



UNIVERSITÉ LIBRE DE BRUXELLES



Faculté de Lettres, Traduction et Communication

S'appuyer sur l'IA pour identifier les biais dans la rédaction des offres d'emploi

Marina AUBERT

Mémoire présenté sous la direction de
Sébastien DE VALERIOLA en vue de
l'obtention du titre de Master en Sciences
et Technologies de l'Information et de la
Communication

Année académique 2023–2024

Résumé du mémoire

AUBERT Marina - LTC Master STIC - 2023-2024

Titre : S'appuyer sur l'IA pour identifier les biais dans la rédaction des offres d'emploi

Mots-clés : biais, discrimination, offre d'emploi, ia, qualité des données, machine learning, visualisation

L'objectif de ce mémoire de master est d'identifier les biais des offres d'emploi via l'IA. Après un état des lieux sur les différents types de biais existants dans les offres d'emploi, nous avons voulu identifier quel vocabulaire attirerait plus de candidatures de femmes ou d'hommes. A cet effet, nous avons développé un prototype de machine learning, basé sur les données issues des fiches formation de Bruxelles Formation et des candidatures qu'elles ont reçues, et un algorithme d'apprentissage supervisé ; celui-ci, développé en R, utilise la méthode du RandomForest. Par différentes techniques de visualisation, nous avons pu améliorer ce prototype. La liste de vocabulaire générée par notre prototype s'est avérée similaire aux études déjà connues.

Nous en avons conclu des différentes approches évoquées et expérimentées que, les offres d'emploi n'étant pas générées à la volée, il serait probablement plus efficace de former les spécialistes du recrutement à la rédaction inclusive, plutôt que de mobiliser périodiquement du personnel spécialisé en optimisation d'algorithmes d'intelligence artificielle.

Remerciements

À mes fils pour leur compréhension et leur soutien, que cette reprise d'études à 43 ans puisse les inspirer : il n'est jamais trop tard

À Sébastien De Valeriola, Isabelle Boydens et Laurence Dierickx pour leur guidance et leur inspiration

À Morad Chaboune et Ibrahim Ouassari pour les mises en contact

À Leila Maidane pour son temps et son inspiration

À mes relectrices et relecteurs

À Maxime, la chatte, pour son accompagnement quotidien

À mes prédecesseuses qui m'ont permis d'atteindre mon objectif

À moi-même pour avoir eu le courage d'affronter la situation et cru en ma réussite

Table des matières

1	Introduction	6
1.1	Exploiter utilement ma reprise d'études	6
1.2	Construire ma question de recherche	7
1.3	Contextualiser la question de recherche	9
1.4	Traiter la question de recherche	10
2	État de l'art	12
2.1	Définition des concepts	12
2.1.1	La qualité des données	12
2.1.2	Le biais et le biais de données	12
2.1.3	Le traitement automatique des données par IA	13
2.2	Biais dans les offres d'emploi	14
2.2.1	Les biais genrés	16
2.2.2	Les biais de niveau de vocabulaire	18
2.2.3	Les biais d'annotation	18
2.3	Biais générés par le traitement automatique des données par IA	19
2.3.1	Le défi de la qualité des données	21
2.3.2	Le défi de la pertinence et de la représentativité des données	26
2.3.3	Le défi de la transformation des données et l'opacité des choix des algorithmes	34
2.4	Biais de l'utilité de l'IA	36
2.5	Biais des sources	38

2.6	Discussion	38
2.6.1	Détection des biais des offres d'emploi	38
2.6.2	Détection des biais du traitement automatique des données	44
2.6.3	Détection des autres biais	44
2.6.4	Synthèse de la discussion	44
2.7	Conclusion de l'état de l'art	45
3	Etude de cas	47
3.1	Méthodologie	47
3.1.1	Collecte des données	47
3.1.2	Vérification et correction de la qualité des données	51
3.1.3	Algorithme d'apprentissage supervisé	52
3.1.4	Vérification des résultats prédictifs via la visualisation	56
3.2	Vérification et correction de la qualité des données	59
3.2.1	Exploration et préparation des données	60
3.2.2	Évaluation de la pertinence des données	62
3.3	Algorithme d'apprentissage supervisé	63
3.3.1	Import et nettoyage des données	64
3.3.2	Paramétrage des données	65
3.3.3	Construction de la boucle <i>Randomforest</i>	66
3.4	Vérifier les résultats prédictifs via la visualisation	69
3.4.1	Scénarios d'hyperparamétrages	69
3.4.2	Scénarios sans et avec lemmatisation	74
3.4.3	Scénarios des pourcentages des mots les plus rares	76
3.5	Discussion	78
3.5.1	Consolidation du modèle	85
3.6	Conclusion de l'étude de cas	86
4	Recommandations	87
4.1	Atouts du prototype dans la réalité	87

4.2	Limites du prototype dans la réalité	87
5	Conclusion	89
5.1	Hypothèse	89
5.2	Traitement de la question de recherche	89
5.3	Limites de la question de recherche	91
5.4	Difficultés rencontrées	91
5.5	Perspectives	92
6	Annexes	93
6.1	Code du prototype	93
6.1.1	Fichier 1 : 10 nettoyage-paramétrage.R	93
6.1.2	Fichier 2 : 11. algorithme.R	96
6.1.3	Fichier 3 : 12 visualisation.R	98
6.1.4	Fichier 4 : 20 lemmatisation.R	100
6.1.5	Fichier 5 : 21 paramétrage.R	102
6.1.6	Fichier 6 : 30 distant reading.R	104
6.2	Bibliographie	107

1 Introduction

Ce mémoire s'inscrit dans ma volonté de mettre en place des leviers pour contribuer à un monde plus juste. L'heureuse découverte des humanités numériques lors du Master STIC me permet dans ce mémoire de Master d'identifier, grâce au *Machine Learning* (ou ML, une des techniques de l'Intelligence Artificielle (IA)), le vocabulaire d'une offre d'emploi qui la fera pencher vers un public ou vers un autre.

1.1 Exploiter utilement ma reprise d'études

Ce mémoire de Master s'inscrit dans la cadre de ma reprise d'études, à l'âge de 43 ans. Je travaillais alors à Bruxelles Formation, comme Digital expert, depuis 2018.

En 2021, j'avais notamment initié et dirigé la mise en place de l'Open Data en API du catalogue Dorifor des fiches formation. L'Open Data de Dorifor permet désormais à plusieurs sites web de publier l'extrait des données qui correspondent à leur mission spécifique : Actiris sélectionne les activités liées aux formations, pour les personnes en recherche d'emploi; la Cité des métiers de Bruxelles publie l'ensemble du catalogue des formations et des activités qui y sont liées; Bruxelles Formation affiche uniquement ses fiches formations, et celles de ses partenaires conventionnés; le Pôle Formation Emploi Digitalcity.brussels reprend les formations de son secteur et l'intègre à son catalogue généraliste; le Pôle Formation Emploi Technicity.brussels extrait brièvement les intitulés des formations de son secteur. Il est prévu que les autres Pôles Formation Emploi intègrent également les fiches Dorifor dans leurs catalogues en ligne via l'API.

En 2023, je participais au premier objectif stratégique du Contrat de gestion 2023-2027¹ qui était de “faciliter l'accessibilité, la lisibilité et la simplification de l'information, de l'inscription, des processus et des services, tant pour les usagers que pour les prescripteurs (conseillers d'Actiris, CPAS, OISP, BAPA, . . .)”.

Une fois les tuyaux de l'API mis en place, je devais donc m'attaquer à l'accessibilité de son contenu. Je disposais déjà de deux outils en ligne : un outil d'évaluation du

1. Contrat de gestion 2023-2027 de Bruxelles Formation, Bruxelles Formation, URL : <https://bruxellesformation.brussels/> (visité le 11/02/2024).

niveau de langage, édité par Sapiens UX, et un dictionnaire des synonymes inclusifs, édité par l’Inclupédie. Comme dernier filet de sécurité à ajouter aux CMS d’encodage, je réfléchissais à utiliser la prédiction du ML pour guider les équipes dans les biais qu’elles pourraient introduire dans leurs choix de mots et leurs tournures de phrases : l’orientation de mon sujet de mémoire était donc fixée.

1.2 Construire ma question de recherche

Pour éviter de connecter directement mon mémoire d’étudiante à mon emploi, je voulais m’exercer sur un sujet connexe, qui pourrait néanmoins devenir utile aux institutions bruxelloises pour lesquels je travaillais.

D’une part, j’avais lu dans la presse² qu’Actiris, sous l’impulsion de son ministre de tutelle, avait fait appel au laboratoire bruxellois FARI, cofondé par l’ULB et la VUB, pour utiliser l’IA. L’objectif était de tenter d’extraire via ML les compétences demandées dans les offres d’emploi, afin de trouver une correspondance avec les compétences encodées par les personnes en recherche d’emploi dans MyActiris. J’ai pris contact avec le cabinet du ministre, également en charge de Bruxelles Formation, pour évaluer quel sujet de mémoire pourrait apporter sa pierre à l’édifice. En juillet 2022, le cabinet m’a indiqué que la lutte contre les discriminations à l’embauche faisait partie des objectifs prévus, et nous en avons conclu que ma contribution via l’identification des éléments discriminants dans les offres d’emploi par le ML serait la bienvenue.

D’autre part, les médias français s’étaient fait l’écho en septembre 2022³ d’une plainte syndicale selon laquelle une majorité d’offres d’emploi publiée par Pôle Emploi n’était pas pertinente : la CGT affirmait que 76% des offres d’emploi diffusées par Pôle Emploi étaient illégales car ne respectaient pas la législation, elles étaient inexistantables pour la recherche d’emploi, car obsolètes, incomplètes, incohérentes, ou erronées. Les sites web privés, “agences publicitaires de l’emploi”, fournissent à Pôle

2. Pierre-François LOVENS : La Région bruxelloise va se doter d’un pôle d’excellence en intelligence artificielle, in : La Libre.be, 25 jan. 2022, (visité le 05/03/2023).

3. France TELEVISIONS : Pôle emploi : une majorité d’offres non-fiables ?, 2022, URL : <https://www.francetvinfo.fr> (visité le 05/03/2023).

Emploi 90% des offres non fiables, souvent laissées en ligne mais déjà pourvues, surtout dans le bâtiment (en 2024, ce taux est descendu à 61%⁴).

Par ailleurs, j'avais été interpellée par la présentation au FNRS de Laurence Dierickx sur « Analyse critique et amélioration de la qualité de l'information numérique » du 18/05/2022⁵, lorsqu'elle avait présenté le cycle de vie des données en ML d'Alberto Marocchino (Figure 1).

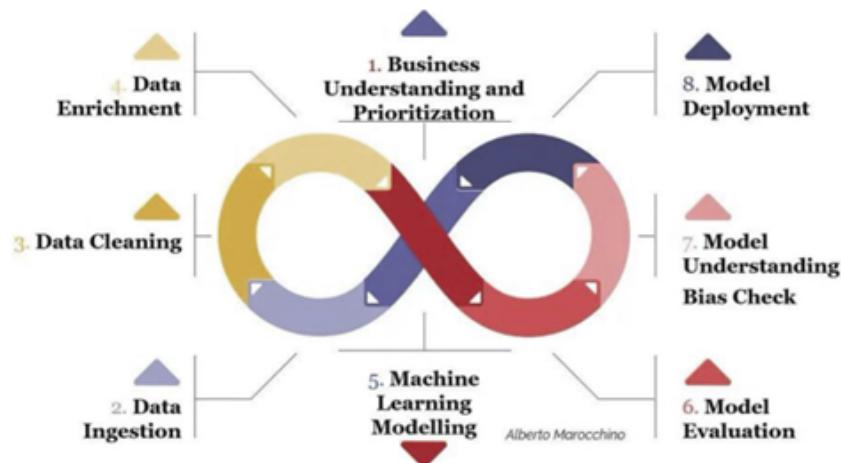


FIGURE 1 – Cycle de vie des données en ML

Ce schéma m'a paru illustrer de manière particulièrement pertinente les défis professionnels que je rencontrais. En fin de présentation, Laurence Dierickx a suggéré un certain nombre d'idées de recherche. J'ai décidé de m'emparer du thème de l'identification des indicateurs de qualité pour chaque étape d'un processus de ML, et de me pencher particulièrement sur le cas des offres d'emploi en français.

Via mon réseau de connaissances bruxellois, j'ai appris que la startup Interskillar s'était justement spécialisée sur la question. J'ai pris contact avec sa CEO, Leila Maidane, qui m'a expliqué leur principe d'évaluation des biais d'une offre d'emploi : plus une offre d'emploi reçoit des profils de candidature similaires, plus elle est biaisée ; et plus une offre d'emploi reçoit des profils de candidature diversifiés, moins elle est bai-

4. Cyprien BOGANDA : Le scandale des offres bidons : 61 % des offres d'emploi de France Travail sont illégales - L'Humanité, in : 18 jan. 2024, (visité le 19/01/2024).

5. Laurence DIERICKX : Apprentissage automatique : les challenges de la qualité des données dans la perspective d'une adéquation aux usages, Réunion du groupe de contact FNRS « Analyse critique et amélioration de la qualité de l'information numérique », 2022.

sée. Elle m'a également fourni une liste de critères discriminants permettant d'évaluer la similarité des profils.

Je disposais enfin de tous les éléments me permettant de formuler ma question de recherche :

Quels sont les indicateurs à considérer pour détecter les biais dans les offres d'emploi, en utilisant un processus de ML?

1.3 Contextualiser la question de recherche

Avant toute chose, il convient de préciser les différentes étapes menant à un recrutement, afin d'isoler les termes de l'étude.

On peut ainsi dissocier : la formulation du besoin d'embaucher (qualités, compétences, expérience), la rédaction de l'offre d'emploi (langage, forme de l'offre), la publication de l'offre d'emploi (canaux de diffusion), le processus de sélection (critères primaires) et le processus de recrutement (critères secondaires).

La mixité des publics dans un secteur d'emploi améliore sa créativité et sa productivité⁶. Il existe donc de véritables enjeux publics à diminuer les biais d'emploi, qui créent une scission entre les publics "traditionnels" (hommes blancs jeunes) et les publics "nouveaux" (toutes et tous les autres)⁷.

L'IA est actuellement utilisée dans les recrutements, et y reproduit ces biais systémiques⁸. Plusieurs leviers d'action convergents sont en cours de construction pour limiter ces biais, notamment des engagements éthiques⁹, des législations¹⁰ (notamment européennes¹¹), des normes, des règles métier, ...

6. Jonathan D. OSTRY et al. : Economic Gains From Gender Inclusion : New Mechanisms, New Evidence, in : IMF eLibrary 2018, ISBN : 9781484337127, (visité le 11/02/2024).

7. Océane COUILLAUD : Le Genre Dans l'Education Entrepreneuriale : Une Analyse Exploratoire Inspirée De La Fouille De Texte, 2020, (visité le 22/07/2023).

8. Philippe BESSE : Déetecter,évaluer les risques des impacts discriminatoires des algorithmes d'IA, mai 2020, (visité le 11/02/2024).

9. AI Assessment Tool, URL : <https://altaia.ai4belgium.be/> (visité le 28/03/2023).

10. Amanda LEVENDOWSKI : How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, in : Washington Law Review 93.2 (2018), p. 579-630, (visité le 05/08/2023).

11. EU AI Act : first regulation on artificial intelligence | Topics | European Parliament, (visité le

Par ailleurs, les biais qui amènent un public à postuler ou pas à une offre d'emploi sont multifactoriels. Pour ce mémoire, nous nous intéressons uniquement au langage utilisé dans la rédaction de l'offre d'emploi elle-même, et écartons donc également les (nombreuses) études liées au processus même du recrutement.

1.4 Traiter la question de recherche

Ce mémoire est divisé en plusieurs parties.

Dans la partie de l'état de l'art, nous définirons les concepts des termes liés à la question de recherche, nous énumérerons les biais dans les offres d'emploi, et les biais générés par le traitement automatique des données. Nous évaluerons également les biais de l'utilité même de l'Intelligence Artificielle pour traiter notre question de recherche, ainsi que les biais ajoutés par les sources citées dans ce mémoire. Enfin, nous discuterons de ces différents types de biais, et identifierons les solutions déjà existantes pour les détecter. Les résultats de cette discussion nous guideront dans le développement de l'étude de cas. Les concepts à définir seront la qualité des données, le biais et le biais de données, ainsi que le traitement automatique de corpus. Parmi les biais dans les offres d'emploi, nous aborderons les biais genrés, les biais de niveau de vocabulaire et les biais d'annotation. Pour les biais générés par le traitement automatique des données, nous suivrons les défis identifiés par Joseph F. Hair Jr. et Marko Sarstedt¹², à savoir le défi de la qualité des données, le défi de la pertinence et de la représentativité des données, et le défi de la transformation des données et l'opacité des choix des algorithmes.

Dans la partie de l'étude de cas, nous développerons un prototype d'Intelligence Artificielle afin de tenter de répondre à la question de recherche. Nous décrirons la méthodologie utilisée sur l'ensemble des étapes de développement du prototype : pour la collecte des données, la vérification et la correction de la qualité des données, l'algorithme

11/02/2024).

12. Joseph F. Jr. HAIR/Marko SARSTEDT : Data, measurement, and causal inferences in machine learning : opportunities and challenges for marketing, in : Journal of Marketing Theory and Practice 2021, (visit  le 25/09/2023).

rithme d'apprentissage supervisé, et comment vérifier les résultats prédictifs via la visualisation. Chaque partie sera ensuite développée. Pour la vérification et la correction de la qualité des données, nous détaillerons l'exploration et la préparation des données, puis l'évaluation de la pertinence des données. L'algorithme d'apprentissage supervisé sera abordé en trois étapes : l'étape du nettoyage des données, celle du paramétrage des données, et enfin celle de la construction de la boucle d'apprentissage. Enfin, la vérification des résultats prédictifs via la visualisation abordera les scénarios d'hyperparamétrage, les scénarios sans et avec lemmatisation, et les scénarios de pourcentages des mots les plus rares. Nous discuterons ensuite des résultats du prototype, mis en perspectives avec des sources extérieures.

Enfin, nous formulerais des recommandations d'utilisation du prototype dans la réalité, avec ses atouts et ses limites.

2 État de l'art

Le sujet de ce mémoire se trouve au croisement entre la qualité des données, la définition de ce qu'est un biais, et les problématiques du traitement automatique des données. A partir des normes et références disponibles, l'état de l'art devra identifier les biais possibles dans ces trois domaines, et évaluer lesquels sont mesurables, et comment.

2.1 Définition des concepts

2.1.1 La qualité des données

Dans le livre de référence Data Quality¹³, la qualité des données est définie comme l'adaptation des données à leur utilisation ("fitness for use"), selon 4 critères : leur qualité intrinsèque (exactitude, objectivité, crédibilité, réputation), leur accessibilité (accessibilité, sécurité d'accès), leur contexte (pertinence, valeur ajoutée, actualité, exhaustivité, quantité d'informations), leur représentativité (interprétabilité, facilité de compréhension, représentation concise, représentation cohérente, facilité de manipulation).

Selon Laurence Dierickx, "la problématique de la qualité des données concerne l'ensemble du processus : jeu de données en entrée, représentation du modèle, évaluation et précision, recherche du meilleur modèle"¹⁴

Dans ce mémoire, le concept "fitness for use" sera central.

2.1.2 Le biais et le biais de données

Dans sa thèse de doctorat¹⁵, Nicolas Brault étudie le concept de biais en épidémiologie. Il distingue in fine le biais psychologique, comme antithèse de la notion d'objec-

13. Richard Y. WANG/Mostapha ZIAD/Yang W. LEE : Data Quality, t. 23 (Advances in Database Systems), Boston 2002, (visité le 21/09/2023).

14. DIERICKX : Apprentissage automatique : les challenges de la qualité des données dans la perspective d'une adéquation aux usages (cf. note 5).

15. Nicolas BRAULT : Le concept de biais en épidémiologie, thèse de doct., Université Sorbonne Paris Cité, 2017, (visité le 08/09/2023).

tivité ; le biais statistique, qui "renvoie à une erreur systématique de mesure qui n'est pas seulement liée à l'instrument" ; le biais épidémiologique, qui hiérarchise des plans d'expérience en fonction de leur niveau de preuve (il donne comme exemple l'essai clinique randomisé, plus probant qu'une étude cas-témoins) ; le biais d'interprétation, soit "toutes les explications alternatives possibles autre que l'explication en termes de causalité et de hasard, (...) y compris le biais de confusion".

En 2021, Milagros Miceli et al.¹⁶ précisent que "le biais de données a été défini comme "une distorsion systématique des données" qui peut être mesurée en "comparant un échantillon de données de travail avec des échantillons de référence provenant de sources ou de contextes différents". Cette définition contient un principe important : il existe une valeur de vérité absolue dans les données et le biais est une "distorsion" par rapport à cette valeur. Les données, tout comme la vérité, sont le produit de relations sociales subjectives et asymétriques. La documentation peut aider à "diagnostiquer les sources de biais" et peut potentiellement "atténuer les biais indésirables dans les systèmes d'apprentissage automatique".

Dans les parties suivantes, nous nous attacherons à préciser quels sont les biais intéressants pour répondre à notre question de recherche.

2.1.3 Le traitement automatique des données par IA

Dans *Machine Learning and Artificial Intelligence*¹⁷, Ameet V. Joshi contextualise l'IA :

Dans une perspective moderne, lorsque nous parlons d'IA, nous entendons des machines capables d'effectuer une ou plusieurs des tâches suivantes : comprendre le langage humain, effectuer des tâches mécaniques impliquant des manœuvres complexes, résoudre des problèmes complexes basés sur l'ordinateur et impliquant éventuellement de grandes quantités de données en un temps très court et revenir avec des réponses à la ma-

16. Milagros MICELI/Julian POSADA/Tianling YANG : Studying Up Machine Learning Data : Why Talk About Bias When We Mean Power?, in : arXiv.org; Ithaca 2021, (visité le 25/09/2023).

17. Ameet V JOSHI : Machine Learning and Artificial Intelligence, Cham 2023, (visité le 21/09/2023).

nière d'un humain, etc. (...) Le terme *Machine learning* (ML) a été inventé en 1959 par Arthur Samuel dans le contexte de la résolution d'un jeu de dames par une machine. Ce terme fait référence à un programme informatique capable d'apprendre à produire un comportement qui n'est pas explicitement programmé par l'auteur du programme. Au contraire, il est capable de montrer un comportement dont l'auteur peut être totalement inconscient. (...) Les méthodes de la théorie de ML sont essentielles pour construire des systèmes artificiellement intelligents.

Dans ce mémoire, nous nous interrogerons sur les avantages et les inconvénients du traitement automatique des données par IA pour répondre à notre question de recherche.

2.2 Biais dans les offres d'emploi

La construction de nos habitudes de notre expression individuelle s'est réalisée dans un certain contexte culturel, socio-économique et de position genrée. Notre manière de nous exprimer reflète généralement ce contexte. De la même manière, la formulation des offres d'emploi révèle les habitudes du cadre dans lequel elles sont rédigées. Et qui dit habitudes, dit biais.

Dans *Semantics derived automatically from language corpora contain human-like biases*¹⁸, Caliskan et al. le confirment :

L'idée générale selon laquelle les corpus de textes capturent la sémantique, y compris les stéréotypes culturels et les associations empiriques, est connue depuis longtemps dans le domaine de la linguistique de corpus.

Il existe des études et des articles qui listent les termes excluants pour les offres d'em-

18. Aylin CALISKAN/Joanna J. BRYSON/Arvind NARAYANAN : Semantics derived automatically from language corpora contain human-like biases, in : Science 356.6334 (2017), Publisher : American Association for the Advancement of Science, p. 183-186, (visité le 05/08/2023).

ploi : vocabulaire sexué et ethnique¹⁹, le tutoiement²⁰, les contenus trompeurs ou abusifs²¹, les mots-clés classistes²².

Le Dr. Moses Isooba, directeur général du Forum national des ONG d'Ouganda, fait d'ailleurs partie d'une équipe de développement de langue et de prototype lexical utilisant l'IA pour développer la communication inclusive, "qui est plus ou moins implicitement néocoloniale, sexiste ou raciste²³. L'idée est d'explorer l'application de l'IA pour redresser et remplacer le lexique du jargon péjoratif, les idiomes et la terminologie « imposés » par les organisations de la société civile impliquées dans l'aide complexe internationale et le secteur du développement.". Par exemple :

"Les termes tels que « desk officer » (Responsable de secteur) ou « in the field » (sur le terrain) évoquent l'époque coloniale. Les organisations humanitaires internationales font souvent référence à leurs bureaux dans les pays du Sud comme «country offices » (bureaux de pays) semblables à des avant-postes coloniaux «acting as the remaining rope tying us to the colonial ship » (qui fait office d'unique lien nous rattachant au navire colonial)"

Dans son rapport "Qualité des données et intelligence artificielle - atténuer les biais et les erreurs pour protéger les droits fondamentaux"²⁴, l'Agence de l'Union européenne pour les droits fondamentaux mentionne : "Les données générées sur l'internet sont nécessairement non représentatives de certains groupes de la population. Il s'agit notamment des pays du sud et de l'est de l'UE, des familles les plus pauvres, des personnes âgées et des personnes ayant un faible niveau d'éducation, en particulier les

19. Nikhil GARG et al. : Word embeddings quantify 100 years of gender and ethnic stereotypes, in : Proceedings of the National Academy of Sciences 115.16 (2018), Publisher : Proceedings of the National Academy of Sciences, E3635-E3644, (visité le 05/03/2023).

