Guilherme Alves da Silva

Brésilien, 32 ans
Résidant en France
http://guilhermealves.eti.br

1 Parcours profissionnel

2023-	Co-founder et développeur FixOut, Nancy
2023-2024	Ingénieur maturation SATT Sayens, Nancy
2022-2023	Attaché Temporaire d'Enseignement et de Recherche (ATER) ENSGSI, Université de Lorraine, Nancy
2021-2022	Attaché Temporaire d'Enseignement et de Recherche (ATER) à mi-temps IDMC, Université de Lorraine, Nancy
2018-2021	Doctorant INRIA (LORIA, CNRS, Université de Lorraine), Nancy
2018	Data scientist 7waves, Sorocaba, Brésil
2017	Trainee Algar Telecom, Uberlândia, Brésil

2 Diplômes

2022 **Doctorat en Informatique**, Université de Lorraine

Titre	Traitement hybride pour l'équité algorithmique *Hybrid processing for algorithmic fairness*	
Directeur	Miguel Couceiro	Pr. Université de Lorraine, LORIA
Co-Directeur	Amedeo Napoli	DR Emérite, CNRS, LORIA
Présidente	Anne Boyer	Pr. Université de Lorraine, LORIA
Rapporteures	Marie-Jeanne Lesot	MCF HDR, Sorbonne Université, LIP6
	Katharina Simbeck	Pr. HTW Berlin
Examinatrice	Fatiha Saïs	Pr. Université Paris Saclay, LISN
Invités	Catuscia Palamidessi	DR Inria Saclay, LIX
	Francis Colas	CR Inria Nancy-Grand Est, LORIA

2017 **Master en Informatique**, Universidade Federal de Uberlândia, Brésil

Mémoire	Évolution de la semi-supervision dans le cadre du clustering en ligne *Evolução da semissupervisão em detecção online de agrupamentos*	
Encadré par	Maria Camila Barioni	Universidade F. de Uberlândia
Examinateurs	Ana Paula Appel	IBM Research
	Bruno Travençolo	Universidade F. de Uberlândia

2015 **Licence en Informatique**, Universidade Federal de Uberlândia, Brésil

2010 Ensino médio (eq. **Baccalauréat**), Escola Estadual Messias Pedreiro, Brésil

3 Recherche

3.1 Thèse

Résumé. Les décisions algorithmiques sont actuellement utilisées quotidiennement. Ces décisions reposent souvent sur des algorithmes d'apprentissage automatique (*machine learning* – ML) qui peuvent produire des modèles complexes et opaques. Des études récentes ont soulevé des problèmes d'iniquité en révélant des résultats discriminatoires produits par les modèles ML contre des minorités et des groupes non privilégiés. Comme les modèles ML sont capables d'amplifier la discrimination en raison de résultats injustes, cela révèle la nécessité d'approches qui découvrent et suppriment les biais inattendues.

L'évaluation de l'équité et l'atténuation de l'iniquité sont les deux tâches principales qui ont motivé la croissance du domaine de recherche en *équité algorithmique* (*algorithmic fairness*). Plusieurs notions utilisées pour évaluer l'équité se concentrent sur les résultats et sont liées à des caractéristiques sensibles (par exemple, le sexe et l'éthinicité) par des mesures statistiques. Bien que ces notions aient une sémantique distincte, l'utilisation de ces définitions de l'équité est critiquée pour sa compréhension réductrice de l'équité, dont le but est essentiellement de mettre en œuvre des rapports d'acceptation/non-acceptation, ignorant d'autres perspectives sur l'inégalité et l'impact sociétal. *Process fairness* (équité des procédures) est au contraire une notion d'équité subjective, centrée sur le processus qui conduit aux résultats.

Pour atténuer ou supprimer l'iniquité, les approches appliquent généralement des interventions en matière d'équité selon des étapes spécifiques. Elles modifient généralement soit (1) les données avant l'apprentissage, soit (2) la fonction d'optimisation, soit (3) les sorties des algorithmes afin d'obtenir des résultats plus équitables. Récemment, les recherches sur l'équité algorithmique ont été consacrées à l'exploration de combinaisons de différentes interventions en matière d'équité, ce qui est désigné dans cette thèse par le *traitement hybride de l'équité*. Une fois que nous essayons d'atténuer l'iniquité, une tension entre l'équité et la performance apparaît, connue comme le compromis équité/précision.

Cette thèse se concentre sur le problème du compromis équité/précision, puisque nous sommes intéressés par la réduction des biais inattendues sans compromettre les performances de classification. Nous proposons donc des méthodes basées sur les ensembles pour trouver un bon compromis entre l'équité et la performance de classification des modèles ML, en particulier les modèles de classification binaire. De plus, ces méthodes produisent des classifieurs d'ensemble grâce à une combinaison d'interventions sur l'équité, ce qui caractérise les approches de traitement hybride de l'équité.

