la méthode de pré-traitement comme on peut le voir sur la figure 14.

### 2.4.4 Analyse et interprétation

La méthode de pré-traitement *Correlation Remover* et de post-processing *Threshold Optimization* ont permis d'atteindre une équité parfaite, cependant les modèles ont perdu toute capacité de prédiction. La méthode d'in-processing, la pénalité d'équité, a permis d'améliorer l'équité du modèle de manière significative, mais sans dégrader la performance du modèle, ce qui est un bon compromis.

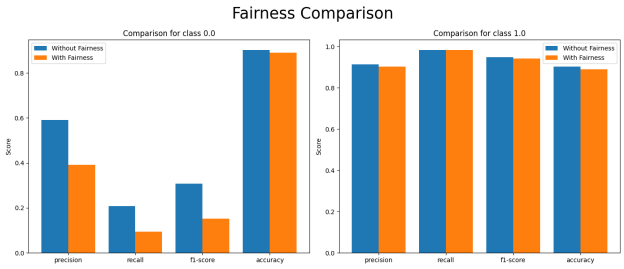



FIGURE 14 – Résultat de la méthode de *Threshold Optimization*.

On a pu observer que grâce à ces différentes méthodes, on a pu améliorer l'équité des modèles, mais à chaque fois au prix d'une perte de performance. Cela soulève donc quelques questions éthiques :

— Un modèle peut-il être performant et juste ?
— Quelles variables sont utiles ou problématiques ?
— Faut-il toujours un seul modèle pour tout le monde ?

**Comment améliorer les résultats ?**

Pour chaque méthode, il est possible d'améliorer les résultats en ajustant l'intensité de la décorrélation ou le seuil de tolérance pour tenter de préserver la performance du modèle. Sur la méthode d'in-processing, cela avait déjà été fait, un coefficient $\lambda$ avait été ajouté dans la fonction de perte pour ajuster l'importance de la pénalité d'équité. Je l'avait également optimisé pour trouver le meilleur compromis entre performance et équité, mais il est possible de l'optimiser encore plus finement.

# 3 Réflexions personnelles

## 3.1 Comparaison avec le stage de 3e

Recherche privée vs recherche publique

## 3.2 Activtés spécifiques au domaine de la recherche publique

Soutenance de thèse, coding sprint, LaTeX

### 3.3 Conclusion

# 4 Remerciements

# Références

[1] INRIA. Vidéo de présentation de l'inria, 28/06/2025. https://www.youtube.com/watch?v=CuJ_39LY1rY.

[2] SODA. Présentation de l'équipe soda, 28/06/2025. https://www.inria.fr/fr/soda.

[3] Amnesty International. Intelligence artificielle : les sept choses qu'on ne vous dit pas, 28/06/2025. https://www.amnesty.fr/actualites/intelligence-artificielle-les-sept-choses-qu-on-ne-vous-dit-pas.