20. Alex ALBER : Tutoyer son chef. Entre rapports sociaux et logiques managériales, in : Sociologie du travail 61.1 (2019), Number : 1 Publisher : Association pour le développement de la sociologie du travail, (visité le 13/05/2023).

21. ALSHYMI : offre d'emploi : les avantages qui sont des obligations légales. Twitter, 4 fév. 2023, (visité le 05/03/2023).

22. Sophie GOURION : Red flags : ces offres d'emploi qui font fuir les candidats (et spécialement les femmes), 2021, (visité le 05/03/2023).

23. Pierre-Emmanuel FARRET : Utiliser l'intelligence artificielle pour « décoloniser » le langage, in : Global Voices Online, 20 mars 2023, (visité le 21/03/2023).

24. Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights, European Union Agency for Fundamental Rights, 2019.

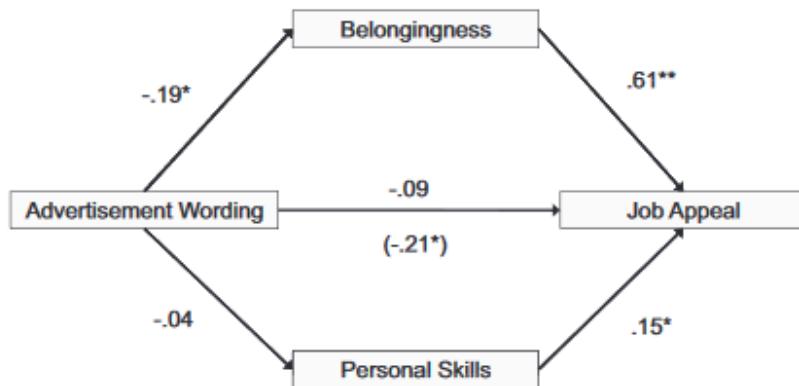


Figure 4. Standardized path estimates from Study 5 (all female participants). Advertisement wording coded as 1 = masculine wording and 0 = feminine wording. * $p < .05$. ** $p < .001$.

FIGURE 2 – Impact du vocabulaire genré sur l’attractivité d’une offre d’emploi sur les femmes

femmes.”

Nous allons maintenant nous concentrer sur 3 types de biais : les biais genrés, les biais de niveau de vocabulaire, les biais d’annotation du vocabulaire.

2.2.1 Les biais genrés

En 2011, Danielle Gaucher et Justin Friesen²⁵ ont démontré que le vocabulaire des offres d’emploi genré a un réel effet d’exclusion sur les femmes (-13.95%), et participe au renforcement des stéréotypes de genre des métiers (Figure 2) :

“En moyenne, 23 % des représentations mentales sont féminines après l’utilisation d’un générique masculin, alors que ce même pourcentage est de 43 % après l’utilisation d’un générique épicène. La différence varie presque du simple au double.”

En 2020, Océane Couillaud, dans son mémoire de Master “Le genre dans l’éducation entrepreneuriale : une analyse exploratoire inspirée de la fouille de textes”²⁶, s’est appuyée sur l’étude de 2011 et sur celle de Sandra Bem de 1974. Elle écrit :

25. Danielle GAUCHER/Justin FRIESSEN/Aaron C. KAY : Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality, 2011, (visité le 27/03/2023).

26. COUILAUD : Le Genre Dans l’Education Entrepreneuriale (cf. note 7).

Une des recherches fondatrices sur les mots considérés comme masculin et féminin est celle de Bem (1974). Sandra Bem, psychologue américaine, a développé le *Bem Sex-Role Inventory* (BSRI), un index dont les items sont utilisés pour décrire la masculinité et la féminité. Bem se base sur la conception du genre comme construction sociale, c'est-à-dire que la masculinité et la féminité sont vues comme deux construits différents représentant l'internalisation de comportements, ou de traits de personnalité, jugés comme désirables pour un homme, ou pour une femme, par la société (Bem, 1974). Pour déterminer ceci, l'examineur demande au sujet d'évaluer, pour chaque trait de personnalité, dans quelle mesure ce trait est désirable, grâce à une échelle Likert à 7 points d'ancrage. Ainsi, dans le BSRI, un item est considéré comme étant masculin s'il est plus désirable pour un homme de posséder ce trait que pour une femme. Par exemple, les mots « *aggressive* », « *dominant* », « *willing to take a stand* », représentent des items masculins du BSRI. À l'inverse, un item est qualifié de féminin s'il est jugé plus désirable pour une femme que pour un homme dans notre société, comme les mots « *compassionate* », « *loyal* » et « *sensitive to the needs of others* ». En d'autres termes, le BSRI représente ce que la société considère généralement comme des traits typiquement masculins ou féminins.(...)

Malgré sa date de publication, cet index reste largement utilisé dans la recherche contemporaine, particulièrement dans les études de genre qui s'intéressent au langage (Ahl, 2004, 2006; Gaucher et al., 2011; Jones & Warhuus, 2018). Par ailleurs, l'étude de Bem a été menée aux États-Unis, et on pourrait s'attendre à ce que les construits changent selon la culture, donc dans un contexte de culture québécoise. Néanmoins, on peut mentionner l'étude de Persson (1999) et celle de Carver, Vafaei, Guerra, Freire et Phillips (2013) qui ont validé l'utilisation du BSRI dans d'autres cultures.

Nous pouvons en conclure que le vocabulaire genré dans les offres d'emploi produit

un effet d'exclusion d'une partie des femmes.

2.2.2 Les biais de niveau de vocabulaire

L'accumulation des enseignements scolaires permet de développer le vocabulaire d'une personne. Ainsi, plus on aura étudié dans une langue, plus on disposera du niveau de vocabulaire adapté.

Une personne qui a acquis des compétences en-dehors d'un enseignement scolaire ou dans un parcours scolaire différent (parcours atypique, dans une autre langue, voire dans la même langue mais dans un pays à la culture distante...), dispose mécaniquement d'un niveau de vocabulaire différent de celui imposé par le cadre de l'offre d'emploi, et peut donc rencontrer des difficultés à en comprendre tous les termes²⁷. Par ailleurs, les personnes victimes d'un ou plusieurs handicaps sociaux (minorités, neuro atypes, troubles dys, post-AVC²⁸...) peuvent également rencontrer des difficultés de compréhension²⁹ de ces offres d'emploi³⁰.

2.2.3 Les biais d'annotation

Les offres d'emploi mal annotées ou annotées avec un vocabulaire excluant³¹ freinent leur diffusion en ligne par manque de fiabilité³², d'indexabilité et par manque de découvrabilité.

Dans *Data Statements for Natural Language Processing : Toward Mitigating System Bias and Enabling Better Science*³³, Emily Bender et Batya Friedman expliquent :

27. Référentiel des compétences initiales, Enseignement.be, (visité le 18/02/2024).

28. Didier CASTIEL/Pierre-Henri BRÉCHAT : 1. « L'économie de la discrimination » de Gary Becker : une approche au service d'une politique de réduction des handicaps sociaux, in : Solidarités, précarité et handicap social (Hors collection), Rennes 2010, p. 81-92, (visité le 18/02/2024).

29. M.-L. LOPEZ : Les "Handicapés sociaux" et leur resocialisation : Diversité des pratiques et ambiguïté de leurs effets, t. 2, Company : Persée - Portail des revues scientifiques en SHS Distributor : Persée - Portail des revues scientifiques en SHS Institution : Persée - Portail des revues scientifiques en SHS Label : Persée - Portail des revues scientifiques en SHS Publisher : Editions Médecine et Hygiène, 1978, (visité le 18/02/2024).

30. Michel ABHERVÉ : Il faut ouvrir le débat sur le "handicap social". Les blogs d'Alternatives Économiques, 10 fév. 2014, (visité le 18/02/2024).

31. Emmanouil Antonios PLATANIOS et al. : Learning from Imperfect Annotations, 2020, (visité le 05/03/2023).

32. Isabelle BOYDENS : Open Data et eGovernment, in.

33. Emily M. BENDER/Batya FRIEDMAN : Data Statements for Natural Language Processing : Toward

Les déclarations de données devraient nous aider, en tant que domaine, à aborder les questions éthiques de l'exclusion, de la surgénéralisation et de la sous-exposition (Hovy et Spruit, 2016). En outre, comme les déclarations de données permettent de mieux cibler nos ensembles de données et leurs populations représentées, elles devraient également nous aider, en tant que domaine, à traiter les questions scientifiques de généralisation et de reproductibilité. L'adoption de cette pratique nous permettra de mieux comprendre et décrire nos résultats et, en fin de compte, d'améliorer la science et l'ingénierie et de les rendre plus éthiques. (...) Les déclarations de données ne résolvent pas en elles-mêmes l'ensemble du problème de la partialité. Elles constituent plutôt une infrastructure habilitante essentielle. Prenons par analogie cet exemple de Friedman (1997) sur l'accès à la technologie et à l'emploi pour les personnes handicapées. (...) De même, en ce qui concerne les biais dans la technologie NLP, si nous ne nous engageons pas en faveur des déclarations de données ou d'une pratique similaire pour rendre explicites les caractéristiques des ensembles de données, nous compromettrons à nous seuls la capacité du domaine à traiter les biais.

Nous pouvons conclure de ces sources que les biais d'annotation peuvent limiter l'accès aux offres d'emploi, et donc générer un biais d'accessibilité.

2.3 Biais générés par le traitement automatique des données par IA

Pour pouvoir répondre à la question de recherche et "s'appuyer sur l'IA pour identifier les biais dans la rédaction des offres d'emploi", il est nécessaire d'appliquer un certain nombre de traitements à leurs données. Ce chapitre s'attache à identifier les types de biais qui pourraient être générés par ces traitements.

Dans leur article "*Data, measurement, and causal inferences in machine learning* :

Mitigating System Bias and Enabling Better Science, in : Transactions of the Association for Computational Linguistics 6 (2018), p. 587-604, (visité le 08/08/2023).

*opportunities and challenges for marketing*³⁴, Joseph F. Hair Jr. et Marko Sarstedt incitent à la plus grande prudence dans les biais générés par le ML :

“S'il ne fait aucun doute que les données numériques apportent d'énormes avantages, leur utilisation s'accompagne également de défis considérables. Ceux-ci concernent **(1) la qualité des données, (2) la densité des données, (3) leur efficacité pour mesurer des phénomènes non observés, et (4) la transformation.** (...) L'appel à l'augmentation de la facilité d'utilisation des données suggère que l'avantage fréquemment exprimé selon lequel une approche de l'enquête scientifique axée sur les données offre un compte rendu plus objectif des phénomènes devrait être considéré avec prudence. La pratique réelle de la science implique une centaine de décisions subjectives - l'objectivité pure est une caractérisation erronée (Brownstein et al., 2019). Cela vaut non seulement pour le choix et la mise en œuvre des modèles de recherche ou le choix des mesures, mais aussi pour la conception des algorithmes ou la manière dont les données sont *scrappées*. Ou, comme le note Hales (2013), "tout test statistique ou algorithme d'apprentissage automatique exprime un point de vue sur ce qu'est un modèle ou une régularité, et toute donnée a été collectée pour une raison basée sur ce qui est considéré comme approprié à mesurer". Un algorithme trouvera un type de modèle et un autre en trouvera un autre. Un ensemble de données mettra en évidence certains modèles et pas d'autres."

Nous nous proposons d'emprunter la grille d'analyse développée par Hair et Sarstedt pour nous interroger sur (1) la qualité des données traitées par le ML, sur (2) l'efficacité et la densité des données traitées (leur pertinence et leur représentativité), et enfin (3) sur la transformation des données et l'opacité des critères de choix de modèle par les algorithmes.

Pour rappel, dans le cadre de ce mémoire, nous suivons le postulat de la startup

34. HAIR/SARSTEDT : Data, measurement, and causal inferences in machine learning : opportunities and challenges for marketing (cf. note 12).

Interskillar : les biais d'une offre d'emploi se mesurent à la diversité des profils qui y répondent, plus l'offre d'emploi est biaisée, plus les profils des candidatures sont homogènes. Nous devons donc considérer que le processus de traitement automatique des données concerne tant les offres d'emploi en elles-mêmes, que les profils qui y répondent.

2.3.1 Le défi de la qualité des données

Pour ce premier défi, nous allons comparer les critères de qualité des données mis en perspectives du *Big data*, puis mis en perspectives par la *Business Intelligence*, soit les deux extrêmes du spectre de la qualité des données : du cadre le moins maîtrisé (le *Big data*) au cadre le mieux maîtrisé (la *Business Intelligence*).

Les critères de qualité des données mis en perspectives du *Big data*

Dans son article “*Data Quality Issues Concerning Statistical Data Gathering Supported by Big Data Technology.*”³⁵, Jacek Maślankowski identifie plusieurs catégories de critères de qualité de données, avec leurs différentes dimensions et leur importance au niveau du *Big data* (Table 1). Il appelle *Big data* "la technologie qui permet de collecter des données à partir de sources de données non structurées, principalement des sites web". Ce tableau nous offre une grille de lecture sur la qualité probable des données étudiées.

Utilisons cette grille pour analyser les critiques émises par les syndicats sur les offres d'emploi venant du *Big data*, comme celles publiées par Pôle Emploi³⁶. Selon la grille, ces offres d'emploi rencontrent probablement des difficultés à rencontrer des critères satisfaisants de précision, d'objectivité, de pertinence et de réputation : les syndicats leur reprochent en effet leur incohérence, leur incomplétude et leurs erreurs intrinsèques (dimension de précision). Atteindre les exigences des syndicats sur ces points semble difficile. A l'opposé, la grille indique que les critères contextuels, de re-

35. Jacek MAŚLANKOWSKI : *Data Quality Issues Concerning Statistical Data Gathering Supported by Big Data Technology*, in : Stanislaw KOZIELSKI et al. (éd.) : *Beyond Databases, Architectures, and Structures (Communications in Computer and Information Science)*, Cham 2014, p. 92-101.

36. TELEVISIONS : Pôle emploi : une majorité d'offres non-fiables? (cf. note 3).

TABLE 1 – Catégories de critères de qualité de données, vues par Jacek Maślankowski

Catégorie	Dimension	Niveau d'importance pour le <i>Big data</i>
Critères intrinsèques	Précision, objectivité, pertinence, réputation	Haute, plutôt difficile à atteindre
Critères contextuels	Valeur ajoutée, pertinence, datation, complétude, quantité appropriée de données	Haute, facile à atteindre
Critères de représentativité	Interprétabilité, facilité de compréhension, cohérence de représentativité, brieveté de représentativité	Moyenne, facile à atteindre
Critères d'accessibilité	Accessibilité, sécurité d'accès	Moyenne, facile à atteindre

présentativité, et d'accessibilité sont faciles à atteindre pour des offres d'emploi publiées sur le *Big data*. Les syndicats critiquant des offres d'emploi obsolètes (dimension datation) devraient pouvoir être plus facilement entendus par Pôle Emploi. Dans cet exemple, le *Fitness for use* de la qualité des données n'est donc pas atteint.

Les profils de candidature répondant aux offres d'emploi proviendront par définition de sources personnelles qualifiées hors *Big data*, et ne peuvent en conséquence être analysées selon cette grille d'analyse.

Il est par ailleurs à noter que dans le tableau de Maślankowski, la dimension "facilité de compréhension" élude totalement les biais évoqués dans ce mémoire (biais générés, biais de niveau de vocabulaire, biais d'annotation). Il le justifie de la manière suivante : "Il est assez facile d'y parvenir car, dans ce document, l'accent est mis sur les données qui sont filtrées à partir des sites web pour produire les résultats qui sont strictement définis"; selon notre compréhension, "les résultats strictement définis" sont des données structurées intégrables au système d'information existant. Selon l'auteur, l'expression "facilité de compréhension" recouvre donc surtout la notion de "facilité de manipulation" ou "facilité de réutilisation". Cette dimension, telle que présentée, sort donc du champ de ce mémoire (et révèle les biais de représentation de l'auteur).

Les critères de qualité des données mis en perspectives par la *Business Intelligence*

Alors que le *Big data* est un espace peu maîtrisé fournissant des données à la qualité

aléatoire, la *Business Intelligence* exige la manipulation de données stratégiques, donc de très hautes qualité et dans un espace extrêmement précis et maîtrisé.

A partir du point de vue de la Business Intelligence³⁷, Luisa Franchina et Federico Sergiani confirment que l'importance de la qualité des données pour le traitement automatique s'est largement accrue. L'article mentionne notamment :

- le coût pharaonique de la mauvaise qualité de données (évalué par IBM en 2016 à 3,1 trillions de dollars, uniquement pour les États-Unis, et touchant environ 40% des entreprises), avec des impacts sur les pertes de revenus et sur la diminution de la qualité de la relation clientèle;
- le danger supérieur du manque de données : données incomplètes, inconsistantes, inadéquates, dupliquées, datées;
- le risque de prendre des décisions sur des données non pertinentes ou non fiables;
- les coûts additionnés de collecter puis d'essayer de réconcilier des données de sources différentes;
- le gain de temps pour le développement et la mise en production.

Parmi les différentes méthodes d'évaluation de la qualité des données énumérées, nous retiendrons :

- La démarche de gestion de la qualité des données, qui s'appuie sur une matrice de faux positifs / faux négatifs pour déterminer un ratio suffisant de pertinence, et un pourcentage d'instances correctement classifiées;
- L'évaluation de données de *Big data*, qu'ils résument à la fraîcheur des données (la datation de Maślankowski), la pertinence, la complétude, et la cohérence;
- Les critères utilisé par le *Admiralty System* de l'OTAN : la fiabilité, et la crédibilité des données (la réputation de Maślankowski);
- La liste de vérification des bibliothécaires contre les *fake news* : l'authenticité de la preuve (qui se rapproche de la réputation vue par Maślankowski) , fiabilité

37. Luisa FRANCHINA/Federico SERGANI : High Quality Dataset for Machine Learning in the Business Intelligence Domain, in : Yixin BI/Rahul BHATIA/Supriya KAPOOR (éd.) : Intelligent Systems and Applications (Advances in Intelligent Systems and Computing), Cham 2020, p. 391-401.

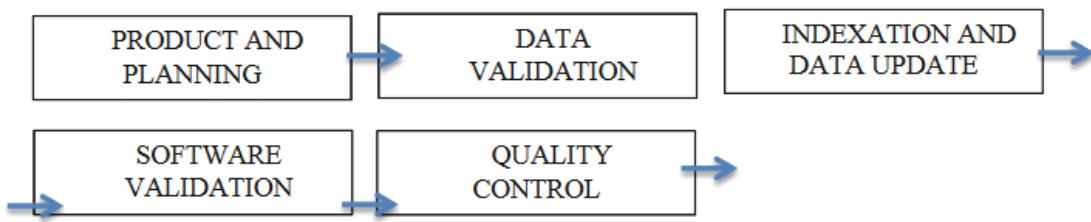


Fig. 1. Product development at Bureau Van Dijk

FIGURE 3 – Processus de contrôle qualité au Bureau Van Dijk

de la méthode de collecte, précision du mode de représentation des données par les analystes ; ce système semble le plus reconnu au monde, et certains universitaires tentent de l'automatiser avec des algorithmes sémantiques et du ML supervisé.

- La liste de vérification des services secrets canadiens pour évaluer l'open source : information-clé, précision, autorité, objectivité

Pour permettre aux organisations de prendre des décisions stratégiques éclairées, l'article insiste sur l'importance en *Business Intelligence* de s'appuyer sur un fournisseur spécialisé, unique et fiable de données intégrées, qui aura déjà fait le travail de collecte, de nettoyage, de validation, d'indexation, et de mise à jour. Il prend comme exemple le Bureau Van Dijke ("BVD"), un fournisseur de données de catégorie "A". L'article précise (Figure 3) :

"Le processus suit, avant que les données n'entrent dans le logiciel, un raffinement précis opéré par des analystes de données experts. Il convient de mentionner que la base de données BVD contient plus de 150 flux provenant de différentes sources et s'appuie sur plus de 75 fournisseurs d'informations dans le monde entier, qui fournissent des données brutes aux analystes de BVD. Les données sont obtenues directement auprès des entreprises déclarantes ou des registres officiels (dans la mesure où, dans la plupart des pays, les entreprises sont tenues par la loi de soumettre ce type de données aux registres officiels). Dans certains cas, ces informations sont collectées auprès de fournisseurs associés et vérifiés qui collectent des

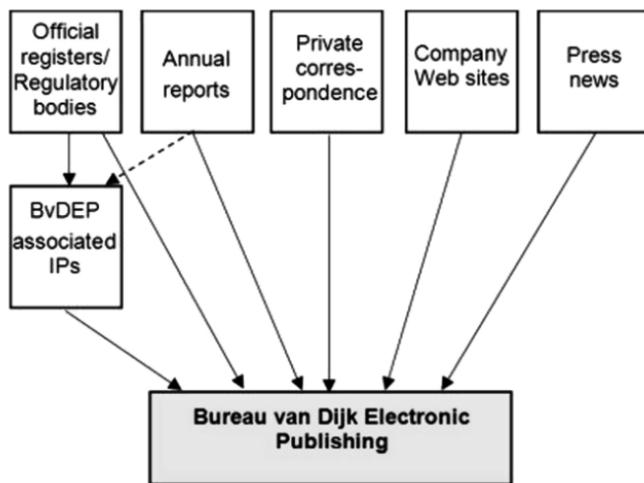


Fig. 2. Sources input at Bureau Van Dijk

FIGURE 4 – Types de sources de données transmises au Bureau Van Dijk

données auprès de sources officielles pour BVD (Figure 4)."

L'article explique que les données brutes sont quotidiennement nettoyées, standardisées, enrichies. Elles sont ensuite comparées pour validation avec les données précédentes, puis avec des données financières et économiques. Chaque jeu unique de données possède ses propres règles d'intégration. Enfin, les 200 millions d'enregistrements sont indexés; cette opération dure plus de 48 heures. Ces données, fournies aux formats API et REST, permettent aux clients de réaliser leurs propres tableaux de bord de *Business Intelligence*, et des modèles prédictifs sur de très larges quantités de données.

Avec un tel effort investi sur la qualité des données, l'article confirme que l'attente de faux positifs ou de faux négatifs est quasiment nulle : la complétude est parfaitement atteinte, et avec toutes les annotations du secteur. Dans cet exemple, le *Fitness for use* de la qualité des données est donc atteint.

2.3.2 Le défi de la pertinence et de la représentativité des données

Dans cette partie, nous aborderons la pertinence des données avec l'identification et le traitement des anomalies, ainsi que la représentativité des données, avec les évolutions de l'éthique et des concepts.

La pertinence des données

Dans son livre *Data Quality : The Accuracy Dimensions*³⁸, Jack E. Olson décrit plusieurs possibilités de tests pour évaluer la pertinence des données collectées.

Avec son schéma *Breakdown of data within a set of data* (Figure 5), il identifie d'une part les données pertinentes (*Accurate data*), et d'autre part les données non pertinentes (*Inaccurate data*).

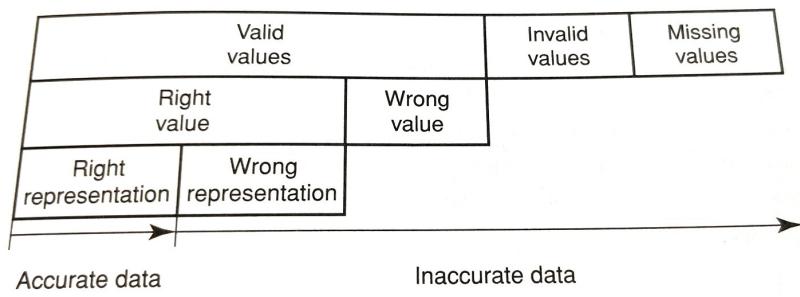


FIGURE 2.1 Breakdown of data within a set of data.

FIGURE 5 – Ventilation des données au sein d'un ensemble de données

Pour les données pertinentes, il cite les valeurs valides, justes et représentatives (*right representation*, soit les vrais positifs). Pour les données non pertinentes, il reprend les valeurs valides, justes mais non représentatives (*wrong representation*, soit les faux positifs), les valeurs valides mais fausses, les valeurs non valides, et les valeurs manquantes.

Pour les offres d'emploi et les profils de candidats, le taux de biais des sources de données peut être calculé comme le rapport entre les valeurs valides, justes mais non représentatives (*wrong representation*, faux positifs) sur l'ensemble des valeurs valides et justes. On retombe alors sur la matrice de faux positifs / faux négatifs de la démarche

38. Jack E. OLSON : Data Quality : The Accuracy Dimension, Google-Books-ID : x8ahL57VOtcC, 2003.

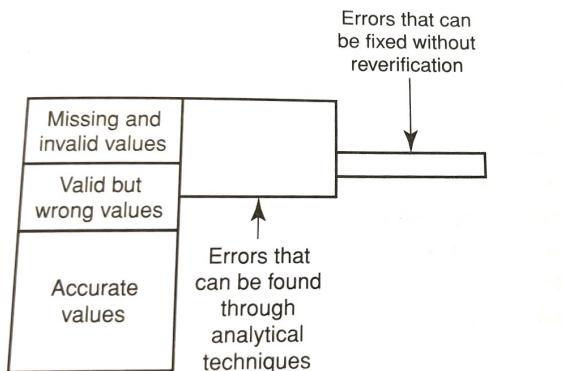


FIGURE 2.2 Chart of accurate/inaccurate values and those that are findable and fixable.

FIGURE 6 – Tableau des valeurs pertinentes/non-pertinentes, et de celles qui peuvent être trouvées et corrigées

de gestion de la qualité des données, évoquée par Luisa Franchina et Federico Sergiani³⁹.

Dans son *Chart of accurate/inaccurate values and those that are findable and fixable* (Figure 6), Olson⁴⁰ différencie les valeurs pertinentes, non-pertinentes et celles qui sont trouvables et corigeables.

