Digital Bridge RDC

Biais et éthique de la science des données

December 2022

Introduction

- La science des données joue un rôle crucial dans de nombreuses industries, et les décisions prises sur la base des données ont un impact significatif sur la société.
- Il est donc important de s'assurer que les pratiques de la science des données sont équitables, impartiales et éthiquement saines.

Biais

Biais dans les données

- **Biais:** Préjugés en faveur ou contre une chose, une personne ou un groupe par rapport à un autre, généralement d'une manière considérée comme injuste.
- Le biais dans les données fait référence aux distorsions systématiques et involontaires des données qui peuvent conduire à des conclusions incorrectes ou injustes.
- Des biais peuvent survenir à différentes étapes du processus de collecte, de préparation et d'analyse des données.

Types de biais

- Il existe de nombreux types de biais qui peuvent se produire dans la science des données, y compris le biais d'échantillonnage, le biais de sélection et le biais de mesure.
- Comprendre les différents types de biais est important pour identifier et atténuer leurs effets.

Biais d’échantillonnage

- En statistique, le biais d’échantillonnage est un biais dans lequel un échantillon est recueilli de telle sorte que certains membres de la population visée ont une probabilité d’échantillonnage plus faible ou plus élevée que d’autres.
- Il en résulte un échantillon biaisé d’une population (ou de facteurs non humains) dans lequel tous les individus, ou les instances, n’étaient pas également susceptibles d’avoir été sélectionnés
- Si cela n’est pas pris en compte, les résultats peuvent être attribués à tort au phénomène étudié plutôt qu’à la méthode d’échantillonnage.

Biais de sélection

- Le biais de sélection est le biais introduit par la sélection d'individus, de groupes ou de données à analyser de telle sorte qu'une randomisation appropriée n'est pas réalisée, omettant ainsi de s'assurer que l'échantillon obtenu est représentatif de la population destinée à être analysée.
- On l'appelle parfois l'effet de sélection. L'expression « biais de sélection » fait le plus souvent référence à la distorsion d'une analyse statistique, résultant de la méthode de prélèvement des échantillons.
- Si le biais de sélection n'est pas pris en compte, certaines conclusions de l'étude peuvent être fausses.

Biais de mesure

- Le biais de mesure fait référence à une erreur systématique dans la façon dont une variable est mesurée, ce qui peut entraîner des données inexactes ou trompeuses.
- Ce type de biais peut se produire lorsque le processus de mesure est défectueux ou lorsque les outils de mesure ne sont pas correctement calibrés.
- Il existe plusieurs types de biais de mesure, notamment:
  - **Biais de réponse:** Cela se produit lorsque les réponses à un sondage ou à un questionnaire ne sont pas exactes ou honnêtes, en raison de facteurs tels que la désirabilité sociale ou le manque de compréhension.
  - **Biais de l'observateur:** Cela se produit lorsque la personne qui recueille les données a des idées préconçues ou des attentes qui influencent la façon dont elle observe et enregistre les données.
  - **Biais instrumental:** Cela se produit lorsque les outils ou instruments de mesure ne sont pas correctement étalonnés ou ne sont pas adaptés à la tâche.
  - Le biais de mesure peut avoir de graves conséquences, car il peut conduire à des résultats inexacts ou trompeurs qui peuvent avoir une incidence sur des décisions importantes.
- Pour éviter les biais de mesure, il est important d'utiliser des outils de mesure bien calibrés et de suivre les procédures d'échantillonnage et de collecte de données appropriées.

Atténuation des biais

- Il existe plusieurs stratégies pour atténuer les biais dans la science des données, notamment:
  - Assurer la diversité dans la collecte et la représentation des données
  - Utilisation de techniques d'échantillonnage appropriées
  - Validation et nettoyage des données
  - Utilisation d'algorithmes et de modèles impartiaux

Questions Ethiques

Questions éthiques en science des données

- La science des données peut également soulever des préoccupations éthiques liées à la confidentialité, à la sécurité et au potentiel d'abus ou d'utilisation abusive des données.
- Il est important de tenir compte des implications éthiques des projets de science des données et de prendre les mesures appropriées pour protéger les informations sensibles et prévenir les préjudices.

Meilleures pratiques pour une science éthique des données

- Voici quelques bonnes pratiques pour une science éthique des données:
  - Obtention d'un consentement éclairé pour la collecte de données
  - Protection de la vie privée des individus
  - Assurer la sécurité des données
  - Faire preuve de transparence quant à l'utilisation et aux finalités des données

Exemples

Préjugés et questions éthiques – exemples (1)

- **Scandale Cambridge Analytica:** Dans ce cas, une société de conseil politique appelée Cambridge Analytica a utilisé l'exploration de données et le profilage psychologique pour influencer les résultats des élections.
- L'entreprise a obtenu les données de millions d'utilisateurs de Facebook à leur insu ou sans leur consentement, ce qui a suscité d'importantes préoccupations concernant la vie privée et le potentiel d'abus de données personnelles. (https://en.wikipedia.org/wiki/Cambridge_Analytica)

Préjugés et questions éthiques – exemples (2)

- **Le « Projet Maven » de Google:** Dans ce cas, Google a développé un système d'intelligence artificielle pour le Pentagone qui a été utilisé pour analyser les images de drones.
- Le projet a soulevé des préoccupations concernant l'utilisation de l'IA à des fins militaires, ainsi que le potentiel de résultats biaisés en raison de la nature biaisée des données utilisées pour former le système. (https://www.wired.com/story/googles-project-maven-drone-ai-raises-questions-of-killer-robots/)

Préjugés et questions éthiques – exemples (3)

- **L'algorithme d'embauche biaisé d'Amazon :** En 2018, Amazon a développé un algorithme d'embauche biaisé qui favorisait les hommes par rapport aux femmes.
- L'algorithme était basé sur les données des décisions d'embauche passées de l'entreprise, qui étaient principalement masculines, ce qui a conduit à un modèle biaisé qui favorisait les candidats masculins. (https://www.cnn.com/2018/10/10/tech/amazon-hiring-algorithm-bias/index.html)

Préjugés et questions éthiques – exemples (4)

- **Chatbot Tay de Microsoft:** En 2016, Microsoft a publié un chatbot appelé « Tay » qui a été conçu pour apprendre et interagir avec les utilisateurs sur les médias sociaux. Cependant, le chatbot a rapidement été impliqué dans la controverse après avoir été formé sur des données biaisées et offensantes et avoir commencé à cracher des messages haineux et inappropriés. (https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist)

Partialité et éthique : Conclusion

- Les préjugés et les questions éthiques dans la science des données peuvent avoir de graves conséquences et il est important d'être conscient de ces problèmes afin de garantir des pratiques de données équitables et responsables.
- En suivant les meilleures pratiques et en étant conscients des risques potentiels, les scientifiques des données peuvent aider à s'assurer que les données sont utilisées au profit de la société.