Nous proposons FixOut (*FaIrness through eXplanations and feature dropOut*), un framework centré sur l'humain et agnostique vis-à-vis des modèles qui améliore l'équité des processus sans compromettre les performances de classification. Il reçoit en entrée un classificateur pré-entraîné (modèle original), un ensemble de données, un ensemble de caractéristiques sensibles et une méthode d'explication,

et il produit un nouveau classificateur qui dépend moins des caractéristiques sensibles. Pour évaluer la dépendance d'un modèle pré-entraîné aux caractéristiques sensibles, FixOut utilise des explications pour estimer la contribution des caractéristiques aux résultats des modèles. S'il s'avère que les caractéristiques sensibles contribuent globalement aux résultats des modèles, alors le modèle est considéré comme injuste. Dans ce cas, il construit un groupe de classificateurs plus justes qui sont ensuite agrégés pour obtenir un classificateur d'ensemble. Nous montrons l'adaptabilité de FixOut sur différentes combinaisons de méthodes d'explication et d'approches d'échantillonnage. Nous évaluons également l'efficacité de FixOut par rapport au *process fairness* mais aussi en utilisant des notions d'équité standard bien connues disponibles dans la littérature. De plus, nous proposons plusieurs améliorations telles que l'automatisation du choix des paramètres de FixOut et l'extension de FixOut à d'autres types de données.

Mots-clés

Évaluation de l'équité · atténuation de l'iniquité · explications · compromis équité/exactitude.

3.2 Publications et communications

- Publications sur HAL : cv.hal.science/guilherme-alves
- Publications sur dblp : dblp.org/pid/128/3513.html
- Publications sur Google Scholar : scholar.google.com/citations?user=JRyHYPsAAAAJ

Journaux

1. G. Alves. F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, S. Zhioua. Survey on Fairness Notions and Related Tensions. EURO Journal on Decision Processes, v. 11, article 100033, 2023. [doi]
2. C.Z. Felício, K.V.R. Paixão, G. Alves, S. Amo, P. Preux. Exploiting Social Information in Pairwise Preference Recommender System. Journal of Information and Data Management (JIDM), v. 7, p. 99-115, 2016. [pdf]
3. S. Amo, M.L.P. Bueno, G. Alves, N.F. Silva. Mining User Contextual Preferences. Journal of Information and Data Management (JIDM), v. 4, p. 37-46, 2013. [pdf]

Communications internationales avec actes

1. G. Alves, M. Amblard, F. Bernier, M. Couceiro, A. Napoli Reducing Unintended Bias of ML Models on Tabular and Textual Data. IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), 2021, Porto, p. 1-10. [doi]
2. G. Alves, V. Bhargava, M. Couceiro, A. Napoli. Making ML models fairer through explanations: the case of LimeOut. Analysis of Images, Social Networks and Texts (AIST), 2020, Moscou, p. 3-18. [doi]
3. C.Z. Felício, C.M.M. Almeida, G. Alves, F.S.F. Pereira, K.V.R. Paixão, S. Amo. VP-Rec: A Hybrid Image Recommender Using Visual Perception Network. IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), 2016, San José, p. 70-77. [doi]

4. C.Z. Felício, C.M.M. Almeida, G. Alves, F.S.F. Pereira, K.V.R. Paixão, S. Amo. Visual Perception Similarities to Improve the Quality of User Cold Start Recommendations. 29th Canadian Conference on Artificial Intelligence. 2016, Victoria, p. 96-101. [doi]

5. S. Amo, M.L.P. Bueno, G. Alves, N.F. Silva. CPrefMiner: An Algorithm for Mining User Contextual Preferences Based on Bayesian Networks. IEEE 24th International Conference on Tools with Artificial Intelligence (ICTAI), 2012, Athènes, p. 114-121. [doi]

Communications nationales (France) avec actes

1. G. Alves, M. Couceiro, A. Napoli. Sélection de mesures de similarité pour la classification de données catégorielles. 20ème Conférence Extraction et Gestion de Connaissances (EGC), 2020, Bruxelles, p. 325-332. [pdf]

Communications nationales (Brésil) avec actes

1. G. Alves, M.C.N. Barioni, E. Faria. A Framework for Online Clustering Based on Evolving Semi-Supervision. 32nd Brazilian Symposium on Databases (SBBD), 2017, Uberlândia, p. 16-27. [doi]

2. C.Z. Felício, K.V.R. Paixão, G. Alves, S. Amo. Social PrefRec framework: leveraging recommender systems based on social information. 3rd Symposium on Knowledge Discovery, Mining, and Learning (KDMiLe), 2015, Petrópolis, p. 66-73. [pdf]

Autres communications

1. G. Alves, M. Couceiro, A. Napoli. Towards a Constrained Clustering Algorithm Selection. Société Francophone de Classification (SFC) Actes des 26èmes Rencontres, Nancy, 2019, p. 99-104. [pdf]

Pré-publications

1. G. Alves, F. Bernier, V. Bhargava, M. Couceiro, A. Napoli. Studying the impact of feature importance and weighted aggregation in tackling process fairness.