Parmi les valeurs erronées, il distingue les erreurs qui peuvent être trouvées à travers les techniques analytiques, et les erreurs qui peuvent être corrigées sans revérification.

Dans le cas des offres d’emploi de Pôle Emploi critiquées par les syndicats, les offres d’emploi obsolètes ou incomplètes font partie des erreurs qui peuvent être automatiquement corrigées sans revérification, en étant tout simplement dépubliées, voire supprimées. Les offres d’emploi incohérentes ou erronées doivent quant à elles être analysées par du personnel qualifié avant de pouvoir être modifiées.

Mais ces types d’erreurs ne font pas partie du champ qui nous intéresse dans ce mémoire : pour notre objectif *Fitness for use*, nous devons ajouter à ce dernier schéma le cas des anomalies et des offres d’emploi biaisées, qui ne sont pas des erreurs à proprement parler, mais représentent pourtant bien des données à corriger. On parle donc ici du biais de représentativité des données.

39. FRANCHINA/SERGIANI : High Quality Dataset for Machine Learning in the Business Intelligence Domain (cf. note 37).

40. OLSON : Data Quality (cf. note 38).

Anomalies de sélection Dans *Data Quality - Concepts, Methodologies and Techniques*⁴¹, Carlo Batini et Monica Scannapieco mentionnent que les anomalies se retrouvent au niveau de la sélection des données retenues. Ce sont soit des données manquantes, soit des données mal encodées ou mal annotées.

Mustafa Alabadia et al.⁴² ont réalisé une synthèse sur les atouts et les limites du ML pour compenser les données manquantes au corpus. Celle-ci indique que les approches ML peuvent traiter des taux d'absence élevés tout en maintenant une faible erreur, quelle que soit la taille de l'ensemble de données. En outre, les méthodes d'imputation par ML se sont avérées capables de gérer l'incertitude et les données bruyantes tout en tenant compte des liens et des relations entre les données.

Pour les offres d'emploi comme pour les candidatures, nous pouvons donc en conclure que le traitement par ML d'un jeu comportant des données manquantes ne génère pas de biais supplémentaires par rapport au traitement par ML d'un autre jeu de données.

Anomalies d'encodage Envisageons maintenant les biais générés par le traitement par ML des anomalies de données mal annotées.

Pour Luisa Franchina et al.⁴³, l'avenir de la *Business Intelligence* est l'efficacité et la rapidité à traiter de grands volumes de données provenant de multiples sources et sous de multiples formats : pour eux, les standards de métadonnées sont cruciaux. L'article précise que le FMI prévoit une utilisation efficace de la combinaison entre données et ML pour les indicateurs de sentiments, les signaux commerciaux, et la détection de fraudes (Figure 8).

Traitements des anomalies Les anomalies peuvent être traitées en amont du système d'information, avec une approche préventive, ou en aval, une fois que les données sont ingérées par le système d'information, avec une approche curative.

41. Carlo BATINI/Monica SCANNAPIECO : Data Quality - Concepts, Methodologies and Techniques (Data-Centric Systems and Applications), 2006, (visité le 05/03/2023).

42. Mustafa ALABADLA et al. : Systematic Review of Using Machine Learning in Imputing Missing Values, in : IEEE Access 10 (2022), Conference Name : IEEE Access, p. 44483-44502.

43. FRANCHINA/SERGIANI : High Quality Dataset for Machine Learning in the Business Intelligence Domain (cf. note 37).

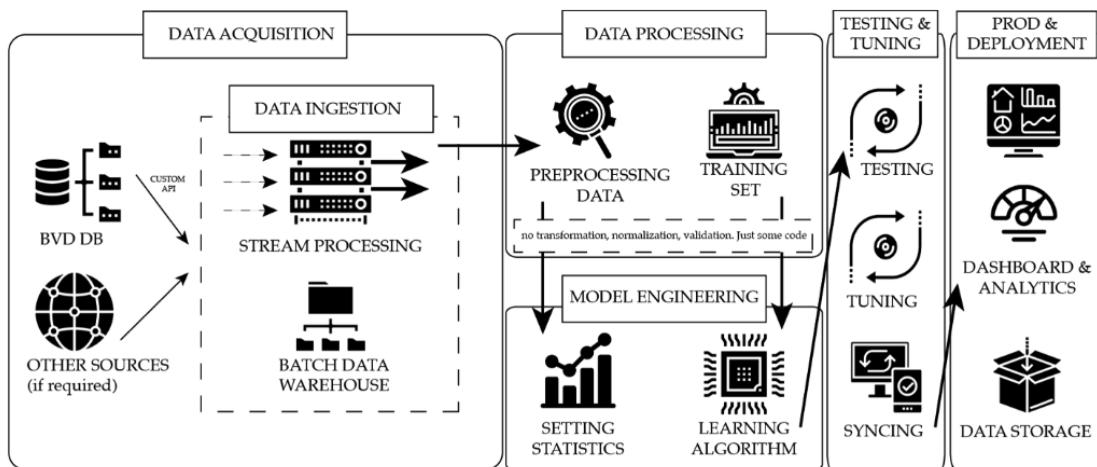


Fig. 4. ML and BI. A possible real-world application.

FIGURE 7 – ML et BI. Une application possible dans le monde réel

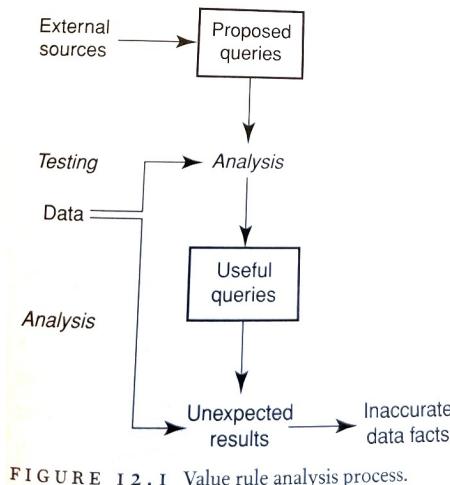


FIGURE 8 – Processus de règles de validation

En approche préventive, Isabelle Boydens et Gani G. Hamiti⁴⁴ recommandent la mise en place d'un système de détection des anomalies. Les anomalies peuvent être rejetées automatiquement à l'encodage quand elles transgressent les contraintes d'intégrité ou des règles métier; elles peuvent également être détectée par un logiciel spécialisé ATMS (*Anomalies & Transactions Management System*).

Pour l'approche corrective, Olson⁴⁵ schématise son processus de règles de validation (*Value rule analysis process*).

44. Isabelle BOYDENS/Gani G. H. HAMITI : Typologie des anomalies, un cadre pour l'action : le cas du machine learning, in : 2022, (visité le 05/03/2023).

45. OLSON : Data Quality (cf. note 38).

Parmi les requêtes proposées (*Proposed queries*) par les sources des données externes, ce processus identifie les données inappropriées ("*innacurate data facts*") suite à l'analyse des requêtes qualifiées d'utiles (*Useful queries*) et d'isoler, parmi les résultats inattendus (*Unexpected results*), ceux qui peuvent être qualifiés de données inappropriées, soit des anomalies.

Pour l'approche corrective encore, Boydens et al.⁴⁶ mentionnent les *data quality tools* et le *backtracking*, pour permettre d'identifier les étapes de perte de qualité des données. Ensemble avec Isabelle Corbesier⁴⁷, ils définissent les *data quality tools* comme interdépendants pour traiter de manière cyclique et itérative, en synergie avec les spécialistes du cœur de métier (le "*Business*"), les processus de *profiling*, de standardisation et de *matching*:

- *profiling* : analyser qualitativement et quantitativement des données pour en évaluer la qualité, isoler ou quantifier des problèmes déjà connus mais dont l'ampleur n'a jamais été évaluée et, souvent, débusquer automatiquement et semi-automatiquement des problèmes inattendus. Exemple : distribution de la longueur des valeurs d'une colonne, inférence de type, vérification ou découverte de dépendances fonctionnelles;
- standardisation : transformer les données en vue de les conformer à un standard défini avec le business ou à un référentiel existant ("*data cleansing*"), pouvant être fourni avec l'outil. Exemple : nettoyage et uniformisation de la représentation des numéros de téléphone, correction, enrichissement et validation d'adresses postales. (...);
- comparaison, détection d'incohérences et dédoublonnage, via des algorithmes de *matching* (qui se déclinent en familles bien spécifiques sur le plan théorique) : détecter les dupliques et incohérences dans les enregistrements au sein d'un jeu de données ou entre plusieurs (is-

46. BOYDENS/HAMITI : Typologie des anomalies, un cadre pour l'action (cf. note 44).

47. Isabelle BOYDENS/Gani G. H. HAMITI/Corbesier I. C. ISABELLE CORBESIER : Data Quality Tools : retours d'expérience et nouveautés, in : 7 déc. 2021, (visité le 09/03/2024).

sus potentiellement de bases de données distinctes, en vue d'une intégration ou dans le cadre d'un reengineering, par exemple). La comparaison se base sur des colonnes discriminantes et des algorithmes tolérants à l'erreur (mesure de la distance d'édition, comparaison de l'empreinte phonétique, etc.), déterminés avec le *business*. Les outils les plus avancés permettent ici de conserver et lier les enregistrements originaux pertinents (après validation par le *Business*) sans les écraser. Les meilleures valeurs identifiées pour chaque colonne serviront à construire le «*survivor* » ou « *golden record* », représentant chaque grappe ainsi repérée et utilisé pour dédoublonner le(s) jeu(x) de données si nécessaire. Notons que la problématique est telle que les règles d'établissement d'un «*golden record* » sont formalisées dans la loi ou dans des règlements administratifs pour certaines sources authentiques, telles que le Registre National ou la Banque Carrefour des Entreprises belges, par exemple. Enfin, vu le nombre de *records* à comparer entre eux et d'opérations associées, des mécanismes de gestion de la performance («*blocking* » ou «*windowing* ») doivent être utilisés de manière itérative dans les opérations de matching d'envergure.

Isabelle Boydens⁴⁸ décrit l'opération de *back tracking* comme consistant à "déetecter, chez l'expéditeur et en partenariat avec celui-ci (...), les éléments à l'origine de la production d'un grand nombre de présomptions d'anomalies systématiques (traitement inadéquat de certaines sources de données, interprétation inadéquate de la législation, erreurs de programmation, etc.). Sur cette base, un diagnostic ainsi que des actions correctrices peuvent être posés (correction de code formel dans les programmes, adaptation de l'interprétation d'une loi, clarification de la documentation administrative ...)."

Nous avons donc vu (1) que les données *Big data* des offres d'emploi obtenaient

48. Isabelle BOYDENS : Data Quality & « back tracking » : depuis les premières expérimentations à la parution d'un Arrêté Royal | Smals Research, (visité le 09/03/2024).

difficilement un niveau satisfaisant de qualité de données, (2) qu'atteindre une qualité maximale comme pour la *Business Intelligence* demandait des investissements très importants, (3) surtout avec un traitement curatif des anomalies, moindre en traitement préventif. En conséquence, la qualité des données des offres d'emploi devrait être exigée dès leur encodage.

La représentativité des données

Pour la représentativité des données, nous allons aborder l'éthique des données et la mobilité des concepts.

Thilo Hagendorff⁴⁹ s'intéresse particulièrement à la qualité éthique des données utilisées par le ML pour le fonctionnement d'applications basées sur le comportement d'utilisation. Il considère les critères de qualité des données traditionnels comme liés aux objectifs business et décisionnels : avec le *Big data*, plus les données considérées sont nombreuses, moins le résultat de la requête sera précis. Les applications apprenantes nourries avec de grandes quantités de données produisent des résultats indifférents à certaines valeurs morales : n <>tout. Les méthodes avancées de ML permettent d'apprendre de petits jeux de données grâce à un processus d'augmentation de la quantité de données d'entraînement par la création de nouvelles données à partir des données existantes. La question est alors surtout de choisir quel sous-groupe ou fraction choisir, ou quelles nuances de données accentuer. Hagendorff ajoute à cette réflexion des critères éthiques de qualité des données :

- dimensions techniques des données : propreté, complétion, objectivité, consistance et fiabilité, datation, véracité, exactitude, coût d'acquisition, erreurs, qualité orthographique ;
- origine des données : auteurs, erreurs, sources de bruit, redondances, fréquences de mise à jour .

Selon lui, la qualité des données ne peut être évaluée que par des algorithmes d'apprentissage supervisés, qui exploitent des données qualifiées et étiquetées.

49. Thilo HAGENDORFF : Linking Human And Machine Behavior : A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning, in : *Minds & Machines* 31.4 (2021), Publisher : Springer Nature, p. 563-593, (visité le 05/03/2023).

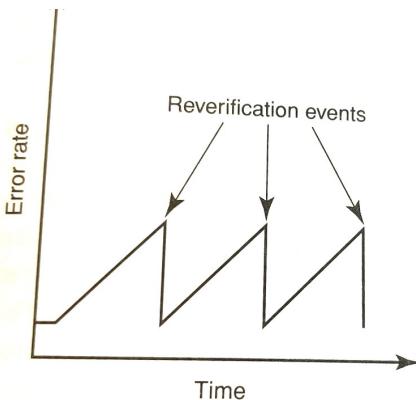


FIGURE 3.2 Accuracy of decayable elements over time.

FIGURE 9 – Précision des éléments dégradables dans le temps

Dans le cadre de nos offres d’emploi, l’auteur confirme ici que, selon lui, l’éthique de données ne peut être garantie que par des algorithmes d’apprentissage supervisés basés sur un corpus restreint, avec le traitement préventif de la qualité des données et de l’étiquetage.

Olson⁵⁰ rappelle dans son schéma *Accuracy of decayable elements over time* (Figure 9) qu’il est nécessaire de planifier des événements de revérification à partir d’un certain seuil de taux d’erreurs. Pour mieux prévoir l’évolution de cette gestion d’anomalies, Isabelle Boydens⁵¹ introduit la mobilité des concepts dans le temps :

D’un point de vue dynamique, une base de données idéale devrait (...) calquer le rythme de ses mises à jour sur la répartition – imprévisible – en « temporalités étagées » des évolutions de la réalité qu’elle appréhende. A ce qui ressemble à une gageure s’ajoute la nécessité, toujours révélée a posteriori, d’intégrer des observations imprévues, interdites a priori par l’hypothèse du monde clos.

Dans le cadre de notre question de recherche, nous pouvons donc conclure que les offres d’emploi et les candidatures doivent être revues régulièrement pour conserver leur *Fitness for use*.

Dans cette partie, nous avons vu que la pertinence et la représentation des don-

50. OLSON : Data Quality (cf. note 38).

51. Isabelle BOYDENS : Les bases de données sont-elles solubles dans le temps ?, in : La Recherche. Hors-série 9 (2002), p. 32-34, (visité le 09/03/2024).

nées collectées pouvaient être garanties par la définition d'une proportion acceptable d'anomalies. Ces anomalies peuvent porter sur des données manquantes, mal encodées, mal annotées, biaisées éthiquement, ou conceptuellement obsolètes. Pour les offres d'emploi et les candidatures, nous pouvons en conclure que le ML est un traitement automatique de données reconnu pour identifier le point de bascule à partir duquel la proportion acceptable d'anomalies est dépassé.

2.3.3 Le défi de la transformation des données et l'opacité des choix des algorithmes

Nous allons ici faire le tour des différents constats quant à l'opacité des choix des IA.

Côté parties prenantes, Stephen C. Slota et al.⁵² ont mené une étude basée sur des entretiens avec des personnes engagées dans le développement technologique, la politique et le droit relatifs à l'IA. Ils en concluent que le manque de contrôle sur les résultats ne concerne pas seulement les données, les algorithmes ou la mise en œuvre, mais nécessite une vision complète du processus pour mieux comprendre les conséquences éthiques et sociales d'une IA.

Ma participation à l'*"IA Week"* des 27 au 31 mars 2023 à Bruxelles m'a permis de comprendre que la législation et les normes européennes étaient en cours de développement, que les organismes développant de l'IA créaient leurs propres indicateurs de qualité, et qu'il n'était pour l'instant pas vraiment possible de corriger les biais des algorithmes; la seule correction peut venir des humains eux-mêmes. Plusieurs programmes s'ouvrent d'ailleurs au niveau européen pour aider la population et les entreprises à mieux comprendre les biais inhérents de l'IA⁵³.

Côté technique, Tolga Bolukbasi et al.⁵⁴ relèvent que les plongements lexicaux (*Word embeddings*) amplifient les biais des données sources. Il est néanmoins possible, via

52. Stephen C. SLOTA et al. : Good systems, bad data? : Interpretations of AI hype and failures, in : Proceedings of the Association for Information Science and Technology 57.1 (2020), e275, (visité le 05/09/2023).

53. Marina AUBERT : AI Week Belgium 2023. Marina Aubert, 31 mars 2023, (visité le 28/04/2024).

54. Tolga BOLUKBASI et al. : Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, in : prépublication Cornell University 2016, (visité le 08/08/2023).

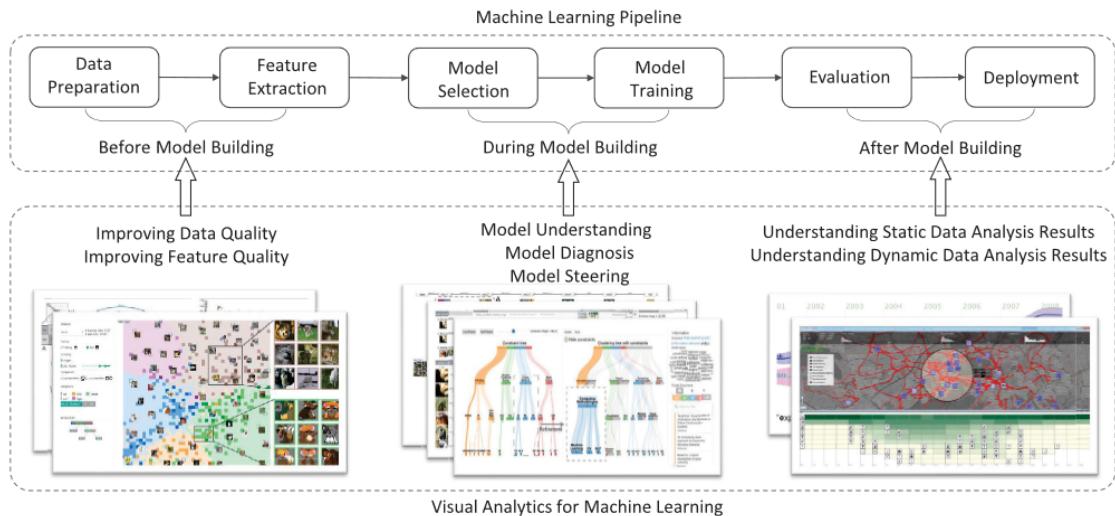


Fig. 1 An overview of visual analytics research for machine learning.

FIGURE 10 – Visualisations du pipeline de ML

une méthodologie spécifique, d'éliminer les stéréotypes de genre, "tels que l'association entre les mots réceptionniste et femme, tout en maintenant les associations souhaitées, telles que les mots reine et femme". Il est donc possible de définir des mesures "pour quantifier les biais de genre directs et indirects dans les plongements, et de développer des algorithmes pour les "débiaiser". Par ailleurs, Rishabh Bhardwaj et al.⁵⁵ ont d'ailleurs constaté que l'omission des composantes du vecteur de mots inappropriés (i.e. réceptionniste et femme) avec le modèle de langage contextuel BERT s'est avérée réduire les préjugés sexistes dans les tâches en aval. La méthode peut être adaptée à l'étude d'autres biais sociaux tels que la race et l'ethnicité.

Yuan, Jun et al⁵⁶ ont réalisé une synthèse des techniques élaborées en 10 ans sur la visualisation des processus de Machine Learning, afin d'en améliorer la transparence (Figure 10, Figure 11).

Dans *AI Failures : A Review of Underlying Issues*⁵⁷, Debarag Narayan Banerjee et Sasanka Sekhar Chandadans ont établi une liste des différents types d'erreurs générées

55. Rishabh BHARDWAJ/Navonil MAJUMDER/Soujanya PORIA : Investigating Gender Bias in BERT, in : prépublication Cornell University 2020, (visité le 22/07/2023).

56. Yuan JUN et al. : A survey of visual analytics techniques for machine learning, in : Computational Visual Media 7.1 (2021), (visité le 05/03/2023).

57. Debarag Narayan BANERJEE/Sasanka Sekhar CHANDA : AI Failures : A Review of Underlying Issues, in : prépublication Cornell University 2020, URL : <http://arxiv.org/abs/2008.04073> (visité le 05/09/2023).

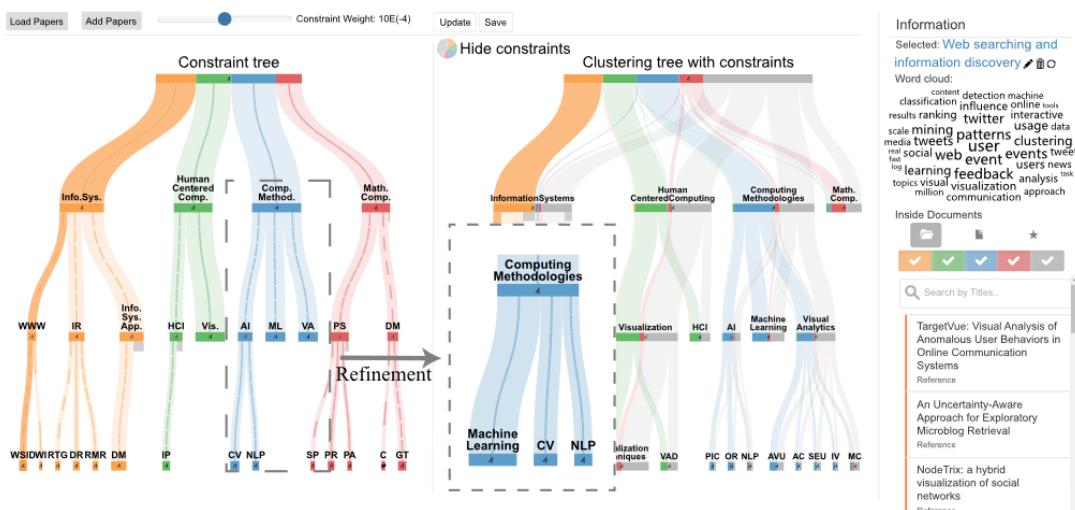


Fig. 6 ReVision, a visual analytics system integrating a constrained hierarchical clustering algorithm with an uncertainty-aware, tree-based visualization to help users interactively refine hierarchical topic modeling results. Reproduced with permission from Ref. [125], © IEEE 2020.

FIGURE 11 – Outil de visualisation ReVision

par les systèmes d’IA : l’oubli d’étapes pour un comportement modélisé, une action inappropriée face à un événement de son environnement, la mauvaise interprétation d’une donnée, le problème généré par le manque de robustesse de son matériel, et notamment de ses capteurs, ainsi que l’incapacité à porter un jugement moral. La publication propose des pistes de solutions à chaque type d’erreur, tout en évoquant le large champ de recherche de l’*AI Safety*.

Même si les résultats de certaines IA peuvent être corrigés de manière empirique, l’opacité des choix des algorithmes ouvre un champ entier de recherche, l’*AI Safety*.

Dans le cadre des offres d’emploi, nous pourrons ajouter au principe d’Interskillar et à l’apprentissage supervisé de la visualisation de contrôle, pour vérifier à chaque étape quel scénario répond le mieux au *Fitness for use*.

2.4 Biais de l’utilité de l’IA

En 2022, Gianluca Bontempi⁵⁸, professeur à l’ULB, rappelait :

Comme l’a souligné Bernoulli en 1738, "aucune mesure valable de la valeur du risque ne peut être donnée sans tenir compte de son utilité".

58. Gianluca BONTEMPI : The AI gap : from good accuracy to bad decisions. The regularity gamble, 2022, URL : <https://datascienceth741.wordpress.com/> (visité le 05/03/2023).

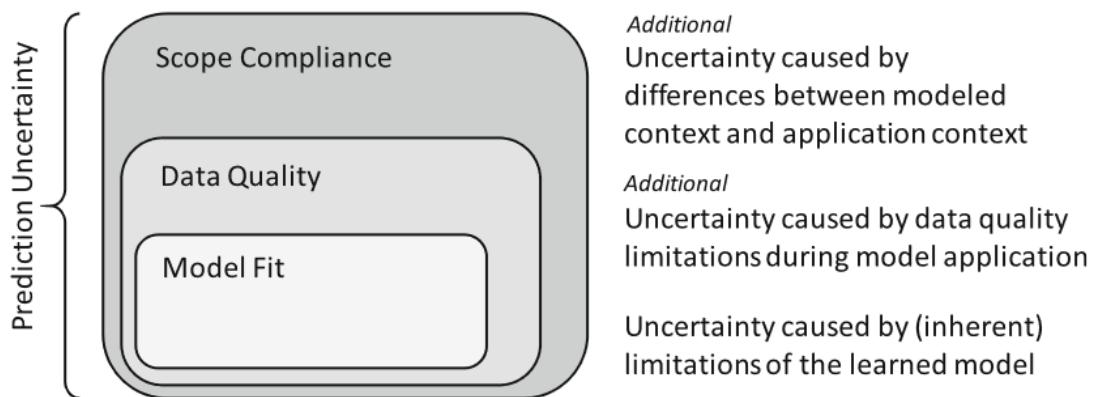


Fig. 1. Onion layer model of uncertainty in AI/ML application outcomes.

FIGURE 12 – Incertitudes introduites par le ML

Pour sa thèse de doctorat en philosophie⁵⁹, Haihua Chen relève même que "améliorer les données pourrait être plus efficace qu'améliorer le modèle".

Dans leur article "*Uncertainty in Machine Learning Applications : A Practice-Driven Classification of Uncertainty*"⁶⁰, Michael Kläs et Anna Maria Vollmer résument les incertitudes de prédictions de l'AI/ML dans un schéma (Figure 12).

Aux incertitudes intrinsèques du modèle, sont ajoutées les incertitudes liées aux limites de la qualité des données durant l'application du modèle, et les incertitudes causées par les différences entre le contexte modélisé et le contexte d'application. La création de modèles d'évaluation des incertitudes est actuellement un champ de recherche en soi⁶¹.

Utiliser l'IA pour débiaiser des offres d'emploi comporte le risque d'introduire de nouveaux biais : on entre alors dans un cycle sans fin, dont la seule issue est d'établir un ou des seuils à partir desquels débiaiser par l'IA sortirait du *Fitness for use*⁶².

59. Haihua CHEN/Jiangping CHEN/Junhua DING : Data Evaluation and Enhancement for Quality Improvement of Machine Learning, in : IEEE Transactions on Reliability 70.2 (2021), Conference Name : IEEE Transactions on Reliability, p. 831-847.

60. Michael KLÄS/Anna Maria VOLLMER : Uncertainty in Machine Learning Applications : A Practice-Driven Classification of Uncertainty, in : Barbara GALLINA et al. (éd.) : Computer Safety, Reliability, and Security (Lecture Notes in Computer Science), Cham 2018, p. 431-438.

61. Lisa JÖCKEL et al. : Conformal Prediction and Uncertainty Wrapper : What Statistical Guarantees Can You Get for Uncertainty Quantification in Machine Learning?, in : Jérémie GUIOCHEZ et al. (éd.) : Computer Safety, Reliability, and Security. SAFECOMP 2023 Workshops, Cham 2023, p. 314-327.

62. Harald FOIDL/Michael FELDERER : Risk-based data validation in machine learning-based software systems, in : 27 août 2019, p. 13-18.

2.5 Biais des sources

Les sources mêmes de ce mémoire et les outils utilisés pour le rédiger pourraient comporter des biais, que nous relevons ici, mais que nous ne nous attacherons pas à analyser.

La bibliographie de ce mémoire comporte des sources scientifiques et non-scientifiques. Elles représentent un certain biais de représentation :

- Biais de représentation de genre : 16 autrices pour 33 auteurs ; à cause de l'effet Matilda⁶³, nous avons considéré qu'une source non signée était rédigée par une femme.
- Biais de représentation géographique : 19 sources provenant d'Amérique du Nord, 3 Asie, 28 Europe, 1 Australie ; aucune d'Afrique ni d'Amérique du Sud.

Par ailleurs, le fait que seules des sources en français et en anglais aient été consultées est un biais en soit.

Les traductions vers le français ont été réalisées avec l'aide de DeepL. Outre les sources déjà citées, le développement du prototype a reçu le concours de mon directeur de mémoire, mais aussi de l'aide des forums en ligne, de blogs et de ChatGPT.

2.6 Discussion

Les biais envisagés pour l'étude de cas devront tenir compte à la fois des ressources disponibles (données sources, temps et équipement informatique), mais aussi éviter de reproduire des outils déjà existants.

2.6.1 Détection des biais des offres d'emploi

Pour détecter les biais genrés, il existe des outils en ligne, Proprec, Textio et Ongig, qui offrent un service de “dégenrification” des offres d'emploi .

63. Silvia KNOBLOCH-WESTERWICK/Carroll J. GLYNN/Michael HUGE : The Matilda Effect in Science Communication : An Experiment on Gender Bias in Publication Quality Perceptions and Collaboration Interest, in : Science Communication 35.5 (2013), Publisher : SAGE Publications Inc, p. 603-625, (visité le 14/05/2023).

This advert is masculine-coded

This job ad uses more words that are subtly coded as feminine than words that are subtly coded as masculine (according to the research). Fortunately, the research suggests this will have only a slight effect on how appealing the job is to men, and will encourage women applicants.

Of course, there are plenty of other factors that affect the diversity of applicants for this role, and of the people who end up being hired. These include the company's reputation for inclusiveness, its culture, and the behaviour and prejudices (both conscious and unconscious) of the interviewers.

Masculine-coded words in this ad

- analyser
- analysez
- analyses
- autonome
- confidentialité

[See the full list of masculine-coded words](#)

Feminine-coded words in this ad

- aménager
- gardiennage
- (responsables)

[See the full list of feminine-coded words](#)

FIGURE 13 – Propec

- Proprec (Figure 13) est basé sur l'article de Danielle Gaucher et Justin Friesen de 2011⁶⁴. Il arrive à analyser une offre d'emploi en français : par exemple, ici issue du site web d'Actiris.

NB : la capture d'écran a été réalisée le 18/02/2024 ; le service a depuis disparu.

- Les sites web Textio (Figure 14) et Onging (Figure 15) proposent un service similaire payant et uniquement en anglais. Il ne permettent pas de tester l'application en ligne. Voici des captures d'écran issues du site web.

64. GAUCHER/FRIESEN/KAY : Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality (cf. note 25).

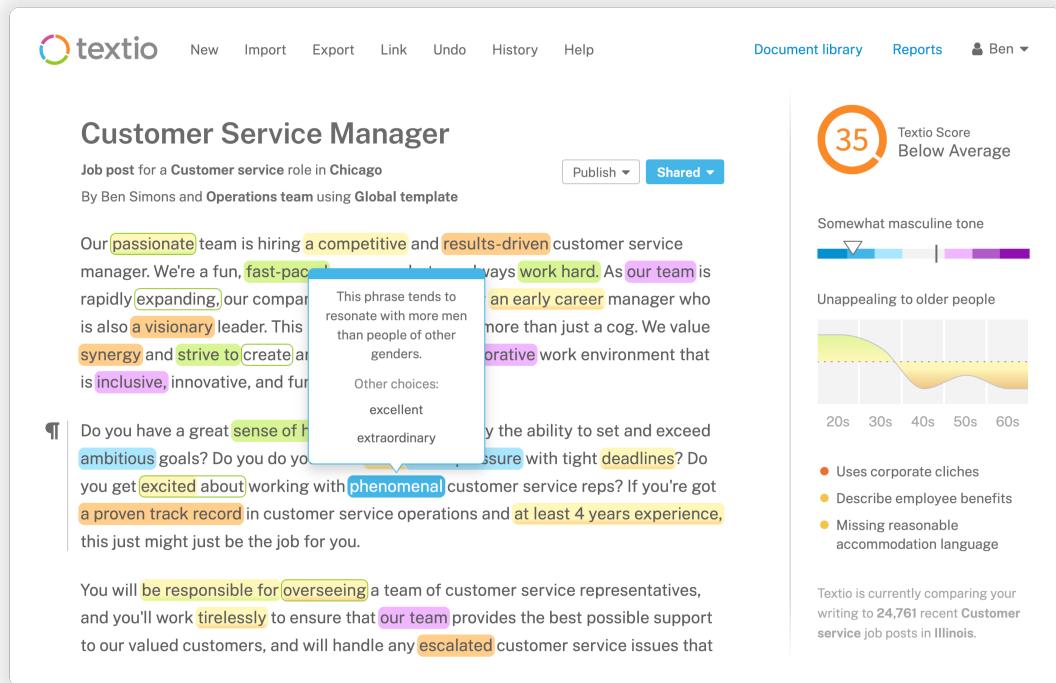


FIGURE 14 – Textio

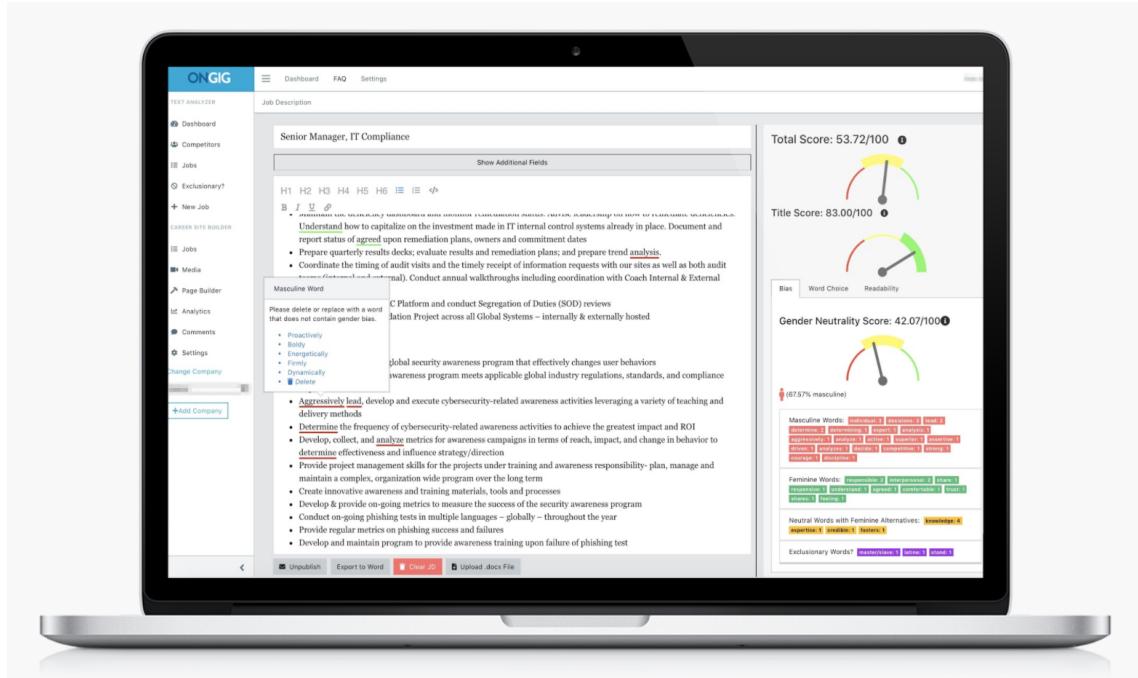


FIGURE 15 – Ongig

Pour détecter les biais de niveau de vocabulaire, la plateforme DeepFLE (Figure 16) évalue depuis 2023 le niveau d’accessibilité du français en utilisant le Cadre euro-

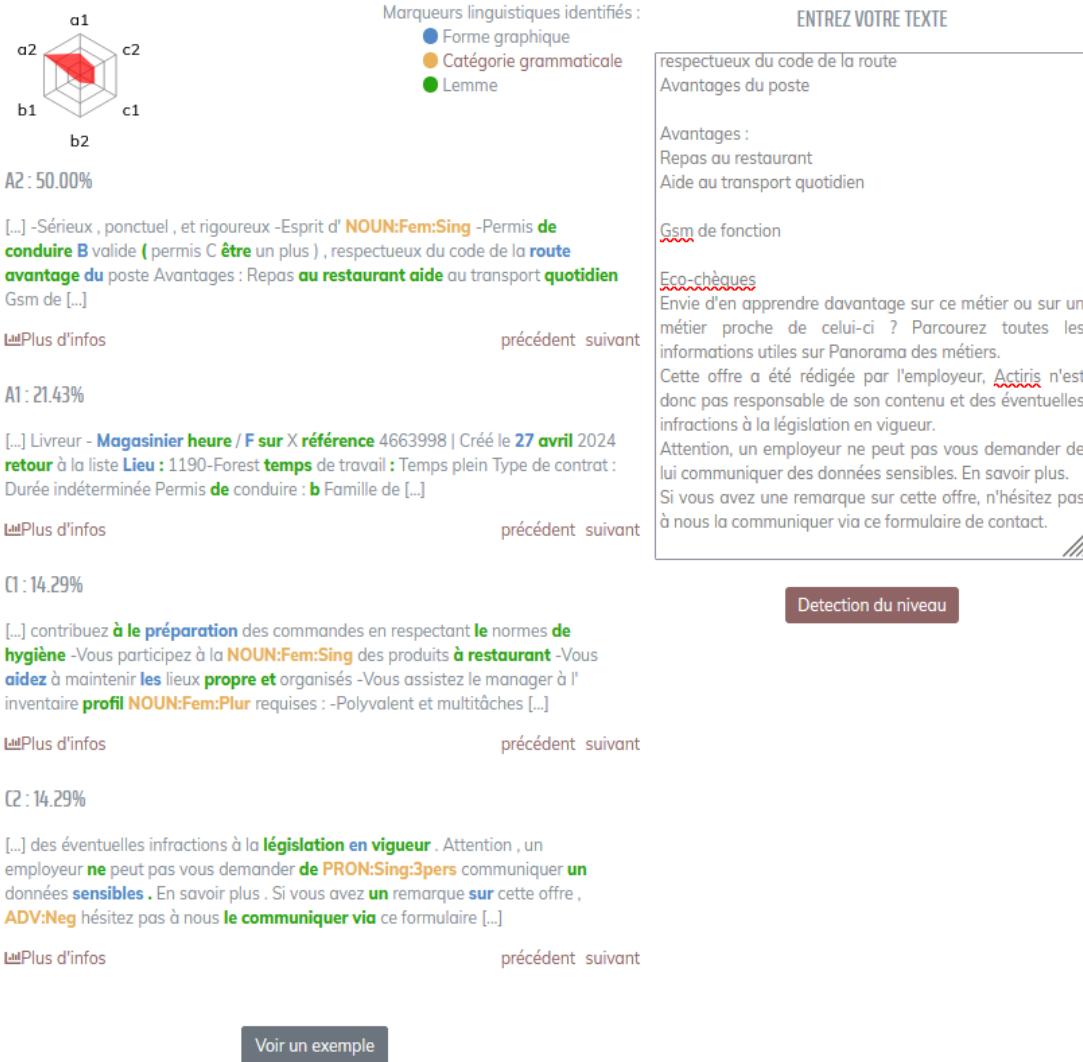


FIGURE 16 – DeepFLE

péen commun de référence pour les langues du Conseil de l'Europe : de A1 pour les débutants, à C2 pour les utilisateurs avancés⁶⁵. Ici une analyse DeepFLE d'une offre d'emploi en français issue du site web d'Actiris

Pour détecter les biais d'annotation, Nicole M. Brown et al⁶⁶ ont pu combiner des méthodes statistiques (topic modeling, Divergence Kullback-Leibler et similarité cosinus), l'intermediate reading (par tableau de données, Figure 17), et le distant reading

65. Simona RUGGIA/Laurent VANNI : DeepFLE : la plateforme pour évaluer le niveau d'un texte selon le CECRL, in : Dialogues et cultures, Dialogues et cultures 2021, Publisher : Fédération internationale des professeurs de français, p. 235-254, (visité le 18/02/2024).

66. Nicole M. BROWN et al. : In Search of Zora/When Metadata Isn't Enough : Rescuing the Experiences of Black Women Through Statistical Modeling, in : ISSN : 1938-6389, (visité le 28/10/2023).

Table 1. Key word list.

Word	Number of times word assigned to topic
Slave	163,897
Master	26,152
Free	24,126
Slaveri	22,843
Negro	11,207
Freedom	9,812
Owner	8,016
Sold	7,750
State	7,632
Property	6,623
Law	4,784
Sell	4,548
Sale	3,669
Children	3,483
White	3,159
Condition	3,101
Liberty	3,054
Persons	2,984
Purchas	2,907
Hold	2,592
Emancipation	2,358
Made	2,356
Labor	2,342

FIGURE 17 – Intermediate reading

(par treemap, Figures 18 et 19) pour ajouter des annotations manquantes ou biaisées à partir d'annotations indirectes sur la présence des noires américaines dans les archives.



FIGURE 18 – Distant reading : annotations



FIGURE 19 – Distant reading : annotations indirectes

2.6.2 Détection des biais du traitement automatique des données

Pour évaluer la qualité des données, nous pourrons utiliser la matrice de faux positifs / faux négatifs.

Pour identifier les anomalies, nous devrons emprunter les méthodes de traitement correctif de la qualité des données, avec les *data quality tools*: profiling, standardisation, matching; la méthode du back tracking est hors de portée d'un mémoire de master. Comme standard de métadonnées, le référentiel des compétences professionnelles Competent a été développé par le VDAB à partir du référentiel français ROMEv3. Il est actuellement en train de devenir central chez Actiris et au Forem. Les institutions belges ont écarté le standard européen ESCO.

Pour éviter les biais éthiques, nous pourrons utiliser un algorithme d'apprentissage supervisé.

Pour pallier à la mobilité des concepts, nous devrons planifier un processus de réévaluation périodique des contraintes d'intégrité, des règles métier, et des concepts. Dans le cadre de ce mémoire, cela signifie qu'une révision des critères d'acceptabilité du moteur de règles du ML devra être intégré au processus d'amélioration continue du ML.

Pour contrôler les biais de la transformation des données et de l'opacité des choix des algorithmes, nous pourrons générer des graphiques de visualisation.

2.6.3 Détection des autres biais

Nous considérons que nous pourrons contrer les biais de l'utilité de l'IA et des sources de ce mémoire uniquement en tenant compte de ces paramètres dans l'interprétation finale des résultats de ce mémoire.

2.6.4 Synthèse de la discussion

Afin d'obtenir une vue d'ensemble des différents biais, nous proposons de réaliser un tableau de synthèse des outils évoqués (Table 2). Grâce à ce tableau de synthèse,

Types de biais	Outils disponibles
Genre	plateformes payantes
Niveau de vocabulaire	DeepFLE
Annotations	Visualisation
Pertinence des données	Matrice de faux positifs / faux négatifs
Anomalies	Data quality tools, Competent
Ethique	Algorithme d'apprentissage supervisé
Mobilité des concepts	Révision périodique du moteur de règles ML
Transformation des données et opacité des choix des algorithmes	Visualisation
Utilité de l'IA et sources	Vérification lors de l'interprétation finale

TABLE 2 – Récapitulatif des types de biais et des outils disponibles

nous pouvons en conclure que l'étude de cas répondra à la question de recherche en identifiant les biais de genre via (1) la vérification et correction de la qualité des données (évaluation de la pertinence des données via la matrice de faux positifs/faux négatifs, avec les *data quality tools*, le référentiel Competent), en utilisant (2) un algorithme d'apprentissage supervisé, et en (3) vérifiant ses résultats prédictifs via la visualisation. (4) Les résultats devront ensuite être discutés à la lumière des biais de l'IA et des sources.

2.7 Conclusion de l'état de l'art

Dans ce chapitre, nous avons réalisé un tour exhaustif de l'état des connaissances sur notre question de recherche au moment de la rédaction du mémoire (c'est une matière vivante qui évolue rapidement!). Nous avons d'abord défini les concepts des biais, des biais des données, de la qualité des données et du traitement automatique des données par IA. Nous avons ensuite énuméré les impacts des biais dans les offres d'emploi au niveau des biais genrés, de niveau de vocabulaire et d'annotation. Puis nous avons balayé largement tous les types de biais possiblement générés par le traitement automatique des données. Ces types de biais étaient liés à plusieurs défis : le défi de la qualité des données, le défi de la pertinence et de la représentation des données, et le défi de la transformation des données et de l'opacité des choix des algorithmes. Par souci d'exhaustivité et de précision, nous avons également évoqué les biais de l'uti-

lité de l'IA et les biais des sources de ce mémoire. Enfin, nous avons développé une discussion sur les biais envisagés pour l'étude de cas, et réalisé une synthèse du cadre auquel elle devra répondre. Et justement, la voilà.

3 Etude de cas

Comme formalisé en conclusion de l'état de l'art, l'objectif de cette étude de cas est d'identifier les biais de genre des offres d'emploi via (1) la vérification et correction de la qualité des données (via la matrice de faux positifs/faux négatifs, avec les *data quality tools*, le référentiel Competent), en utilisant (2) un algorithme d'apprentissage supervisé de ML, et en (3) vérifiant ses résultats prédictifs via la visualisation. (4) Nous pourrons ensuite discuter des résultats obtenus à la lumière des biais induits par l'utilité de l'IA et par les sources.

3.1 Méthodologie

Comme déjà mentionné, nous reprenons pour ce mémoire le postulat d'Interskillar : les biais d'une offre d'emploi se mesurent à la diversité des profils qui y répondent, plus l'offre d'emploi est biaisée, plus les profils des candidatures sont homogènes. Les critères discriminants utilisés par Interskillar sont : les compétences qui ont permis la connexion entre l'offre et le profil, l'âge, le niveau de formation, la première et la deuxième nationalités, le sexe, le domaine d'études, l'école, le statut marital, le statut parental, le code postal, le handicap. Pour notre étude de cas, nous avons donc besoin d'accéder à un jeu de données d'offres d'emploi et à un jeu de données correspondantes de candidatures à ces offres d'emploi.

3.1.1 Collecte des données

Mes contacts professionnels m'ont permis de demander à Actiris et au Forem s'ils étaient intéressés pour participer à mon test de biais, le préalable à cette participation étant de me fournir une quantité de données suffisantes pour le ML, et des profils anonymisés. Les deux institutions travaillant avec des consultants externes, elles n'ont pas souhaité utiliser leur budget à l'anonymisation de profils pour un travail d'étudiante.

Je me suis alors tournée vers de grandes entreprises privées, intéressées par le sujet. Après avoir discuté avec des spécialistes du recrutement chez Proximus, j'ai réa-

lisé que ces critères discriminants étaient évidemment protégés au niveau européen par le Règlement Général de Protection des Données (RGPD)⁶⁷, et ne pouvaient pas être collectés de manière structurée ; ces données sont en effet accessibles de manière structurée uniquement aux institutions publiques qui sont en charge du paiement des allocations sociales.

L'autorité de protection des données précise :

Une protection accrue est prévue pour l'utilisation et le traitement des données à caractère personnel sensibles suivantes :

- les catégories particulières de données à caractère personnel telles que mentionnées à l'article 9.1 du RGPD, plus particulièrement :
 - l'origine raciale ou ethnique;
 - les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale;
 - les données génétiques (par exemple une analyse ADN) (voir également le considérant 34 du RGPD);
 - les données biométriques à des fins d'identification unique (par exemple des données d'empreintes digitales ou la reconnaissance faciale ou de l'iris);
 - les données concernant la santé (voir également le considérant 35 du RGPD);
 - les données relatives à la vie sexuelle ou à l'orientation sexuelle.
 - les données à caractère personnel relatives aux condamnations pénales et aux infractions (article 10 du RGPD)

Il serait techniquement possible de traiter automatiquement les contenus des cv de candidature (*parsing*) pour y déceler les critères discriminants, mais mes recherches ont déjà montré que la prédiction par intelligence artificielle de l'analyse des cv est un champ de recherche en soi, qui n'a pas encore donné de résultats probants⁶⁸. Nous

67. Données sensibles | Autorité de protection des données, (visité le 23/03/2024).

68. Marina AUBERT : Les applications XML dans les systèmes de publication des offres d'emploi, ULB, 2021.

excluons donc cette option de cette étude de cas.

Comment contourner cet obstacle ?

Je me suis alors souvenu que le processus d'inscription aux formations de Bruxelles Formation était similaire à celui d'un recrutement, et reprenait un certain nombre des critères discriminants proposés par Interskillar. Après vérification, il s'est avéré que, par le biais des financements du Fonds Social Européen, Bruxelles Formation avait des obligations de mesurer, et donc de collecter les critères discriminants des personnes candidates à ses formations.

Pour suivre une formation à Bruxelles Formation, les personnes en recherche d'emploi doivent d'abord s'inscrire à une séance d'information, en ligne ou via un organisme prescripteur. A l'issue de cette séance d'information, elles doivent passer un test de base (français, calcul) et un entretien. Si ces résultats sont insuffisants pour participer à la formation, une solution intermédiaire de remise à niveau est proposée, ou une solution alternative comme un processus de validation de compétences. Si ces résultats sont suffisants pour participer à la formation, elle devient stagiaire en formation.

Nous pouvons donc considérer comme source alternative à des offres d'emploi les fiches formation de la base de données Dorifor des formations professionnelles en Région bruxelloise, et plus spécifiquement les fiches formation de Bruxelles Formation, ainsi que les profils des personnes qui y ont candidaté. Ces données ont l'avantage d'être uniquement en français, et destinées aux personnes en recherche d'emploi. On peut donc estimer par glissement, que les types de biais des fiches formation sont similaires à ceux des offres d'emploi.

Bruxelles Formation a confirmé son intérêt et son autorisation pour que je réalise mon étude de cas sur ces données pseudonymisées.

La CNIL⁶⁹ précise :

L'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.

69. L'anonymisation de données personnelles, URL : <https://www.cnil.fr> (visité le 23/03/2024).

sible.

L'anonymisation ne doit pas être confondue avec la pseudonymisation. La pseudonymisation est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire.

En pratique, la pseudonymisation consiste à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.). La pseudonymisation permet ainsi de traiter les données d'individus sans pouvoir identifier ceux-ci de façon directe. En pratique, il est toutefois bien souvent possible de retrouver l'identité de ceux-ci grâce à des données tierces : les données concernées conservent donc un caractère personnel. L'opération de pseudonymisation est également réversible, contrairement à l'anonymisation. La pseudonymisation constitue une des mesures recommandées par le RGPD pour limiter les risques liés au traitement de données personnelles.

Pseudonymisation des données de Bruxelles Formation

Avec Bruxelles Formation, il a d'abord fallu évaluer quelles étaient les données disponibles (existantes) et accessibles (RGPD), avec les services juridiques et Etudes et développement. Puis une convention de manipulation des données confiées a été établie (engagement moral).