Présentations lors de rencontres et séminaires

1. Octobre 2023, Journées du Centre Internet et Societé. Paris.
   Algorithmic fairness.

2. Janvier 2022, Séminaire CaféTAL. Nancy.
   Reducing unintended bias of ML models on tabular and textual data.

3. Avril 2021, Journée Perspectives et Défis de l'IA 2021. Online.
   Making ML Models fairer through explanations, feature dropout and aggregation.

4. Novembre 2020, Séminaire MALOTEC. Nancy.
   Towards fairer ML models: beyond LimeOut and process fairness.

5. Janvier 2020, Séminaire MALOTEC. Nancy.
   Similarity measure selection for categorical data clustering.

6. Décembre 2018, Journées QCM-BioChem. Nacy.
   A framework for online clustering based on evolving semi-supervision.

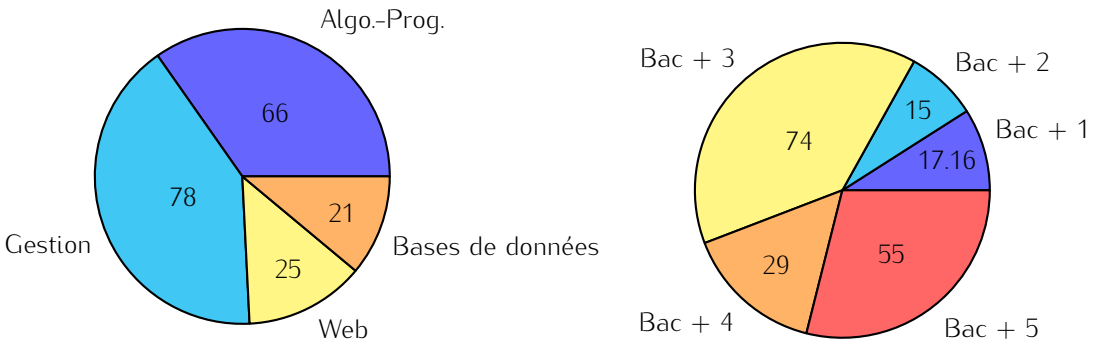
4 Enseignement

Liste des cours dispensés (hors examens et surveillances)

Année	Intitulé	Public	CM	TD	TP	HETD
2020 / 2021 Vacataire	Complexité Algorithmique	L3			13.5	8.91
	Algorithmes pour l'intelligence artificielle	M1 MIAGE		6		6
2021 / 2022 Demi-ATER (Section 27)	Programmation C	L1			26	17.16
	Web avancé	L2			24	15.84
	Complexité Algorithmique	L3			13.5	8.91
	Algorithmes pour l'intelligence artificielle	M1 MIAGE		6		6
	Web sémantique	M1			15	9.9
	Fouille de données	M2 SC, TAL		11.5		11.5
TOTAL (Section 27) : 23h TD - 92h TP soit 84h HETD						
2022 / 2023 ATER (Sections 27-61)	Bases de la gestion de projet	1AI	5	16	18	35.38
	Informatique appliquée pour l'ingénieur	1AI			32	21.12
	Méthodes agiles et gestion de projet Scrum	2AI		6.25	2	7.57
	Project data analyst	3AI		36		36
	Innover pour des Territoires smart	M2 IUVTT		8		8
TOTAL (Sections 27-61) : 5h CM - 101h TD - 159h TP soit 213h HETD						
TOTAL (toutes les sections) : 5h CM - 124h TD - 251h TP soit 297h HETD						

1AI : Première Année Ingeniéur, 2AI : Deuxième Année Ingeniéur, 3AI : Troisième Année Ingeniéur
MIASHS : Mathématiques et Informatique Appliquées aux Sciences Humaines et Sociales
MIAGE : Méthodes Informatiques Appliquées à la Gestion des Entreprises
SCA : Sciences Cognitives — TAL : Traitement Automatique du Langage
IUVTT : Innovation Urbaine pour des Villes & Territoires en Transformation

Répartition des cours dispensés



(a) Répartition par matière

(b) Répartition par niveau

Figure 1: Répartition des heures éq. TD par matière et par niveau.