Avec la data analyst de Bruxelles Formation, Djouaria Ghilani, nous avons ensuite identifié les critères de données adaptés aux besoins de ce mémoire. Enfin, nous avons déterminé les modes de pseudonymisation de chaque type de données sensibles.

Les candidatures retenues ont été encodées entre le 1^{er} janvier 2022 et le 31 mai 2023 inclus. Elles sont nominatives : le champ « nom » n'était pas vide, pour exclure les « annulations techniques » (lorsqu'un formulaire d'inscription est ouvert mais non-soumis). Elles ont été encodées suite à des séances d'information liées à une seule fiche formation à la fois (il existe des séances d'information transversales, qui présentent

plusieurs formations lors de la même séance). Toutefois, une même fiche formation peut être liée à plusieurs séances d'information, qui ont eu lieu à des moments différents. Les inscriptions à ces différentes séances d'information ont été conservées. Elles sont liées à un numéro de registre national à 11 caractères (pour permettre une cohérence lors de l'extraction du sexe et l'année de naissance à partir du registre national). Ces candidatures sont associées à une fiche formation pour laquelle étaient inscrits au moins 50 candidates ou candidats distincts, pour rendre plus difficile la possibilité d'identifier des personnes spécifiques.

L'extraction finale concerne donc 15 777 candidatures, réalisées par 10 522 personnes distinctes, ayant participé à 893 séances d'information distinctes, associées à 84 fiches formation distinctes. A noter que ces candidatures sont incluses, quelle que soit leur issue : elles peuvent avoir été annulées ou non par la suite, la personne peut s'être présentée ou non en séance d'information, peut avoir introduit ou non une demande de formation effective, qui peut avoir débouché ou non sur une entrée réelle en formation. Les éléments du corpus sont édités et disponibles dans un format textuel et sont prêts à l'emploi.

En plus des temps de réunion, ce processus de pseudonymisation des données a coûté environ une journée de travail à la data analyst de Bruxelles Formation, Djouaria Ghilani. Le coût de la pseudonymisation évoqués par Actiris et le Forem est donc réel.

Pour suivre les conclusions de l'état de l'art, nous retiendrons uniquement les données des offres de formation et la donnée sexe des candidatures.

3.1.2 Vérification et correction de la qualité des données

Avant de procéder à l'analyse de la qualité des données reçues, nous devons procéder à leur exploration, et vérifier qu'elles correspondent au *Fitness for use* de cet étude de cas. Si ce n'est pas le cas, il faudra procéder à leur adaptation. Ensuite, nous pourrions reprendre nos conclusions de l'état de l'art, et évaluer leur pertinence et leur qualité (via le *data profiling*), puis procéder à la correction des éventuelles anomalies (via la standardisation et le *data matching*). Il ne sera pas possible d'évaluer la qualité des

méta données via le référentiel Competent, car ces méta données n'existent pas (encore) dans Dorifor.

3.1.3 Algorithme d'apprentissage supervisé

Comme décidé à l'issue de l'état de l'art, nous allons donc utiliser un algorithme d'apprentissage supervisé.

Langage de programmation

Dans le cadre des cours du master STIC, nous avons eu le choix entre développer l'algorithme de ML en Python ou en R. Dans *Comparing programming languages for data analytics : Accuracy of estimation in Python and R*⁷⁰, Chelsey Hill et al. ont réalisé un test comparatif des deux langages sur des tâches similaires. Leur recommandation est d'utiliser R pour les analyses statistiques, et Python pour les analyses non-statistiques.

Dans cette étude de cas, nous voulons prédire si une fiche formation présente des biais genrés. Il s'agit donc d'une probabilité statistique : nous utiliserons donc R.

Méthodologie de ML

Dans *Hands-On Machine Learning with R*⁷¹, Bradley Boehmke et Brandon Greenwell précisent que lorsque l'objectif de l'apprentissage supervisé est de prédire un résultat catégorique, il faut parler d'un problème de classification. Steven Bird, Ewan Klein, et Edward Loper ajoutent dans *Natural Language Processing with Python*⁷² que la classification consiste à choisir la bonne étiquette de classe pour une entrée donnée. "Un classificateur est dit supervisé s'il est construit sur la base de corpus d'apprentissage contenant l'étiquette correcte pour chaque entrée." Ils illustrent dans un schéma (Figure 20) les deux phases nécessaires à l'apprentissage supervisé : la phase d'entraînement du modèle ML, et la phase de prédiction à l'issue de laquelle le modèle est capable de prédire l'étiquette à associer à l'entrée donnée.

70. Chelsey HILL et al. : Comparing programming languages for data analytics : Accuracy of estimation in Python and R, in : WIRES Data Mining and Knowledge Discovery, e1531, (visité le 30/03/2024).

71. Bradley BOEHMKE/Brandon GREENWELL : Hands-On Machine Learning with R, 2020, (visité le 23/04/2023).

72. Steven BIRD/Ewan KLEIN/Edward LOPER : Natural Language Processing with Python, 1^{er} jan. 2009.

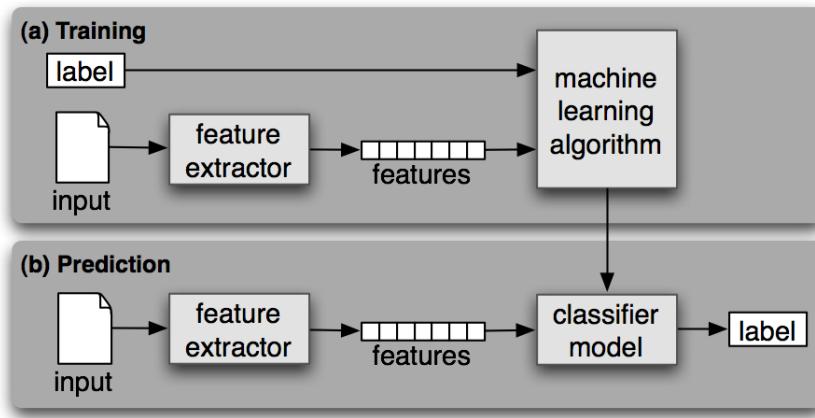


FIGURE 20 – Phases nécessaires à l'apprentissage supervisé

Dans "Hands-On Machine Learning with R"⁷³, Bradley Boehmke et Brandon Greenwell précisent :

Pour bien comprendre la généralisabilité de notre modèle optimal final, nous pouvons diviser nos données en ensembles de données d'entraînement et de données de test : - Ensemble d'entraînement : ces données sont utilisées pour développer des ensembles de caractéristiques, entraîner nos algorithmes, régler les hyperparamètres, comparer les modèles et toutes les autres activités nécessaires pour choisir un modèle final (par exemple, le modèle que nous voulons mettre en production). - Ensemble de test : après avoir choisi un modèle final, ces données sont utilisées pour estimer une évaluation impartiale de la performance du modèle, que nous appelons l'erreur de généralisation.

Dans son exposé au FNRS⁷⁴, Laurence Dierickx avait présenté les différentes options de ML vue par l'entreprise d'IA Techtarget⁷⁵ (Figure 21).

Nous observons que les arbres de décision sont recommandées pour les analyses prédictives. Gopinath Rebala et al. précisent dans *An Introduction to Machine Lear-*

73. BOEHMKE/GREENWELL : Hands-On Machine Learning with R (cf. note 71).

74. DIERICKX : Apprentissage automatique : les challenges de la qualité des données dans la perspective d'une adéquation aux usages (cf. note 5).

75. George LAWTON : What are Machine Learning Models? Types and Examples. Techtarget, (visité le 28/04/2024).

Machine learning models cheat sheet

Supervised learning	Unsupervised learning	Semi-supervised learning	Reinforcement learning
<p>Data scientists provide input, output and feedback to build model (as the definition)</p> <p>EXAMPLE ALGORITHMS:</p> <ul style="list-style-type: none"> Linear regressions <ul style="list-style-type: none"> ▪ sales forecasting ▪ risk assessment Support vector machines <ul style="list-style-type: none"> ▪ image classification ▪ financial performance comparison Decision tree <ul style="list-style-type: none"> ▪ predictive analytics ▪ pricing 	<p>Use deep learning to arrive at conclusions and patterns through unlabeled training data.</p> <p>EXAMPLE ALGORITHMS:</p> <ul style="list-style-type: none"> Apriori <ul style="list-style-type: none"> ▪ sales functions ▪ word associations ▪ searcher K-means clustering <ul style="list-style-type: none"> ▪ performance monitoring ▪ searcher intent 	<p>Builds a model through a mix of labeled and unlabeled data, a set of categories, suggestions and exampled labels.</p> <p>EXAMPLE ALGORITHMS:</p> <ul style="list-style-type: none"> Generative adversarial networks <ul style="list-style-type: none"> ▪ audio and video manipulation ▪ data creation Self-trained Naïve Bayes classifier <ul style="list-style-type: none"> ▪ natural language processing 	<p>Self-interpreting but based on a system of rewards and punishments learned through trial and error, seeking maximum reward.</p> <p>EXAMPLE ALGORITHMS:</p> <ul style="list-style-type: none"> Q-learning <ul style="list-style-type: none"> ▪ policy creation ▪ consumption reduction Model-based value estimation <ul style="list-style-type: none"> ▪ linear tasks ▪ estimating parameters

©2020 TECHTARGET. ALL RIGHTS RESERVED TechTarget

FIGURE 21 – Comparatif des différents types d'IA

ning⁷⁶ :

Un arbre de décision est un modèle d'apprentissage automatique construit à l'aide d'une série de décisions basées sur des valeurs variables afin d'empêtrer une voie ou une autre. Une *Random forest* est un ensemble d'arbres décisionnels qui améliorent la prédiction par rapport à un seul arbre décisionnel.

Chen et al. ont réalisé une synthèse des méthodes de ML utilisées dans leur article *Data evaluation and enhancement for quality improvement of Machine Learning*⁷⁷. Ils indiquent que les Random forests ont obtenu les meilleures performances [parmi les méthodes d'évaluation de la qualité des données]. [Elles] sont également très performantes en matière de classification de textes car elles atténuent les difficultés inhérentes aux données textuelles, telles que la haute dimensionnalité, la rareté et l'espace de caractéristiques bruité.

L'utilisation de la *Randomforest* s'impose donc comme modèle de classification de

76. Gopinath REBALA/Ajay RAVI/Sanjay CHURIWALA : An Introduction to Machine Learning, Cham 2019, (visité le 05/03/2023).

77. CHEN/CHEN/DING : Data Evaluation and Enhancement for Quality Improvement of Machine Learning (cf. note 59).

l'algorithme d'apprentissage supervisé.

Traitements automatiques des données

En suivant la méthode proposée par Boemke et Greenwell⁷⁸, nous procéderons d'abord à l'import, au nettoyage et au paramétrage des données, puis nous distribuerons le jeu de données entre un jeu d'apprentissage (70% des données) et un jeu de test de la prédiction (30% des données), enfin nous construirons et appliquerons aux données notre modèle de classification *Random forest*.

Nous écartons pour l'instant la lemmatisation, que nous considérons comme un des scénarios d'optimisation du prototype ML. Cela permet de conserver un code léger pour cette première version du prototype.

Lemmatisation : Opération consistant à regrouper les formes occurrentes d'un texte ou d'une liste sous des adresses lexicales⁷⁹

Afin d'augmenter la précision des prédictions, Boemke et Greenwell⁸⁰ conseillent également d'ajouter de la validation croisée au jeu d'entraînement :

Les méthodes de rééchantillonnage offrent une approche alternative en nous permettant d'ajuster de manière répétée un modèle intéressant à certaines parties des données d'apprentissage et de tester ses performances sur d'autres parties. La validation croisée k-fold (alias CV k-fold) est une méthode de rééchantillonnage qui divise aléatoirement les données d'apprentissage en k groupes (alias folds) de taille approximativement égale. Le modèle est ajusté sur k-1 groupes, puis le groupe restant est utilisé pour calculer les performances du modèle. Cette procédure est répétée k fois ; à chaque fois, un pli différent est traité comme l'ensemble de validation. Kim (2009) a montré que la répétition de k-fold CV peut contribuer à augmenter la précision de l'erreur de généralisation estimée. Par conséquent, pour les ensembles de données plus petits (disons n < 10 000), un CV 10 fois ré-

78. BOEHMKE/GREENWELL : Hands-On Machine Learning with R (cf. note 71).

79. LEMMATISATION : Définition de LEMMATISATION, URL : <https://www.cnrtl.fr/definition/lemmatisation> (visité le 27/04/2024).

80. BOEHMKE/GREENWELL : Hands-On Machine Learning with R (cf. note 71).

pété 5 ou 10 fois améliorera la précision de votre performance estimée et fournira également une estimation de sa variabilité.

Pour notre prototype, nous réaliserons donc une validation croisée répétée 10 fois.

Pour le développement de l'algorithme, nous utiliserons le logiciel open source RStudio, ainsi que les librairies R suivantes : TM, NLP, Stingr et Randomforest. Afin d'éviter de surcharger la lecture, le code complet du prototype se trouve en annexe du mémoire.

3.1.4 Vérification des résultats prédictifs via la visualisation

Nous disposons de différents moyens de visualisation pour évaluer les performances du prototype ML, puis pour l'optimiser via plusieurs scénarios de paramétrage.

Les modes de vérification des résultats

Dans son article de blog *Interpreting Random Forests - Towards Data Science*⁸¹, Mariya Mansurova explique :

Pour la classification, au lieu de la valeur moyenne, nous utilisons la classe la plus courante comme prédiction pour chaque nœud feuille. Nous utilisons généralement le coefficient de Gini pour estimer la qualité de la division binaire pour la classification. (...) Si notre modèle de classification est parfait, le coefficient de Gini est égal à 0. Dans le pire des cas ($p = 0,5$), le coefficient de Gini est égal à 0,5.

Pour évaluer la qualité de la classification, nous ajouterons donc également au traitement des données le calcul du **coefficient de Gini**. Mansurova ajoute que les *Random-Forest* sont basées sur le concept de l'échantillonnage, et que l'objectif est d'utiliser une prédiction moyenne à partir de ces modèles : comme ils sont indépendants, les erreurs ne sont pas corrélées, et le modèle ne peut donc se suradapter (*overfitting*) en ajoutant des arbres supplémentaires. Il faut juste vérifier qu'on dispose de plusieurs arbres.

81. Mariya MANSUROVA : Interpreting Random Forests. Medium, 8 oct. 2023, (visité le 21/04/2024).

Elle explique également que nous pourrons ensuite calculer **les taux d'erreur OOB** (*Out of bag*) sur la base des prédictions : si l'erreur OOB est beaucoup plus proche de l'erreur sur l'ensemble de validation que de celle pour l'entraînement, cela signifie qu'il s'agit d'une bonne approximation. Nous procéderons donc à la comparaison des taux d'erreur OOB pour chaque scénario de paramétrage du prototype.

Bradley Boehmke et Brandon Greenwell⁸² décrivent le fonctionnement de la **matrice de confusion**, appelée jusqu'ici la matrice faux positifs/faux négatifs.

Lors de l'application de modèles de classification, nous utilisons souvent une matrice de confusion pour évaluer certaines mesures de performance.

Une matrice de confusion est simplement une matrice qui compare les niveaux catégoriels réels (ou événements) aux niveaux catégoriels prédits.

Lorsque nous prédisons le bon niveau, nous parlons d'un vrai positif. Toutefois, si nous prédisons un niveau ou un événement qui ne s'est pas produit, il s'agit d'un faux positif (par exemple, nous avions prédit qu'un client utiliserait un bon de réduction, mais il ne l'a pas fait). À l'inverse, lorsque nous ne prédisons pas un niveau ou un événement et qu'il se produit, il s'agit d'un faux négatif (par exemple, un client dont nous n'avions pas prédit qu'il utiliserait un bon de réduction l'a fait). Nous pouvons extraire différents niveaux de performance pour les classificateurs binaires.

Avec la matrice de confusion, nous allons pouvoir identifier si la classification fonctionne, et comparer les résultats prédictifs des différents scénarios.

Pour évaluer la **précision** des résultats prédictifs du prototype ML, nous calculerons un ratio entre les vrais positifs et la somme des vrais positifs et des faux négatifs de la matrice de confusion, pour chaque validation croisée. Boehmle et Greenwell⁸³ précisent :

Précision : Dans quelle mesure le classificateur prédit-il les événements avec précision ? Cette mesure vise à maximiser le ratio vrais positifs/faux

82. BOEHMKE/GREENWELL : Hands-On Machine Learning with R (cf. note 71).

83. Ibid.

positifs. En d'autres termes, pour le nombre de prédictions que nous avons faites, combien étaient correctes ? Objectif : maximiser

Afin de vérifier les termes identifiés comme discriminants par le prototype, nous allons utiliser la technique du **distant reading**, par la visualisation des premiers termes discriminants. Nicole M. Brown et al.⁸⁴ définissent ainsi le distant reading :

Le distant reading permet aux chercheurs de confirmer les modèles connus et attendus et d'étendre l'analyse de ces thèmes pour voir quelles nouvelles connexions potentielles peuvent être découvertes. Cette première phase de distant reading est essentielle au processus, car elle indique aux chercheurs si le modèle est solide et si nous pouvons aller de l'avant dans notre analyse et nos interprétations.

Ensuite, pour la discussion, nous exporterons la liste complète de ces termes pour les comparer dans la discussion avec les listes d'autres études.

Les modes d'optimisation des résultats

L'optimisation du prototype ML peut être réalisée via des scénarios d'hyperparamétrages de l'algorithme et de traitement des données.

Bradley Boehmke et Brandon Greenwell⁸⁵ expliquent l'hyperparamétrage :

Bien que les *RandomForests* donnent de bons résultats prêts à l'emploi, il existe plusieurs hyperparamètres accordables que nous devons prendre en compte lors de l'entraînement d'un modèle. Bien que nous discutions brièvement des principaux hyperparamètres, Probst et al. (2019) fournissent une discussion beaucoup plus approfondie. Les principaux hyperparamètres à prendre en compte comprennent : (1) le nombre d'arbres dans la forêt (2) Le nombre de caractéristiques à prendre en compte à chaque division (...) (3) La complexité de chaque arbre (4) Le schéma d'échantillonnage (5) La règle de division à utiliser lors de la construction de l'arbre

84. BROWN et al. : In Search of Zora/When Metadata Isn't Enough (cf. note 66).

85. BOEHMKE/GREENWELL : Hands-On Machine Learning with R (cf. note 71).

Pour ce mémoire, nous nous limiterons à différents scénarios de nombre d'arbres, et de nombre de permutations d'arbres. Boehmke et Greenwell⁸⁶ les détaillent ainsi :

Nombre d'arbres

Le premier élément à prendre en compte est le nombre d'arbres de votre forêt aléatoire. Bien qu'il ne s'agisse pas techniquement d'un hyperparamètre, le nombre d'arbres doit être suffisamment important pour stabiliser le taux d'erreur. Une bonne règle empirique consiste à commencer avec 10 fois le nombre de caractéristiques (...); cependant, à mesure que vous ajustez d'autres hyperparamètres (...), un plus grand nombre ou un plus petit nombre d'arbres peut être nécessaire. Un plus grand nombre d'arbres permet d'obtenir des estimations d'erreur et des mesures d'importance variables plus robustes et plus stables; toutefois, l'impact sur le temps de calcul augmente linéairement avec le nombre d'arbres.

Permutations

L'hyperparamètre qui contrôle la caractéristique de randomisation des variables divisées des forêts aléatoires (...) permet d'équilibrer une faible corrélation entre les arbres et une force prédictive raisonnable.

Dans la librairie *RandomForest*, les permutations oscillent entre les valeurs 0 et 2.

Comme scénarios de paramétrage du traitement des données, nous évaluerons également des scénarios sans et avec **lemmatisation**, ainsi que la sélection de différents pourcentages des **mots les plus rares**.

Pour la visualisation des résultats de la prédiction et l'optimisation du prototype ML, nous utiliserons les librairies R suivantes : Randomforest, udpipe, dplyr.

3.2 Vérification et correction de la qualité des données

D'abord, nous allons explorer et préparer les données au *Fitness for use*, puis nous évaluerons la pertinence et la qualité des données du corpus.

86. BOEHMKE/GREENWELL : Hands-On Machine Learning with R (cf. note 71).

TABLE 3 – Architecture de la table fournie

Libellé du champ	Description du champ
id_inscription	Identifiant de l'inscription à la séance d'information
id_usager_dorifor	Identifiant de compte sur la plateforme Dorifor; clé étrangère
id_demande	Identifiant de la candidature; clé étrangère
prescripteur	Organisme à l'origine de l'inscription
sexe via RN	Sexe identifié à partir du numéro de registre national
année naissance via RN	Année de naissance identifiée à partir du numéro de registre national
num référence SI	Identifiant de la séance d'information; clé étrangère
num référence fiche	Identifiant de la fiche formation ; clé étrangère
titre fiche	Titre de la fiche formation
chapeau	Introduction de la fiche formation
organisme	Institution dans laquelle se déroule la séance d'information
objectif	Objectif de la formation
programme	Programme de la formation
pré-requis admin - statut	Statut administratif requis pour pouvoir participer à la formation (liste de choix)
pré-requis admin	Statut administratif requis pour pouvoir participer à la formation (champ libre)
connaître	Connaissances requises pour participer à la formation
remarques	Remarques sur la séance d'information ou sur le formation
code formation BF	Identifiant de la formation Bruxelles Formation; clé étrangère

3.2.1 Exploration et préparation des données

La table fournie par la data analyst de Bruxelles Formation doit être explorée et éventuellement adaptée pour pouvoir répondre aux critères d'utilisation *Fitness for use*.

Exploration des données

Le jeu de données fourni se présente sous la forme d'une feuille de fichier Excel, reprenant 15 812 profils de candidature pseudonymisés, ainsi que le contenu des fiches formation liées (Table 3).

Une candidature [id_inscription] est encodée soit par la candidate ou le candidat lui-même via le module d'inscription Dorifor [id_usager_dorifor] à une séances d'information [num référence SI], soit via un organisme prescripteur [prescripteur] (call

center, Cité des métiers, Actiris, etc.). La data analyst de Bruxelles Formation a supprimé le numéro de registre national des profils, mais nous en a communiqué le sexe [sexe via RN] et l'année de naissance [année naissance via RN].

En effet, le numéro de registre national⁸⁷ est composé de plusieurs parties :

A A M M J J S S S C C

- A A M M J J : le groupe date de naissance avec les 2 derniers chiffres de l'année (A A), 2 chiffres pour le mois (M M), et 2 chiffres pour le jour (J J)
- S S S : le groupe numéro d'ordre est constitué par le rang d'inscription de la personne dans le groupe date de naissance; "A une personne du sexe féminin est attribué un numéro d'ordre pair, à une personne du sexe masculin est attribué un numéro d'ordre impair."
- groupe CC : nombre de contrôle

Chaque profil est lié à une fiche formation [num référence fiche] et à son contenu [titre fiche], [chapeau], [organisme], [objectif], [programme], [pré-requis admin - statut], [pré-requis admin], [remarques], et à une formation [code formation BF].

Préparation des données pour l'analyse

Tout d'abord, il est nécessaire de dénormaliser les données fournies pour créer un jeu de données adapté au *Fitness for use* : nous excluons donc les clés étrangères inutilisées, les noms des prescripteurs, et l'année de naissance, et nous renommons l'identifiant de la fiche formation, les deux champs du pré-requis administratif, et le champ connaître.

Pour chaque enregistrement, on dispose désormais de la clé de profil [id_inscription], de la donnée discriminante [sexe] et des champs de la fiche formation pour laquelle la personne a candidaté : [titre], [chapeau], [objectif], [programme], [pre_requis_liste], [pre_requis_libre], [competences_requises], [remarques].

Comme l'extraction finale annoncé était de 15 777 candidatures, et que le fichier comprenait 15 812 profils, un dédoublonnage était nécessaire, et effectivement, le fi-

87. REGISTRE NATIONAL DES PERSONNES PHYSIQUES, URL : <https://www.ibz.rrn.fgov.be> (visité le 23/03/2024).

chier est redescendu à 15 777 profils.

3.2.2 Évaluation de la pertinence des données

Le numéro de registre national possède une règle d'intégrité intégrée⁸⁸, l'extraction de l'information par ce biais est donc plus fiable que celle encodée manuellement. Les texte des fiches formation sont régulièrement mis à jour par les conseillères et conseillers d'orientation de la Cité des métiers, sur base du descriptif administratif des formations de Bruxelles Formation. Tout comme les profils des candidatures, nous pouvons donc les considérer comme du niveau de qualité de données *Business Intelligence*.

Nous prenons l'hypothèse que notre volume de données est suffisant. Mustafa Alabadia⁸⁹ a d'ailleurs démontré que le traitement par un jeu comportant des données manquantes ne génère pas de biais supplémentaires par rapport au traitement par ML d'un autre jeu de données.

Data profiling : évaluation de la qualité des données

Après avoir importé les fichiers dans OpenRefine, nous réalisons le *data profiling* du jeu de données.

La table des profils de candidatures possède 15 777 enregistrements. Avec le *Text facet*, j'ai pu identifier dans le champ [sexe] 8 834 "F" (pour femme) et 6 943 "H" (pour homme). Avec le filtre *Sort*, j'ai pu vérifier qu'il n'existe pas d'erreurs ni de valeurs nulles parmi la clé étrangère [id_fiche] (ce qui aurait supprimé de facto des données des données pour l'entraînement du prototype d'apprentissage supervisé).

Pour la distribution des données, nous disposons donc d'un ratio femmes-hommes 56%-44%.

La table des fiches formation contient 102 enregistrements. On y retrouve plusieurs types d'anomalies. Plusieurs enregistrements présentent des champs à valeurs nulles :

chapeau : 56

88. REGISTRE NATIONAL DES PERSONNES PHYSIQUES (cf. note 87).

89. ALABADIA et al. : Systematic Review of Using Machine Learning in Imputing Missing Values (cf. note 42).

objectif : 24

pre_requis_liste : 33; mais présente bien 3 valeurs possibles ("Etre chercheur d'emploi", "Etre chercheur d'emploi inoccupé", "Etre chercheur d'emploi complet indemnisé")

pre_requis_libre : 42

competences_requises : 38

Des balises HTML sont présentes dans les champs [programme], [pre_requis_libre], [competences_requises] et [remarques]. Les valeurs nulles et les balises HTML ne sont pas forcément des anomalies à corriger : les règles métier ne nous ont pas été communiquées, nous ne pouvons donc pas en juger.

Gestion des anomalies : standardisation des données

Nous n'avons pas identifié d'anomalies de standardisation : les formats des données numérique ou texte ont bien été respectés.

Gestion des anomalies : *data matching* des incohérences

Sans accès aux règles métier, il ne nous est pas possible d'identifier les anomalies de cohérence. Aucun golden record n'a été défini par Bruxelles Formation, qui considère la qualité des données de Dorifor comme non critique.

Après l'adaptation *Fitness for use* des données, nous n'avons pas relevé d'anomalies. Pendant le nettoyage, il faudra adapter le format du texte pour l'apprentissage supervisé.

3.3 Algorithme d'apprentissage supervisé

Nous importerons d'abord le jeu de données dans RStudio puis procéderons au nettoyage des données. Ensuite, nous distribuerons le jeu de données entre jeu d'apprentissage et jeu de test de la prédiction. Enfin nous construirons et appliquerons aux données notre modèle de classification *Randomforest*.

3.3.1 Import et nettoyage des données

A l'issue de la vérification et de la correction de la qualité des données, nous avons obtenu un jeu de données au format CSV directement importable dans R Studio. Comme identifié ci-avant, il est nécessaire de supprimer les balises HTML et le bruit (ponctuation, mots vides, espaces) avant de pouvoir être traitée par l'algorithme de ML.

Import des données

Dans cette première partie du code, nous devons d'abord nous assurer que l'encodage des caractères est adapté à l'interprétation de R, et procéder à un nettoyage complet des données pour permettre le traitement en R. L'encodage des caractères en UTF-8 est réalisé lors de l'import du fichier CSV :

```
data <- read.csv("traitement-2024-2.csv",
header= TRUE, encoding = "UTF-8")
```

Nettoyage des données

Le nettoyage des données passe par plusieurs phases.

Premièrement, nous transformons les données importée dans un dataframe (données affichées sous forme de tableau⁹⁰). Pour l'apprentissage supervisé, nous devons construire ce dataframe en séparant l'identifiant de chaque fiche dans un champ [doc_id] et sa description concaténée dans un champ [text].

```
colnames(data) [1] <- "doc_id"
data$text = paste(data$id_inscription,data$sexe,
data$titre, data$chapeau, data$objectif, data$programme,
data$pre_requis_liste, data$pre_requis_libre,
data$competences_requises, data$remarques)
data2 = subset(data, select = c(doc_id,text))
```

Ensuite, nous pouvons nettoyer les chaînes de caractères avec la librairie Stringr et des expressions régulières : suppression des balises HTML, puis du bruit généré par les éléments de mise en page du CSV, et la ponctuation.

90. DataframeSource function - RDocumentation, (visité le 01/04/2024).

```

##supprimer le HTML

data2$text = str_replace_all(data2$text,"</?[a-z]+/?>"," ")

##supprimer le bruit

data2$text = str_replace_all(data2$text,"\\n"," ") #sauts de ligne

data2$text = str_replace_all(data2$text,"[^[:^punct:]]"," ")

#ponctuation

```

Nous utilisons également la fonction iconv pour supprimer les accents.

```
data2$text <- iconv(data2$text, from="UTF-8", to="ASCII//TRANSLIT")
```

Pour utiliser les outils de traitement avancé de la librairie TM⁹¹, les données doivent être insérées dans un corpus de documents via la classe DataframeSource.

```
(ds <- DataframeSource(data2))

corpus_fiches <- Corpus(ds)
```

Nous pouvons ainsi terminer d'adapter la forme du texte au traitement automatique : transformation de tous les mots en lettres minuscules, suppression des mots vides en français, des éventuels restes de ponctuation, et des espaces blancs.

```
corpus_fiches <- tm_map(corpus_fiches, content_transformer(tolower))

corpus_fiches <- tm_map(corpus_fiches, removeWords,
                        stopwords("french"))

corpus_fiches <- tm_map(corpus_fiches, removePunctuation)

corpus_fiches <- tm_map(corpus_fiches, stripWhitespace)
```

3.3.2 Paramétrage des données

Afin de fixer un paramétrage de base à ces données avant traitement, nous les importons dans une matrice termes-document.

```
#Import des fiches nettoyées dans une matrice termes-documents

dtm <- DocumentTermMatrix(corpus_fiches)
```

91. Ingo FEINERER : Introduction to the tm Package Text Mining in R, in : <https://cran.r-project.org>.

Nous calibrons le prototype avec 10% des mots les plus rares ; nous reprendrons plus tard ce pourcentage comme un paramètre d'optimisation de l'algorithme.

```
removeCommonTerms <- function (x, pct)
{
  stopifnot(inherits(x, c("DocumentTermMatrix", "TermDocumentMatrix")),
            is.numeric(pct), pct > 0, pct < 1)
  m <- if (inherits(x, "DocumentTermMatrix"))
    t(x)
  else x
  t <- table(m$i) < m$ncol * (pct)
  termIndex <- as.numeric(names(t[t]))
  if (inherits(x, "DocumentTermMatrix"))
    x[, termIndex]
  else x[termIndex, ]
}
```

dtm <- removeCommonTerms(dtm, 0.1) #Pourcentage des mots les plus rares

Afin de faciliter la manipulation des données, nous transformons la matrice dense en matrice creuse.

```
#Transformer la matrice dense en matrice creuse
dtmM = as.matrix(dtm)
dtmD = as.data.frame.matrix(dtmM)
```

Afin de stocker ces données nettoyées et paramétrées, nous les exportons dans un fichier dédié.

```
write.table(dtmD, file="dtmd.txt", row.names=FALSE, sep="\t",
            quote=FALSE)
```

3.3.3 Construction de la boucle *Randomforest*

Avec la librairie NLP, nous importons les données précédemment stockées.

```

#Import des données nettoyées et paramétrées

dtmD <- read.table(file="dtmd.txt", as.is = TRUE, header = TRUE,
sep = "\t", row.names = NULL)

```

Nous y ajoutons une colonne pour les données discriminantes, F ou H. Les données existantes y sont ajoutées; l'algorithme complétera le jeu avec les données prédites.

```

#Ajout à la matrice de la colonne des profils

data <- read.csv2("traitement-2024-2.csv", header = TRUE,
encoding = "UTF-8")

colnames(data)[1] <- "doc_id"

data_sexe = subset(data, select = c(doc_id, sexe))

dtmD[1]

dtmD = transform(dtmD, maPrediction = data_sexe$sexe)

dtmD$maPrediction = factor(dtmD$maPrediction)

```

Avant le lancement de la boucle d'apprentissage, nous ajoutons 10 tours validations croisées, comme recommandé dans la méthodologie. Nous définissons une clé de repère "DO" pour la segmentation du jeu de données pour l'entraînement et le test de la prédition. Et nous déclarons la liste *accuracies* qui nous permettra de stocker la précision des prédictions de chaque scénario.

```

#Paramétrage du prototype

K = 10 #validation croisée

DO = dim(dtmD)[1]

accuracies = array(NA, K)

```

La boucle d'apprentissage et de test rassemble : la segmentation des données entre jeu de données d'entraînement et jeu de données de test de la prédition, le lancement de l'entraînement d'apprentissage et des prédictions, et l'évaluation des tests de prédition. Nous y ajoutons un index Seed, afin de pouvoir reproduire cette partition.

```
#Lancement de la boucle de Random forest
```

```

for (i in 1:K){

  #Segmentation des données pour l'entraînement et le test de
  la prédition

  set.seed(i) #indice de reproductibilité

  indicesTrain = sample(1:D0, size = floor(0.7*D0)) #70% des données
  pour l'entraînement

  indicesTest = setdiff(1:D0, indicesTrain) #30% des données
  pour le test
}

```

Nous paramétrons le *Randomforest* avec 5 arbres et 1,1 permutations.

```

#Lancement de l'entraînement d'apprentissage et des prédictions
modeleRF01 = randomForest(maPrediction ~ ., data=dtmD[indicesTrain,],
  ntree = 5, nPerme=1.1) #Nombre d'arbres et de permutations
predTestRF01 = predict(modeleRF01, newdata = dtmD[indicesTest,])

```

La matrice de confusion matConf compte pour chaque tour de validation croisée la différence entre valeur prédite et valeur attendue. La liste *accuracies* calcule et enregistre le taux de précision, en divisant les vrais positifs par la somme des vrais positifs, des vrais négatifs et des faux positifs.

```

#Evaluation des tests de la prédition
matConf = table(predTestRF01, dtmD$maPrediction[indicesTest])
accurrancies[i] = sum(diag(matConf)) / sum(matConf)
print(c(i,accurrancies[i]))
}

```

A partir des données brutes fournies, nous avons nettoyé, paramétré et traité les données par un algorithme d'apprentissage supervisé de ML. Il faut maintenant vérifier les résultats obtenus et optimiser le prototype.

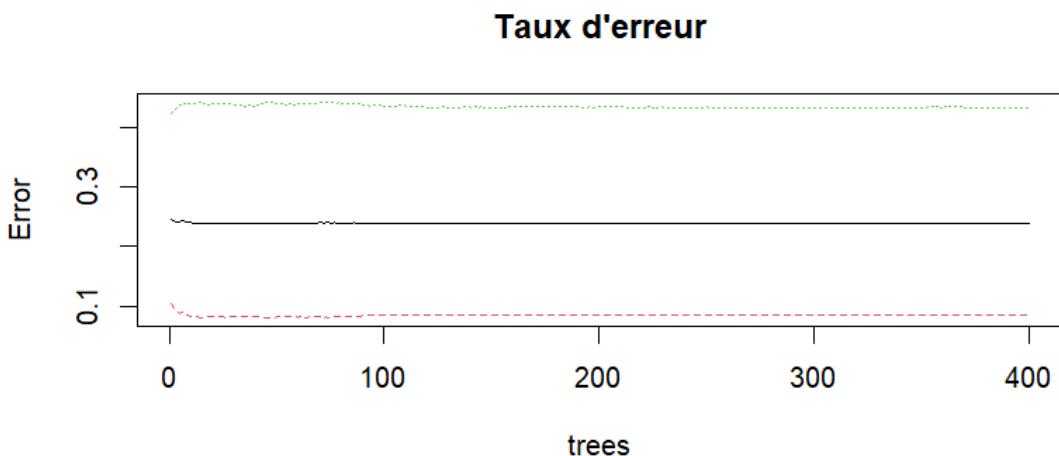


FIGURE 22 – Évolution du taux d'erreurs en fonction du nombre d'arbres

3.4 Vérifier les résultats prédictifs via la visualisation

Avec les scénarios d'hyperparamétrage de l'algorithme, nous allons tenter d'identifier le point de stabilisation des résultats du *RandomForest*, en croisant plusieurs nombres d'arbres et plusieurs permutations d'arbres; nous évaluerons la qualité finale de la classification avec le coefficient de Gini. Avec les scénarios d'alternatives de préparation des données, nous comparerons les résultats sans et avec lemmatisation, puis différents pourcentages des mots les plus rares.

3.4.1 Scénarios d'hyperparamétrages

En testant plusieurs scénarios d'hyperparamétrage, nous allons tenter d'identifier en premier lieu le point à partir duquel le taux d'erreur estimé remonte, et ensuite comparer les taux de précision obtenus pour chaque scénario.

En visualisant les résultats dans un graphique sur le scénario le plus important, soit 400 arbres et 2 permutations, la stabilisation du modèle n'est pas vraiment visible (Figure 22).

```
plot(modeleRF01,main ="Taux d'erreur")
```

Pour l'observer, il est nécessaire de zoomer sur l'intervalle de taux d'erreur 0,2378-0,244 (Figure 23).

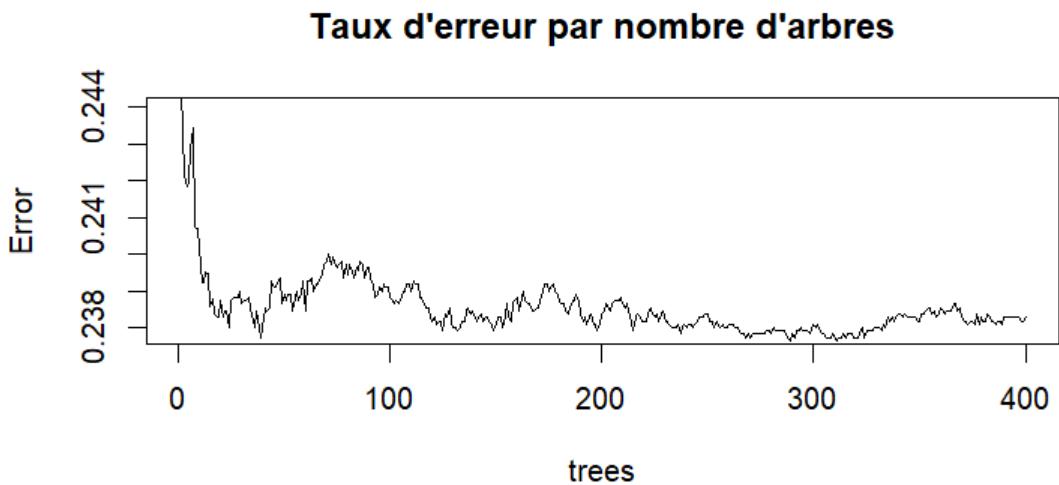


FIGURE 23 – Évolution du taux d'erreurs en fonction du nombre d'arbres : zoom

```
plot(modeleRF01,main ="Taux d'erreur par nombre d'arbres",
      ylim = c(0.2378,0.244))
```

Le taux d'erreur le plus bas apparaît donc aux alentours de 300 arbres.

```

Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,      ],
              Type of random forest: classification
              Number of trees: 300
No. of variables tried at each split: 47

          OOB estimate of error rate: 23.81%
Confusion matrix:
      F     H class.error
F 5673 528 0.08514756
H 2107 2759 0.43300452

```

FIGURE 24 – Matrice de confusion pour le scénario 300 arbres, 0,1 permutations

```

Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,      ],
              Type of random forest: classification
              Number of trees: 300
No. of variables tried at each split: 47

          OOB estimate of error rate: 23.81%
Confusion matrix:
      F     H class.error
F 5673 528 0.08514756
H 2107 2759 0.43300452

```

FIGURE 25 – Matrice de confusion pour le scénario 300 arbres, 1,1 permutations

Nous faisons osciller le nombre de permutations pour identifier le meilleur scénario avec 300 arbres. Nous comparons les matrices de confusion avec les permutations 0,1 (Figure 24), 1,1 (Figure 25), et 2,0 (Figure 26).

modeleRF01

Pour les 3 scénarios, le taux d'erreur OOB reste le même à 23,81%. Nous pouvons donc en conclure que les permutations des arbres n'ont pas d'impact sur les taux d'erreur du prototype.

Dans les matrices de confusion, nous observons que les taux de faux positif F-H et de faux positif H-F sont très différents. Nous retrouvons cette différence dans les taux d'erreur de classification : il y a une bien meilleure classification pour la prédiction des offres de formation pour les femmes (0,08) que pour les hommes (0,43).

```

Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,      ],
              Type of random forest: classification
              Number of trees: 300
No. of variables tried at each split: 47

          OOB estimate of error rate: 23.81%
Confusion matrix:
      F     H class.error
F 5673 528 0.08514756
H 2107 2759 0.43300452

```

FIGURE 26 – Matrice de confusion pour le scénario 300 arbres, 2,0 permutations

Afin de mesurer la qualité de la classification, nous calculons maintenant le coefficient de Gini pour 300 arbres (Figure 27). Pour cela, nous utilisons les librairies tibble et ggplot2.

```
#Hiérarchisation des termes discriminants par coefficient de Gini
impRF01 = importance(modeleRF01)

nomsVar = rownames(impRF01)

impRF01 = impRF01[impRF01[,1]>0,1]

resultatsRF01 = tibble(variable = names(impRF01), \\
coeffGini = impRF01)

resultatsRF01$variable = factor(resultatsRF01$variable,
levels = resultatsRF01$variable[order(resultatsRF01$coeffGini,
decreasing=FALSE)])

resultatsRF02 = resultatsRF01[order(resultatsRF01$coeffGini
decreasing=FALSE),]

#Visualisation des termes discriminants hiérarchisés
par coefficient de Gini
print(resultatsRF02,n=20)

ggplot(resultatsRF02) +
geom_col(aes(x = variable, y = coeffGini)) +
coord_flip()
```

Comme indiqué dans la méthodologie, la classification sera de qualité si le coefficient de Gini est proche de 0, et en tout cas inférieur à 0,5. Il faut donc zoomer pour évaluer les résultats.

```
#Zoom sur le coefficient de Gini
print(length(resultatsRF02[["coeffGini"]]))
resultatsRF03 = subset(resultatsRF02, coeffGini < 0.5)
print(length(resultatsRF03[["coeffGini"]]))
```

Nous excluons du jeu de données les coefficients de Gini supérieurs à 0,5, ce qui réduit

	variable	coeffGini
1	accrocheur	0.00000131
2	wordpress	0.00000182
3	envisagez	0.00000261
4	cuisinier	0.00000372
5	directs	0.00000662
6	correctement	0.0000135
7	deplacer	0.0000181
8	nouveaux	0.0000292
9	x160	0.0000315
10	legendes	0.0000393
11	oil	0.0000417
12	formatage	0.0000435
13	illustrations	0.0000454
14	pensez	0.0000620
15	source	0.0000633
16	export	0.0000726
17	epreuve	0.0000848
18	filigrane	0.0000863
19	comprenant	0.0000969
20	sons	0.000105
	# i 2,209 more rows	

FIGURE 27 – Coefficient de Gini pour 300 arbres (extrait)

le nombre de résultats de 2 229 à 1 659 termes. Nous estimons les meilleurs résultats à un coefficient de Gini inférieur à 0,1.

```
#Zoom sur les résultats de meilleure qualité
resultatsRF04 = subset(resultatsRF02, coeffGini < 0.1)
print(length(resultatsRF04[["coeffGini"]]))
write.csv(resultatsRF04,"resultatsRF04-300-11.csv",row.names = FALSE)
```

Le nombre de résultats est réduit à 973, soit 44% du jeu de données initial. Nous exportons ces résultats pour être comparés en distant reading pendant la discussion.

Afin de mesurer la précision des prévisions du prototype, nous faisons tourner l'algorithme sur 9 scénarios : selon 4 nombres d'arbres (100, 200, 300, 400) et selon 3 valeurs de permutation (0,1, 1,1, 2). Pour chacune des 10 validations croisées, nous obtenons un taux de précision. Nous intégrons ces résultats dans Excel (Figure 28).

Pour chaque scénario, nous réalisons une moyenne des taux de précision, puis identifions quels scénarios obtiennent la meilleure précision : les scénarios avec 100 et 300 arbres obtiennent les meilleurs résultats.

Nous concluons de cette partie que le nombre d'arbres menant au taux d'erreur le plus bas n'est pas nécessairement le seul à amener à la meilleure précision des résultats

Validation croisée	100-01	100-11	100-20	200-01	200-11	200-20	300-01	300-11	300-20	400-01	400-11	400-20
1	7,71079E+14	7,7108E+14	7,7108E+14	7,7108E+14								
2	7,57589E+14	7,5759E+14	7,5759E+14	7,5759E+14								
3	7,60118E+14	7,6012E+14	7,6012E+14	7,6012E+14								
4	7,53162E+14	7,5316E+14	7,5316E+14	7,5316E+14								
5	7,57378E+14	7,5738E+14	7,5738E+14	7,5738E+14								
6	7,52951E+14	7,5126E+14	7,5126E+14	7,5126E+14								
7	7,67074E+14	7,67074E+14	7,67074E+14	7,6602E+13	7,6602E+13	7,6602E+13	7,67074E+14	7,67074E+14	7,67074E+14	7,6707E+14	7,6707E+14	7,6707E+14
8	7,61172E+14	7,6117E+14	7,6117E+14	7,6117E+14								
9	7,65388E+14	7,6539E+14	7,6539E+14	7,6539E+14								
10	7,55481E+14	7,5548E+14	7,5548E+14	7,5548E+14								
Précision moyenne	7,60139E+14	7,60139E+14	7,60139E+14	6,91092E+14	6,91092E+14	6,91092E+14	7,60139E+14	7,60139E+14	7,60139E+14	7,5997E+14	7,5997E+14	7,5997E+14
Meilleures précisions MAX	MAX	MAX	FAUX	FAUX	FAUX	MAX	MAX	MAX	MAX	FAUX	FAUX	FAUX

FIGURE 28 – Analyse des taux de précision

```
Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,           ],
              Type of random forest: classification
              Number of trees: 100
No. of variables tried at each split: 50

          OOB estimate of  error rate: 23.74%
Confusion matrix:
             F      H class.error
F 5675 526 0.08482503
H 2101 2765 0.43177148
```

FIGURE 29 – Matrice de confusion pour le scénario 100 arbres, permutations 1,1

prédictifs, et que la permutation des arbres n'affecte pas le taux d'erreur de prédiction pour ce prototype.

Nous pouvons vérifier cette observation au niveau de la matrice de confusion du scénario 100 arbres et 1,1 permutations (Figure 29). La matrice de confusion du scénario 100 arbres et permutations 1,1 donne un taux d'erreur de 23,74%, soit un taux inférieur aux scénarios à 300 arbres.

Nous pourrons donc réaliser la suite de l'optimisation avec le scénario 100 arbres et 1,1 permutations.

3.4.2 Scénarios sans et avec lemmatisation

Nous allons identifier quel scénario délivre la meilleure précision. Pour le prototype sans lemmatisation, nous reprenons la matrice de confusion obtenue à partir de l'hyperparamétrage 100 arbres et 1,1 permutations (Figure 30).

Pour le prototype avec lemmatisation, nous ajoutons une étape de traitement à l'algorithme. Nous utilisons la librairie udpipe et le modèle french-gsd avec les données nettoyées, puis réindexons les lignes du dataframe.

```
data <- read.csv("data_clean.csv", header = TRUE, encoding = "UTF-8")
```

```

Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,      ],
              Type of random forest: classification
              Number of trees: 100
No. of variables tried at each split: 50

          OOB estimate of  error rate: 23.74%
Confusion matrix:
      F     H class.error
F 5675 526 0.08482503
H 2101 2765 0.43177148

```

FIGURE 30 – Matrice de confusion pour le scénario 100 arbres, permutations 1,1, sans lemmatisation

```

texteAnalyse = udpipe(x = data, "french-gsd", trace = FALSE)
rownames(texteAnalyse)

```

Ensuite, nous réalisons un subset du jeu de données et, avec la librairie dplyr, consolidons les résultats puis les exportons en CSV pour pouvoir ensuite en réintroduire les données dans le processus de paramétrage.

```

data_lem2 <- subset(texteAnalyse, select= c(doc_id,lemma))
data_lem3 <- data_lem2 %>%
  group_by(doc_id) %>%
  summarise(lemma = toString(lemma))

```

En effet, le processus de lemmatisation divise les données en autant de lignes que de lemmes. Il est donc nécessaire de regrouper ces données au bon format, soit un dataframe avec les colonnes doc_id et text.

Enfin, nous pouvons réintroduire les données lemmatisées dans le processus de paramétrage des données, puis dans l'algorithme (cf. supra).

Nous pouvons alors générer la matrice de confusion pour ce nouveau scénario (Figure 31). Avec la lemmatisation, nous observons que le taux d'erreur OOB augmente fortement, passant de 23,74% à 44,24%.

Nous pratiquons la même manipulation que plus haut pour comparer les précisions moyennes (Figure 32). Avec la lemmatisation, la précision moyenne diminue aussi fortement, passant de $7,64 * 10^{14}$ à $5,02 * 10^{14}$.

Nous concluons de ces résultats que la lemmatisation n'apporte aucune optimisation au prototype.

```

Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,      ],
              Type of random forest: classification
              Number of trees: 100
No. of variables tried at each split: 44

          OOB estimate of  error rate: 44.24%
Confusion matrix:
   F   H class.error
F 5771 446  0.07173878
H 4439 387  0.91980937

```

FIGURE 31 – Matrice de confusion pour le scénario 100 arbres, permutations 1,1, avec lemmatisation

Validation croisée	100-11-sanslem	100-11-aveclem
1	7,71079E+14	5,51753E+14
2	7,57589E+14	5,56612E+14
3	7,60118E+14	5,58513E+14
4	7,53162E+14	5,50063E+14
5	7,57378E+14	5,53021E+14
6	7,52951E+14	5,4774E+13
7	7,67074E+14	5,50063E+14
8	7,61172E+14	5,59358E+14
9	7,65388E+14	5,50697E+14
10	7,55481E+14	5,41825E+14
Précision moyenne		7,60139E+14
Meilleures précisions MAX		FAUX

FIGURE 32 – Analyse des taux de précision

3.4.3 Scénarios des pourcentages des mots les plus rares

Pour les mots les plus rares conservés, nous conservons le scénario 100 arbres et 1,1 permutations. Vu les résultats précédents, nous excluons la lemmatisation de ce scénario.

Pour tester ce paramètre d'optimisation, nous allons faire osciller le pourcentage des mots les plus rares. Notre prototype de base utilisait les 10% des mots les plus rares; nous allons tester l'oscillation sur 5% (Figure 33) et sur 15% (Figure 35). Comme précédemment, nous allons comparer les taux d'erreur puis les précisions. Ensuite, nous utiliserons le distant reading pour observer comment les différences de pourcentages de mots se traduisent au niveau des mots eux-mêmes.

Nous pouvons donc reprendre la matrice de confusion de notre scénario 100 arbres, 1,1 permutations et les 10% des mots les plus rares (Figure 34). De manière inattendue, le scénario à 10% des mots les plus rares conserve le meilleur taux d'erreur, à 23,74%.

Comme précédemment, nous comparons les précisions via Excel (Figure 36).

```

Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,      ],
              Type of random forest: classification
              Number of trees: 100
No. of variables tried at each split: 46

          OOB estimate of error rate: 24.11%
Confusion matrix:
      F     H class.error
F 5742  475  0.07640341
H 2187 2639  0.45317033

```

FIGURE 33 – Matrice de confusion pour le scénario 100 arbres, permutations 1,1, avec les 5% des mots les plus rares

```

Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,      ],
              Type of random forest: classification
              Number of trees: 100
No. of variables tried at each split: 50

          OOB estimate of error rate: 23.74%
Confusion matrix:
      F     H class.error
F 5675  526  0.08482503
H 2101 2765  0.43177148

```

FIGURE 34 – Matrice de confusion pour le scénario 100 arbres, permutations 1,1, avec les 10% des mots les plus rares

```

Call:
randomForest(formula = maPrediction ~ ., data = dtmD[indicesTrain,      ],
              Type of random forest: classification
              Number of trees: 100
No. of variables tried at each split: 50

          OOB estimate of error rate: 24.16%
Confusion matrix:
      F     H class.error
F 5718  499  0.08026379
H 2169 2657  0.44944053

```

FIGURE 35 – Matrice de confusion pour le scénario 100 arbres, permutations 1,1, avec les 15% des mots les plus rares

Validation croisée	100-11-005	100-11-01	100-11-015	
1	7,64259E+13	7,64259E+13	7,64259E+13	
2	7,62569E+14	7,61935E+14	7,61935E+14	
3	7,55387E+14	7,57921E+14	7,55387E+14	
4	7,6447E+14	7,6447E+14	7,6447E+14	
5	7,55387E+14	7,56232E+13	7,55387E+14	
6	7,50951E+14	7,51584E+14	7,50106E+14	
7	7,57288E+14	7,57288E+14	7,57288E+14	
8	7,66793E+14	7,65104E+14	7,65104E+14	
9	7,6278E+14	7,6278E+14	7,6278E+14	
10	7,55175E+14	7,54964E+13	7,54964E+13	
	Précision moyenne	6,90722E+14	5,54863E+14	6,22438E+14
	Meilleures précisions	MAX	FAUX	FAUX

FIGURE 36 – Analyse des taux de précision

Cette fois-ci, c'est le scénario à 5% des mots les plus rares qui obtient le meilleur taux de précision.

Pour les scénarios variant au niveau des pourcentages des mots les plus rares, une fois encore le taux d'erreur le plus bas n'obtenait pas le meilleur taux de précision. Le scénario ayant obtenu le meilleur taux de précision est : 100 arbres, 1,1 permutations, 5% des mots les plus rares.

3.5 Discussion

Notre algorithme d'apprentissage supervisé a appris par classification quelles fiches formation avaient plus de probabilités de recevoir des candidatures de femmes ou d'hommes.

Vérification des hypothèses Avec le coefficient de Gini, nous lui avons demandé de générer la liste des termes avec la qualité de prévision la meilleure, soit les termes les plus discriminants, ceux qui orientaient le plus soit vers des femmes, soit vers des hommes. Nous pouvons maintenant comparer ces résultats avec ceux obtenus par Océane Couillaud⁹², par distant reading. Dans son mémoire de master, elle a construit un thésaurus masculin-féminin à partir des études de Bem (1974), Ahl (2004) et Gaucher et al. (2011) (Figures 37 et 38).

92. COUILLAUD : Le Genre Dans l'Education Entrepreneuriale (cf. note 7).

Tableau 3.3 Thésaurus de la masculinité et féminité

Masculinité	Féminité
Active	Affectionate
Adventurous	Cheer*
Aggress*	Child*
Ambitio*	Commit*
Analy*	Communal
Assert*	Compassion*
Astute	Connect*
Athlet*	Considerate
Autonom*	Cooperat*
Boast*	Depend*
Challeng*	Emotiona*
Compet*	Empath*
Confident	Feminine
Courag*	Flatterable
Daring	Gentle
Decide	Gullible
Decision*	Honest
Decisive	Interdependen*
Detached	Interpersona*
Determin*	Kind
Domina*	Kinship
Force*	Loyal*
Foresighted	Modesty
Greedy	Nag
Headstrong	Nurtur*

FIGURE 37 – Thésaurus de la masculinité et féminité (p.1 sur 2)

Tableau 3.3 Thésaurus de la masculinité et féminité (suite et fin)

Hierarch*	Pleasant*
Hostil*	Polite
Impulsive	Quiet*
Independen*	Respon*
Individual*	Sensitiv*
Influential	Shy
Intellect*	Soft-spoken
Intelligent	Submissive
Lead*	Support*
Logic	Sympath*
Manager	Tender*
Masculine	Together*
Objective	Trust*
Opinion	Understand*
Optimistic	Warm*
Outspoken	Whin*
Perceptive	Yield*
Persist	
Pilot	
Principle*	
Reckless	
Resolute	
Self-centered	
Self-confiden*	
Self-efficacious	
Self-relian*	
Self-sufficien*	
Strong-willed	
Stubborn	
Superior	
Visionary	

FIGURE 38 – Thésaurus de la masculinité et féminité (p.2 sur 2)

prototype ML	coeffGini	Masculin	Feminin
boites	0,000000002	avis	agréable*
espacebr	0,000000468	domina*	ensemble
training	0,000001204	tétu*	ton*
eme	0,000002468	domina*	dou*
decorateur	0,000004105	décision*	crédule
declarations	0,000005263	décision*	crédule
divers	0,000007229	dirige*	dépendan*
feuilles	0,000008119	domina*	féminin*
insérer	0,000010665	influen*	honnête
validationdescompetences	0,000016359	tétu*	ton*
familiale	0,000019786	domina*	ensemble
determinant	0,000039362	détermin*	dépendan*
diversité	0,000048390	dirige*	dépendan*
organisées	0,000051516	optimiste	nourri*
sources	0,000053406	résolu*	soumis*
tugsten	0,000060665	tétu*	ton*
proximité	0,000065399	principe	poli
partenaires	0,000081154	optimiste	parent*
jobcoachs	0,000091962	intelligent*	interpersonnel
cesi	0,005365083	avis	céder
assimilees	0,014211179	assert*	agréable*
probleme	0,022711381	principe	poli
cire	0,024828381	avis	chaleur
maintenance	0,025384505	logique	loya*
transversale	0,047138597	tétu*	ton*
adequate	0,084964959	acti*	#N/A
illustrator	0,092075367	hostil*	honnête
chimiste	0,099577920	avis	chaleur
styles	0,101120429	résolu*	soutien*
administratifbr	0,109523601	acti*	#N/A

FIGURE 39 – Comparaison prototype-thésaurus

Nous allons comparer nos résultats à ceux de ce thésaurus. Préalablement, nous devons le traduire en français, puis en traiter le contenu pour identifier les similarités avec nos résultats. Nous utilisons DeepL pour la traduction brute, puis adaptions les termes traduits et enfin réalisons un premier comparatif de visualisation dans Excel avec la fonction RECHERCHEV, en activant la suggestion de valeur proche (Figure 39).

Dans cet extrait de la meilleure qualité de termes prédits, nous observons que la reconnaissance par suggestion de valeur proche est faible : nous avons noté en rouge les anomalies, et surligné en jaune les termes reconnus.

Afin d'isoler la liste complète des termes similaires entre les résultats du prototype et le thésaurus, nous utilisons R et la librairie stringr : nous adaptions le format des

termes du thésaurus (accents, casse), appliquons une fonction d'analyse et stockons les résultats dans des dataframes (Figure 40).

```
#Chargement des librairies

library(stringr) #nettoyage des chaînes de caractère

#Import des données

voc_fh <- read.csv2("voc-fh-2.csv", header = TRUE, \\
encoding = "UTF-8")

results <- read.csv("resultatsRF04-300-11.csv", header = TRUE,
encoding = "UTF-8")

#nettoyage

##supprimer les accents

voc_fh$X.U.FEFF.vocf <- iconv(voc_fh$X.U.FEFF.vocf, from="UTF-8",
to="ASCII//TRANSLIT")

voc_fh$voch <- iconv(voc_fh$voch, from="UTF-8",
to="ASCII//TRANSLIT")

##supprimer les majuscules

voc_fh$X.U.FEFF.vocf <- tolower(voc_fh$X.U.FEFF.vocf)

voc_fh$voch <- tolower(voc_fh$voch)

print(voc_fh)

##renommer les colonnes

colnames(voc_fh)[1] <- "vocf"

colnames(voc_fh)[2] <- "voch"

#comparer les listes de termes

## Fonction pour trouver des mots similaires entre deux colonnes

find_similar_words <- function(col1, col2) {

similar_words <- character()
```

```

> print(voc_f)
similar_words
1      compren*
2      conne*
3      dou*
4      emotion*
5      enfan*
6      engage*
7      ton*
8
> print(voc_h)
similar_words
1      acti*
2      decid*
3      decis*
4      determin*
5      gener*
6      superi*
7      sur*

```

FIGURE 40 – Termes similaires entre les résultats du prototype et le thésaurus

```

for (word1 in col1) {
  for (word2 in col2) {
    if (str_detect(word2, paste0("^", word1, ".*"))) {
      similar_words <- c(similar_words, word1)
      break
    }
  }
}

## Créer un dataframe avec les résultats
result_df <- data.frame(similar_words = unique(similar_words))
return(result_df)
}

## Appel de la fonction
voc_f <- find_similar_words(voc_fh$vocf, results$variable)
voc_h <- find_similar_words(voc_fh$voch, results$variable)
print(voc_f)
print(voc_h)

```

Parmi les termes, `voc_f` représente la liste des termes plutôt considérés comme féminins et `voc_h` comme masculins. Cette liste de termes discriminants est donc à la fois reconnue par notre état de l'art et vérifiée par notre étude de cas.

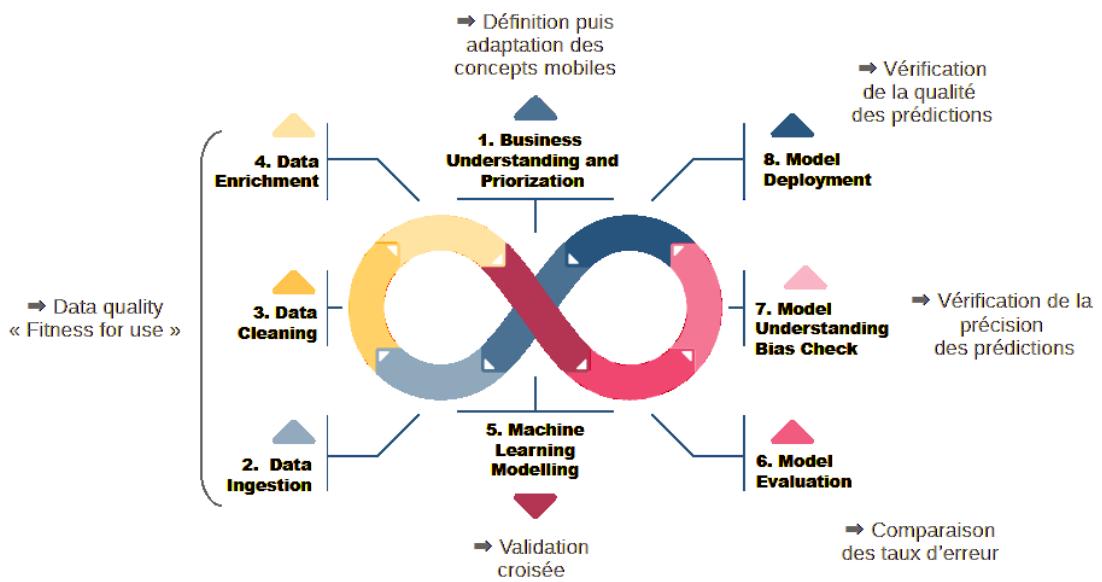


FIGURE 41 – Processus ML vu par Alberto Marocchina, et enrichi de nos indicateurs de qualité

3.5.1 Consolidation du modèle

Suite à la réussite de la démonstration de l'étude de cas, nous pouvons proposer d'enrichir le modèle proposé par Alberto Marocchina avec les différents indicateurs de qualité pour un processus ML (Figure 41).

Nous proposons d'ajouter des indicateurs de qualité à chaque étape du processus ML :

- 1. Compréhension et priorisation du *business* : la définition puis l'adaptation des concepts mobiles
- 2. à 4. Collecte, nettoyage et enrichissement des données : la qualité des données et le *Fitness for use*
- 5. Modélisation ML : la validation croisée
- 6. Evaluation du modèle : la comparaison des taux d'erreur de chaque scénario d'optimisation
- 7. Compréhension du modèle et vérification des biais : la vérification de la précision des prédictions
- 8. Déploiement du modèle : la vérification de la qualité des prédictions, à partir

de nouvelles sources extérieures

Nous concluons de la cohérence de ces résultats que les hypothèses et la méthodologie correspondent au *Fitness for use*.

3.6 Conclusion de l'étude de cas

Dans cette partie, nous avons d'abord décrit notre méthodologie pour la collecte des données, la vérification et la correction de la qualité des données, l'algorithme d'apprentissage supervisé et la vérification des résultats prédictifs via la visualisation. Nous avons ensuite réalisé cette méthodologie en vérifiant et corrigéant la qualité des données collectées, via l'exploration et la préparation des données, puis via l'évaluation de leur pertinence. Puis, nous avons nettoyé et paramétré les données, et construit la boucle de *RandomForest*. Enfin, nous avons vérifié par la visualisation puis optimisé les résultats selon différents scénarios, d'hyperparamétrages, sans et avec lemmatisation, et via des pourcentages des mots les plus rares.

Suite aux résultats positifs de notre discussion, nous pouvons conclure que notre jeu de données répondait bien aux critères du *Fitness for use*, et que notre prototype d'algorithme d'apprentissage supervisé en ML atteignait un taux de précision suffisant pour délivrer des résultats réalistes et utilisables.

4 Recommandations

La question a directement été posée dès la définition de la question de recherche : comment procérons-nous après avoir détecté ces biais ? Nous allons voir comment nous pouvons utiliser le prototype dans la réalité, avec ses atouts et ses limites.

4.1 Atouts du prototype dans la réalité

Comme déjà montré dans la discussion de l'étude de cas, il existe déjà des outils pour guider la rédaction d'offres d'emploi tout en limitant les biais : **biais de genre avec xx, biais de niveau de vocabulaire avec xx,**

Tout comme tout processus d'amélioration continue de la qualité des données, il existe plusieurs manières pour éviter l'ajout de biais (méthodes préventives) et pour corriger les biais une fois qu'ils sont détectés (méthodes curatives).

De manière préventive, il serait intéressant de pouvoir prédire les biais dans les offres d'emploi dès leur encodage, avec des détections automatiques directement actives dans les outils de gestion des contenus (CMS pour les sites web, ATS pour les offres d'emploi). Tout comme nous disposons aujourd'hui d'aide dans la détection de fautes d'orthographes et de grammaire, les personnes qui rédigent les offres d'emploi pourraient ainsi directement corriger leur rédaction. A partir d'un prototype tel que celui développé pour notre état de l'art, nous pourrions ainsi ajouter des contraintes d'intégrité adaptées dès l'encodage.

De manière curative, des formulaires en ligne de signalement au bas des pages web pourraient enrichir ce prototype (cette fonctionnalité de feedback existe sur Dorifor).

4.2 Limites du prototype dans la réalité

De manière pragmatique, nous avons pu observer que cette approche technosolutionniste restait limitée.

En effet, chaque indicateur de vérification demande l'intervention de personnel spécialisé. L'état de l'art nous a démontré l'importance de l'investissement nécessaire

pour travailler avec des données du niveau *Business Intelligence*. Les offres d'emploi n'étant pas des documents générés à la volée, nous pouvons nous demander si la sensibilisation et la formation des spécialistes du recrutement aux biais de rédaction aurait un meilleur rendement.

Par ailleurs, le poids social, notamment encouragé par les syndicats, permet de rappeler aux organisations qu'un contrôle qualité élémentaire des offres d'emploi possède une dimension politique cruciale.

Enfin, rappelons que notre approche occidentale rencontre des résultats qui pourraient être différents dans d'autres cultures.

5 Conclusion

Dans ce mémoire, nous avons tenté d'utiliser l'IA pour débiaiser les offres d'emploi.

5.1 Hypothèse

Nous avons basé notre hypothèse sur le principe énoncé par la startup Interskillar : plus une offre d'emploi reçoit des profils de candidature diversifiés, moins elle est biaisée.

5.2 Traitement de la question de recherche

Dans la partie état de l'art, nous avons défini les concepts, et identifié les différents biais : les biais intrinsèques aux données, les biais générés par le traitement automatique des données, les biais de l'utilité de l'IA, et les biais des sources utilisées. Les biais intrinsèques aux données regroupaient les biais genrés femmes-hommes, conséquence du vocabulaire utilisé ; les biais de niveau de vocabulaire, surtout liés aux parcours atypiques et aux migrations ; les biais d'annotation, conséquences de biais de représentation. Pour les biais générés par le traitement automatique des données, nous avons d'abord identifié les biais de la qualité des données, et mis en perspective les critères de la qualité des données d'un point de vue du *Big data* d'une part, et les critères de la qualité des données d'un point de vue de la *Business Intelligence* d'autre part. Nous en avons retenu la pertinence de la matrice de faux positifs / faux négatifs pour atteindre le *Fitness for use*. Nous avons ensuite étudié le défi de la pertinence et de la représentativité des données. La pertinence des données nous a ramenées à la matrice de faux positifs / faux négatifs, à laquelle nous avons ajouté les anomalies. Pour traiter les différents types d'anomalies, nous avons compris qu'il existait des méthodes préventives et des méthodes curatives. Pour les offres d'emploi, nous avons retenu l'importance des outils *data quality tools* que sont le *profiling*, la standardisation et le *matching*. La représentativité des données était illustrée par des biais d'éthique, et la mobilité des concepts dans le temps. Nous avons également identifié les biais de l'utilité de

l'IA, les biais culturels et évité de tomber dans le technosolutionnisme. Lors d'une discussion sur ces biais, nous avons identifié qu'il existait de généralement des solutions pour les corriger. Afin de répondre à la question de recherche, nous avons retenu que l'étude de cas devrait identifier les biais de genre via la vérification et la correction de la qualité des données (évaluation de la pertinence des données via la matrice de faux positifs/faux négatifs, avec les *data quality tools*, le référentiel Competent), en utilisant un algorithme d'apprentissage supervisé, et en vérifiant ces résultats prédictifs via la visualisation. Nous avons aussi noté que les résultats finaux devraient être discutés à la lumière des biais de l'IA et des sources.

Dans la partie étude de cas, nous avons d'abord décrit la méthodologie utilisée pour répondre au cadre décrit dans la conclusion de l'état de l'art. Nous avons ensuite mis en place cette méthodologie qui nous a amenées à collecter, explorer et nettoyer des données un volume suffisant de fiches formations de Bruxelles Formation et de profils de candidatures femmes-hommes correspondants. Nous avons enfin réalisé un prototype d'algorithme d'apprentissage supervisé utilisant la technique du *Machine Learning* et le modèle de classification *Randomforest*. Nous avons d'abord fixé des paramètres minimaux afin de valider le prototype développé. Puis nous avons réalisé plusieurs scénarios d'optimisation en vérifiant à chaque étape les meilleures options grâce à des techniques de visualisation des indicateurs. Nous avons notamment utilisé : la validation croisée et la précision avec un tableau comparatif des résultats, le taux d'erreur OOB et le coefficient de Gini visualisés dans des graphiques en courbe puis en barre, les vrais positifs et les vrais négatifs avec la matrice de confusion. Nous avons testé des scénarios portant sur le nombre d'arbres, les permutations d'arbres, la lemmatisation, et le pourcentage de mots les plus rares. Nous en avons conclu que le scénario offrant les meilleurs résultats était celui combinant 100 arbres, 1,1 permutations, 5% des mots les plus rares. Ensuite, nous avons discuté de ces résultats en les comparant à des références extérieures, à savoir ceux d'un thésaurus de référence sur les termes plutôt associés aux femmes et sur les termes plutôt associés aux hommes, et ceux d'un schéma sur le processus ML, en proposant de l'enrichir avec des indica-

teurs de qualité à chaque étape. Nous en avons conclu que nos résultats rencontraient pleinement ces références extérieures, et répondaient ainsi à la question de recherche.

Enfin, nous avons formulé des recommandations d'utilisation du prototype dans la réalité, avec ses atouts et ses limites. Nous avons proposé de l'intégrer aux outils d'encodage des offres d'emploi, tout en notifiant des coûts humains de la maintenance d'un tel outil; nous avons évoqué la piste peut-être plus rentable de la sensibilisation et la formation des spécialistes du recrutement aux biais de rédaction des offres d'emploi.

5.3 Limites de la question de recherche

Pour répondre pleinement au principe de la startup Interskillar, nous aurions pu étendre la question de recherche à la comparaison de la diversité des profils des candidatures à la diversité des personnes résidant dans la zone géographique concernée par l'offre d'emploi. Ainsi, une offre d'emploi débiaisée pour un poste travaillant sur un contexte bruxellois devrait recevoir des candidatures de profils diversifiés proportionnellement à la diversité bruxelloise. Ceci aurait pu être réalisé grâce aux statistiques publiées par l'Office belge de statistique Statbel : la diversité bruxelloise est connue par sexe, par niveau d'éducation, par origine... On pourrait ainsi estimer le coefficient de discrimination de chaque offre d'emploi.

5.4 Difficultés rencontrées

Lors du travail sur ce mémoire, nous avons principalement rencontré deux difficultés.

La première, d'ordre conceptuelle, s'est présentée lorsque nous avons dû associer les vocabulaires de qualité des données et le vocabulaire statistique et informatique du ML pour construire la méthodologie. **Le temps de compréhension pour réaliser cette partie a pris un temps de compréhension certain.**

La seconde, d'ordre pragmatique, est apparue en fin de parcours, lors du traitement automatique des données selon les différents scénarios. Ces scénarios ont d'abord pris

du temps à être programmés, et encore plus de temps à être calculés : générer des résultats a pris des journées et des nuits entières. Il était donc crucial de disposer d'un matériel informatique qui pouvait être mobilisé pour cette tâche.

5.5 Perspectives

Avec plus de temps, il aurait été possible de réaliser des prédictions sur base de catégories croisées, par exemple “femmes handicapées”, et réaliser ensuite une comparaison des résultats avec les (nombreuses) études sur les biais de sélection des candidats.

J'ai d'ailleurs appris lors d'un entretien avec une représentante du recrutement de Proximus le 27/03/2023, que son équipe avait réalisé un travail de dégenrage des offres d'emploi (sur base de l'étude de 2011⁹³), et avait par la suite identifié que le processus de recrutement discriminait les femmes lors des entretiens de mise en situation. Un travail est en cours pour lever les biais de ces entretiens.

93. GAUCHER/FRIESEN/KAY : Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality (cf. note 25).

6 Annexes

6.1 Code du prototype

Le code est réparti en 6 fichiers. Nous n'avons pas repris ici le code avec les différents scénarios de paramétrage; ces scénarios sont repris dans la partie Vérifier les résultats prédictifs via la visualisation.

6.1.1 Fichier 1 : 10 nettoyage-paramétrage.R

```
#Chargement des librairies

library(stringr) #nettoyage des chaînes de caractère
library(NLP) #nettoyage du dataframe (nécessaire à la librairie tm)
library(tm) #nettoyage du dataframe
library(SnowballC) #création de matrice

#Import des données

data <- read.csv2("traitement-2024-2.csv", header = TRUE,
                  encoding = "UTF-8")
options(max.print=2) #limitations du nombre des résultats
                     pour les impressions de test
data

#Concaténation des chaînes de caractère dans un datafram
colnames(data)[1] <- "doc_id"
data$text = paste(data$titre, data$chapeau, data$objectif,
                  data$programme, data$pre_requis_liste, data$pre_requis_libre,
                  data$competences_requises, data$remarques)
data$text

data2 = subset(data, select = c(doc_id, text))
data_sexe = subset(data, select = c(doc_id, sexe))
```

```

class(data2) #vérification du format des données

options(max.print=2) #limitations du nombre des résultats pour les

impressions de test

print(data2[1]) #vérification de la transformation en dataframe
print(data2[2]) #vérification de la transformation en dataframe

#Nettoyage des chaînes de caractère concaténées dans le dataframe
##supprimer le HTML et les caractères parasites pour la suite

du traitement

data2$text = str_replace_all(data2$text,"</? [a-z]+/?>"," ")
##supprimer le bruit

data2$text = str_replace_all(data2$text,"\\n"," ") #sauts de ligne
data2$text = str_replace_all(data2$text,"[^ ]*[<>] [^ ]*"," ") #HTML
data2$text = str_replace_all(data2$text,"[^'[:^punct:]]"," ")\\
#ponctuation

data2$text <- iconv(data2$text, from="UTF-8", to="ASCII//TRANSLIT")

#supprimer les accents

print(data2$text)

##Export du texte pour la lemmatisation

print(data2)

write.csv(data2,"data_clean.csv")

#Création d'un corpus de documents à partir du dataframe

(ds <- DataframeSource(data2))

corpus_fiches <- Corpus(ds)

inspect(corpus_fiches)

meta(corpus_fiches) #imprime uniquement les metadata

```

```

#Nettoyage du corpus

corpus_fiches <- tm_map(corpus_fiches, content_transformer(tolower))
corpus_fiches <- tm_map(corpus_fiches, removeWords, stopwords("french"))
corpus_fiches <- tm_map(corpus_fiches, removePunctuation)
corpus_fiches <- tm_map(corpus_fiches, stripWhitespace)

inspect(corpus_fiches)
meta(corpus_fiches)
options(max.print=10) #adaptation des limitations du nombre des
résultats pour les impressions de test
summary(corpus_fiches) #vérification des formats repris par le corpus

#Import des fiches nettoyées dans une matrice termes-documents
dtm <- DocumentTermMatrix(corpus_fiches)
inspect(dtm)

##Paramétrage de la rareté des mots
removeCommonTerms <- function (x, pct)
{
  stopifnot(inherits(x, c("DocumentTermMatrix", "TermDocumentMatrix")),
            is.numeric(pct), pct > 0, pct < 1)
  m <- if (inherits(x, "DocumentTermMatrix"))
    t(x)
  else x
  t <- table(m$i) < m$ncol * (pct)
  termIndex <- as.numeric(names(t[t]))
  if (inherits(x, "DocumentTermMatrix"))
    x[, termIndex]
}

```

```

    else x[termIndex, ]
}

inspect(dtm)

dtm <- removeCommonTerms(dtm, 0.1) #Pourcentage à adapter pour
#l'optimisation

inspect(dtm)

#Transformation de la matrice dense en matrice creuse
dtmM = as.matrix(dtm)

dtmD = as.data.frame.matrix(dtmM)

#Export des données pour stockage
write.table(dtmD, file="dtmd.txt", row.names=FALSE, sep="\t",
quote=FALSE)

print(dtmD[1:1])

class(dtmD)

```

6.1.2 Fichier 2 : 11. algorithme.R

```

#Librairies

library(randomForest)

#Import des données nettoyées et paramétrées
dtmD <- read.table(file="dtmd-010.txt", as.is = TRUE, header = TRUE,
sep = "\t", row.names = NULL)

options(max.print=10) #limitations du nombre des résultats
#pour les impressions de test

```

```

#Ajout à la matrice de la colonne des profils

data <- read.csv2("traitement-2024-2.csv", header = TRUE,
                  encoding = "UTF-8")

colnames(data) [1] <- "doc_id"

data_sexe = subset(data, select = c(doc_id, sexe))

dtmD[1]

dtmD = transform(dtmD, maPrediction = data_sexe$sexe)

dtmD$maPrediction = factor(dtmD$maPrediction)

#is.factor(dtmD$maPrediction)

#dim(dtmD)

#dtmD = dtmD[1:15811,1913:2013]#test sur les 100 derniers \\
enregistrements      => TEST

#Paramétrage du prototype

K = 10 #validation croisée

D0 = dim(dtmD) [1]

accuracies = array(NA, K)

#Lancement de la boucle de Random forest

for (i in 1:K){

  #Segmentation des données pour l'entraînement et le test

  de la prédition

  set.seed(i) #indice de reproductibilité

  indicesTrain = sample(1:D0, size = floor(0.7*D0)) #70% des données

  pour l'entraînement
}

```

```

indicesTest = setdiff(1:D0, indicesTrain) #30% des données\\
pour le test

#Lancement de l'entraînement d'apprentissage et des prédictions
modeleRF01 = randomForest(maPrediction ~ ., data=dtmD[indicesTrain,],
                           ntree = 100, nPerme=11) #Paramètre à optimiser : nombre d'arbres,
                           et nPerme

predTestRF01 = predict(modeleRF01, newdata = dtmD[indicesTest,])

#Evaluation des tests de la prédiction
matConf = table(predTestRF01, dtmD$maPrediction[indicesTest])
accuracies[i] = sum(diag(matConf)) / sum(matConf)
print(c(i, accuracies[i])) # le calibrage augmente à mesure
qu'on séloigne de 0.5

}

#Visualisation des résultats
modeleRF01 #caractéristiques de la matrice dense

class(modeleRF01)
write.csv(accuracies,"accuracies-100-11-010.csv")
save(modeleRF01,file = "modeleRF01-100-11-010.RData")

```

6.1.3 Fichier 3 : 12 visualisation.R

```

#Librairies

library(randomForest)

library(randomForestExplainer)

library(tibble)

library(ggplot2)

```

```

#library(rpart)
#library(rpart.plot)

#Import des résultats du Random Forest
modeleRF01 = get(load("modeleRF01-100-11-005.RData"))

#Visualisation des taux d'erreur OOB et de la matrice de confusion
plot(modeleRF01,main ="Taux d'erreur par nombre d'arbres")
plot(modeleRF01,main ="Taux d'erreur par nombre d'arbres",
      ylim = c(0.2378,0.244))

modeleRF01

#Hiérarchisation des termes discriminants par coefficient de Gini
round(importance(modeleRF01), 2)
class(round(importance(modeleRF01), 2))
impRF01 = importance(modeleRF01)

#impRF01

nomsVar = rownames(impRF01)
impRF01 = impRF01[impRF01[,1]>0,1]
resultatsRF01 = tibble(variable = names(impRF01), coeffGini = impRF01)
#resultatsRF01

resultatsRF01$variable = factor(resultatsRF01$variable, levels =
                                 resultatsRF01$variable[order(resultatsRF01$coeffGini,
                                 decreasing=FALSE)])
resultatsRF02 = resultatsRF01[order(resultatsRF01$coeffGini,
                                 decreasing=FALSE),]

#Visualisation des termes discriminants hiérarchisés par le
#coefficient de Gini

```

```

print(resultatsRF02,n=20)

ggplot(resultatsRF02) +
  geom_col(aes(x = variable, y = coeffGini)) +
  coord_flip()

#Zoom sur le coefficient de Gini

print(length(resultatsRF02[["coeffGini"]]))

resultatsRF03 = subset(resultatsRF02, coeffGini < 0.5)

print(length(resultatsRF03[["coeffGini"]]))

#Zoom sur les résultats de meilleure qualité

resultatsRF04 = subset(resultatsRF03, coeffGini < 0.1)

print(resultatsRF04,n=20)

print(length(resultatsRF04[["coeffGini"]]))

write.csv(resultatsRF04,"resultatsRF04-300-11.csv",row.names = FALSE)

#Visualisation des pourcentages de mots rares

plot_min_depth_distribution(modeleRF01)

#Visualisation de la précision des prédictions

accuracies <- read.csv("accuracies-400-20.csv", header = TRUE,
  encoding = "UTF-8")

colnames(accuracies) <- c("validation_croisee", "accuracy")

accuracies

ggplot(accuracies) +
  geom_col(aes(x=validation_croisee, y = accuracy))

```

6.1.4 Fichier 4 : 20 lemmatisation.R

```

#LIBRARIES

library(udpipe)

```

```

udpipe_download_model("french-gsd")
library(dplyr)

#ETAPE 0

##IMPORTER LES DONNEES

data <- read.csv("data_clean.csv", header = TRUE, encoding = "UTF-8")

##LEMMATISER

texteAnalyse = udpipe(x = data, "french-gsd", trace = FALSE)
#/?\ gros job

View(texteAnalyse)

##EXPORTER LES RESULTATS

tb <- data.frame(texteAnalyse)
write.csv(tb, "lem.csv" )

####pour éviter une énième lémmatisation

texteAnalyse <- read.csv("lem.csv", header = TRUE, sep = ",",
                         quote = "\"", fill = TRUE, comment.char = "",encoding = "UTF-8")
View(texteAnalyse)

rownames(texteAnalyse) <- NULL #réindexer les lignes

##EXPORTER LES RESULTATS NETTOYES

tb <- data.frame(texteAnalyse)
write.csv(tb, "lem_clean.csv" )

##CONSOLIDER POUR LE PROTOTYPE

data_lem <- read.csv("lem_clean.csv",

```

```

            header = TRUE, encoding = "UTF-8")

data_lem2 <- subset(data_lem, select= c(doc_id,lemma))

nb_doc_id <- length(unique(data_lem2[["doc_id"]]))

doc_id_df <- c(unique(data_lem$doc_id))

nb_doc_id

data_lem3 <- data_lem2 %>%
  group_by(doc_id) %>%
  summarise(text = paste(c(lemma), collapse = " "))

data_lem3

##EXPORTER POUR LE PROTOTYPE

tb <- data.frame(data_lem3)

write.csv(tb, "lem2.csv" )

```

6.1.5 Fichier 5 : 21 paramétrage.R

Division du fichier 10 nettoyage-paramétrage.R pour pouvoir y intégrer les résultats de la lemmatisation

```

#Chargement des librairies

library(stringr) #nettoyage des chaînes de caractère

library(NLP) #nettoyage du dataframe (nécessaire à la librairie tm)

library(tm) #nettoyage du dataframe

library(SnowballC) #création de matrice

#Import des données

#sans lemmatизации

data2 <- read.csv("data_clean.csv", header = TRUE, encoding = "UTF-8")

# OU

```

```

# avec lemmatisation

#data2 <- read.csv("lem2.csv", header = TRUE, encoding = "UTF-8")

#Création d'un corpus de documents à partir du dataframe
(ds <- DataframeSource(data2))

corpus_fiches <- Corpus(ds)

#inspect(corpus_fiches)

#meta(corpus_fiches) #imprime uniquement les metadata

#Nettoyage du corpus

corpus_fiches <- tm_map(corpus_fiches, content_transformer(tolower))
corpus_fiches <- tm_map(corpus_fiches, removeWords, stopwords("french"))
corpus_fiches <- tm_map(corpus_fiches, removePunctuation)
corpus_fiches <- tm_map(corpus_fiches, stripWhitespace)

#inspect(corpus_fiches)

#meta(corpus_fiches)

#options(max.print=10) #adaptation des limitations du nombre des
résultats pour les impressions de test

#summary(corpus_fiches) #vérification des formats repris par le corpus

#Import des fiches nettoyées dans une matrice termes-documents
dtm <- DocumentTermMatrix(corpus_fiches)

#inspect(dtm)

##Paramétrage de la rareté des mots

removeCommonTerms <- function (x, pct)
{
  stopifnot(inherits(x, c("DocumentTermMatrix",

```

```

"TermDocumentMatrix")), is.numeric(pct), pct > 0, pct < 1)

m <- if (inherits(x, "DocumentTermMatrix"))

  t(x)

else x

t <- table(m$i) < m$ncol * (pct)

termIndex <- as.numeric(names(t[t]))

if (inherits(x, "DocumentTermMatrix"))

  x[, termIndex]

else x[termIndex, ]

}

#inspect(dtm)

dtm <- removeCommonTerms(dtm, 0.1) #Pourcentage à adapter

pour l'optimisation

#inspect(dtm)

#Transformation de la matrice dense en matrice creuse

dtmM = as.matrix(dtm)

dtmD = as.data.frame.matrix(dtmM)

#Export des données pour stockage

write.table(dtmD, file="dtmd-010.txt", row.names=FALSE, sep="\t",

quote=FALSE)

#print(dtmD[1:1])

#class(dtmD)

```

6.1.6 Fichier 6 : 30 distant reading.R

```

#Chargement des librairies

library(stringr) #nettoyage des chaînes de caractère

```

```

#Import des données

options(max.print=1000)

voc_fh <- read.csv2("voc-fh-2.csv", header = TRUE, encoding = "UTF-8")

class(voc_fh)

results <- read.csv("resultatsRF04-300-11.csv", header = TRUE,
                     encoding = "UTF-8")

print(results)

#nettoyage

##supprimer les accents

voc_fh$X.U.FEFF.vocf <- iconv(voc_fh$X.U.FEFF.vocf, from="UTF-8",
                                 to="ASCII//TRANSLIT")

voc_fh$voch <- iconv(voc_fh$voch, from="UTF-8", to="ASCII//TRANSLIT")

##supprimer les majuscules

voc_fh$X.U.FEFF.vocf <- tolower(voc_fh$X.U.FEFF.vocf)

voc_fh$voch <- tolower(voc_fh$voch)

print(voc_fh)

##renommer les colonnes

colnames(voc_fh)[1] <- "vocf"

colnames(voc_fh)[2] <- "voch"

voc_fh

#comparer les listes de termes

## Fonction pour trouver des mots similaires entre deux colonnes

find_similar_words <- function(col1, col2) {

  similar_words <- character()

  for (word1 in col1) {

```

```

for (word2 in col2) {

  if (str_detect(word2, paste0("^", word1, ".*")))) {

    similar_words <- c(similar_words, word1)

    break
  }
}

## Créer un dataframe avec les résultats

result_df <- data.frame(similar_words = unique(similar_words))

return(result_df)
}

## Appel de la fonction

voc_f <- find_similar_words(voc_fh$vocf, results$variable)

voc_h <- find_similar_words(voc_fh$voch, results$variable)

print(voc_f)

print(voc_h)

```

6.2 Bibliographie

La bibliographie est divisée en sources scientifiques et en sources non-scientifiques.

Sources scientifiques

ALABADLA, Mustafa et al. : Systematic Review of Using Machine Learning in Imputing Missing Values, in : IEEE Access 10 (2022), Conference Name : IEEE Access, p. 44483-44502.

ALBER, Alex : Tutoyer son chef. Entre rapports sociaux et logiques managériales, in : Sociologie du travail 61.1 (2019), Number : 1 Publisher : Association pour le développement de la sociologie du travail, (visité le 13/05/2023).

AUBERT, Marina : Les applications XML dans les systèmes de publication des offres d'emploi, ULB, 2021.

BANERJEE, Debarag Narayan et Sasanka Sekhar CHANDA : AI Failures : A Review of Underlying Issues, in : prépublication Cornell University 2020, URL : <http://arxiv.org/abs/2008.04073> (visité le 05/09/2023).

BATINI, Carlo et Monica SCANNAPIECO : Data Quality - Concepts, Methodologies and Techniques (Data-Centric Systems and Applications), 2006, (visité le 05/03/2023).

BENDER, Emily M. et Batya FRIEDMAN : Data Statements for Natural Language Processing : Toward Mitigating System Bias and Enabling Better Science, in : Transactions of the Association for Computational Linguistics 6 (2018), p. 587-604, (visité le 08/08/2023).

BESSE, Philippe : Déetecter,évaluer les risques des impacts discriminatoires des algorithmes d'IA, mai 2020, (visité le 11/02/2024).

BHARDWAJ, Rishabh, Navonil MAJUMDER et Soujanya PORIA : Investigating Gender Bias in BERT, in : prépublication Cornell University 2020, (visité le 22/07/2023).

- BIRD, Steven, Ewan KLEIN et Edward LOPER : Natural Language Processing with Python, 1^{er} jan. 2009.
- BOEHMKE, Bradley et Brandon GREENWELL : Hands-On Machine Learning with R, 2020, (visité le 23/04/2023).
- BOLUKBASI, Tolga et al. : Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, in : prépublication Cornell University 2016, (visité le 08/08/2023).
- BOYDENS, Isabelle : Les bases de données sont-elles solubles dans le temps?, in : La Recherche. Hors-série 9 (2002), p. 32-34, (visité le 09/03/2024).
- BOYDENS, Isabelle et Gani G. H. HAMITI : Typologie des anomalies, un cadre pour l'action : le cas du machine learning, in : 2022, (visité le 05/03/2023).
- BOYDENS, Isabelle, Gani G. H. HAMITI et Corbesier I. C. ISABELLE CORBESIER : Data Quality Tools : retours d'expérience et nouveautés, in : 7 déc. 2021, (visité le 09/03/2024).
- BRAULT, Nicolas : Le concept de biais en épidémiologie, thèse de doct., Université Sorbonne Paris Cité, 2017, (visité le 08/09/2023).
- BROWN, Nicole M. et al. : In Search of Zora/When Metadata Isn't Enough : Rescuing the Experiences of Black Women Through Statistical Modeling, in : ISSN : 1938-6389, (visité le 28/10/2023).
- CALISKAN, Aylin, Joanna J. BRYSON et Arvind NARAYANAN : Semantics derived automatically from language corpora contain human-like biases, in : Science 356.6334 (2017), Publisher : American Association for the Advancement of Science, p. 183-186, (visité le 05/08/2023).
- CASTIEL, Didier et Pierre-Henri BRÉCHAT : 1. « L'économie de la discrimination » de Gary Becker : une approche au service d'une politique de réduction des handicaps sociaux, in : Solidarités, précarité et handicap social (Hors collection), Rennes 2010, p. 81-92, (visité le 18/02/2024).
- CHEN, Haihua, Jiangping CHEN et Junhua DING : Data Evaluation and Enhancement for Quality Improvement of Machine Learning, in : IEEE Transactions on Reliability 70.2 (2021), Conference Name : IEEE Transactions on Reliability, p. 831-847.

- COUILAUD, Océane : Le Genre Dans l'Education Entrepreneuriale : Une Analyse Exploratoire Inspirée De La Fouille De Texte, 2020, (visité le 22/07/2023).
- FOIDL, Harald et Michael FELDERER : Risk-based data validation in machine learning-based software systems, in : 27 août 2019, p. 13-18.
- FRANCHINA, Luisa et Federico SERGIANI : High Quality Dataset for Machine Learning in the Business Intelligence Domain, in : Yixin BI, Rahul BHATIA et Supriya KAPOOR (éd.) : Intelligent Systems and Applications (Advances in Intelligent Systems and Computing), Cham 2020, p. 391-401.
- GARG, Nikhil et al. : Word embeddings quantify 100 years of gender and ethnic stereotypes, in : Proceedings of the National Academy of Sciences 115.16 (2018), Publisher : Proceedings of the National Academy of Sciences, E3635-E3644, (visité le 05/03/2023).
- HAGENDORFF, Thilo : Linking Human And Machine Behavior : A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning, in : Minds & Machines 31.4 (2021), Publisher : Springer Nature, p. 563-593, (visité le 05/03/2023).
- HAIR, Joseph F. Jr. et Marko SARSTEDT : Data, measurement, and causal inferences in machine learning : opportunities and challenges for marketing, in : Journal of Marketing Theory and Practice 2021, (visité le 25/09/2023).
- HILL, Chelsey et al. : Comparing programming languages for data analytics : Accuracy of estimation in Python and R, in : WIREs Data Mining and Knowledge Discovery, e1531, (visité le 30/03/2024).
- JÖCKEL, Lisa et al. : Conformal Prediction and Uncertainty Wrapper : What Statistical Guarantees Can You Get for Uncertainty Quantification in Machine Learning?, in : Jérémie GUIOCHE et al. (éd.) : Computer Safety, Reliability, and Security. SAFECOMP 2023 Workshops, Cham 2023, p. 314-327.
- JOSHI, Ameet V : Machine Learning and Artificial Intelligence, Cham 2023, (visité le 21/09/2023).
- JUN, Yuan et al. : A survey of visual analytics techniques for machine learning, in : Computational Visual Media 7.1 (2021), (visité le 05/03/2023).

KLÄS, Michael et Anna Maria VOLLMER : Uncertainty in Machine Learning Applications : A Practice-Driven Classification of Uncertainty, in : Barbara GALLINA et al. (éd.) : Computer Safety, Reliability, and Security (Lecture Notes in Computer Science), Cham 2018, p. 431-438.

KNOBLOCH-WESTERWICK, Silvia, Carroll J. GLYNN et Michael HUG : The Matilda Effect in Science Communication : An Experiment on Gender Bias in Publication Quality Perceptions and Collaboration Interest, in : Science Communication 35.5 (2013), Publisher : SAGE Publications Inc, p. 603-625, (visité le 14/05/2023).

LEMMATISATION : Définition de LEMMATISATION, URL : <https://www.cnrtl.fr/definition/lemmatisation> (visité le 27/04/2024).

LEVENDOWSKI, Amanda : How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, in : Washington Law Review 93.2 (2018), p. 579-630, (visité le 05/08/2023).

LOPEZ, M.-L. : Les "Handicapés sociaux" et leur resocialisation : Diversité des pratiques et ambiguïté de leurs effets, t. 2, Company : Persée - Portail des revues scientifiques en SHS Distributor : Persée - Portail des revues scientifiques en SHS Institution : Persée - Portail des revues scientifiques en SHS Label : Persée - Portail des revues scientifiques en SHS Publisher : Editions Médecine et Hygiène, 1978, (visité le 18/02/2024).

MAŚLANKOWSKI, Jacek : Data Quality Issues Concerning Statistical Data Gathering Supported by Big Data Technology, in : Stanislaw KOZIELSKI et al. (éd.) : Beyond Databases, Architectures, and Structures (Communications in Computer and Information Science), Cham 2014, p. 92-101.

MICELI, Milagros, Julian POSADA et Tianling YANG : Studying Up Machine Learning Data : Why Talk About Bias When We Mean Power?, in : arXiv.org; Ithaca 2021, (visité le 25/09/2023).

OLSON, Jack E. : Data Quality : The Accuracy Dimension, Google-Books-ID : x8ahL57V0tcC, 2003.

- OSTRY, Jonathan D. et al. : Economic Gains From Gender Inclusion : New Mechanisms, New Evidence, in : IMF eLibrary 2018, ISBN : 9781484337127, (visité le 11/02/2024).
- PLATANIOS, Emmanouil Antonios et al. : Learning from Imperfect Annotations, 2020, (visité le 05/03/2023).
- REBALA, Gopinath, Ajay RAVI et Sanjay CHURIWALA : An Introduction to Machine Learning, Cham 2019, (visité le 05/03/2023).
- RUGGIA, Simona et Laurent VANNI : DeepFLE : la plateforme pour évaluer le niveau d'un texte selon le CECRL, in : Dialogues et cultures, Dialogues et cultures 2021, Publisher : Fédération internationale des professeurs de français, p. 235-254, (visité le 18/02/2024).
- SLOTA, Stephen C. et al. : Good systems, bad data? : Interpretations of AI hype and failures, in : Proceedings of the Association for Information Science and Technology 57.1 (2020), e275, (visité le 05/09/2023).
- WANG, Richard Y., Mostapha ZIAD et Yang W. LEE : Data Quality, t. 23 (Advances in Database Systems), Boston 2002, (visité le 21/09/2023).

Sources non scientifiques

ABHERVÉ, Michel : Il faut ouvrir le débat sur le "handicap social". Les blogs d'Alternatives Économiques, 10 fév. 2014, (visité le 18/02/2024).

AI Assessment Tool, URL : <https://altai.ai4belgium.be/> (visité le 28/03/2023).

ALSHYMI : offre d'emploi : les avantages qui sont des obligations légales. Twitter, 4 fév. 2023, (visité le 05/03/2023).

BOGANDA, Cyprien : Le scandale des offres bidons : 61 % des offres d'emploi de France Travail sont illégales - L'Humanité, in : 18 jan. 2024, (visité le 19/01/2024).

BONTEMPI, Gianluca : The AI gap : from good accuracy to bad decisions. The regularity gamble, 2022, URL : <https://datascience741.wordpress.com/> (visité le 05/03/2023).

BOYDENS, Isabelle : Data Quality & « back tracking » : depuis les premières expérimentations à la parution d'un Arrêté Royal | Smals Research, (visité le 09/03/2024).

Idem : Open Data et eGovernment, in.

Contrat de gestion 2023-2027 de Bruxelles Formation, Bruxelles Formation, URL : <https://bruxellesformation.brussels/> (visité le 11/02/2024).

Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights, European Union Agency for Fundamental Rights, 2019.

DataframeSource function - RDocumentation, (visité le 01/04/2024).

DIERICKX, Laurence : Apprentissage automatique : les challenges de la qualité des données dans la perspective d'une adéquation aux usages, Réunion du groupe de contact FNRS « Analyse critique et amélioration de la qualité de l'information numérique », 2022.

Données sensibles | Autorité de protection des données, (visité le 23/03/2024).

EU AI Act : first regulation on artificial intelligence | Topics | European Parliament, (visité le 11/02/2024).

FARRET, Pierre-Emmanuel : Utiliser l'intelligence artificielle pour « décoloniser » le langage, in : Global Voices Online, 20 mars 2023, (visité le 21/03/2023).

FEINERER, Ingo : Introduction to the tm Package Text Mining in R, in : <https://cran.r-project.org>.

GOURION, Sophie : Red flags : ces offres d'emploi qui font fuir les candidats (et spécialement les femmes), 2021, (visité le 05/03/2023).

L'anonymisation de données personnelles, URL : <https://www.cnil.fr> (visité le 23/03/2024).

LAWTON, George : What are Machine Learning Models? Types and Examples. Techtarget, (visité le 28/04/2024).

LOVENS, Pierre-François : La Région bruxelloise va se doter d'un pôle d'excellence en intelligence artificielle, in : La Libre.be, 25 jan. 2022, (visité le 05/03/2023).

MANSUROVA, Mariya : Interpreting Random Forests. Medium, 8 oct. 2023, (visité le 21/04/2024).

Référentiel des compétences initiales, Enseignement.be, (visité le 18/02/2024).

REGISTRE NATIONAL DES PERSONNES PHYSIQUES, URL : <https://www.ibz.rrn.fgov.be> (visité le 23/03/2024).

TELEVISIONS, France : Pôle emploi : une majorité d'offres non-fiables?, 2022, URL : <https://www.francetvinfo.fr> (visité le 05/03/2023).