Institut National de Statistique
et d'Economie Appliquée

Centre des Etudes Doctorales
Sciences, Ingénierie
et Développement Durable

Avis de soutenance de thèse de Doctorat

Madame Lamyae BENHLIMA

Doctorant au laboratoire de recherche

« Méthodes Appliquées en Statistique, Actuariat, Finance et Economie Quantitative »

(MASAFEQ)

Spécialité : ***Statistique Appliquée et Actuariat***

Soutiendra publiquement sa thèse de Doctorat

Le samedi **13 juin 2026 à 10h**

à la salle de conférence de l'INSEA

Intitulé de la thèse

« Etude de l'impact de la pondération des individus dans les méthodes d'apprentissage automatique (Machine Learning) »

Devant le jury composé de :

Président :

Pr. Aomar IBORK, PES, Université Cadi Ayyad - Marrakech

Directeur de thèse :

Pr. Mohammed EL Haj Tirari PES, INSEA-Rabat

Membres du jury :

Pr. Abdelhadi AKHARIF, Université Abdelmalek Essadi FST Tanger, Rapporteur
Pr. Badreddine BENYACOUB, MCH – INSEA, Rapporteur
Pr. Ahmed OUAZZA, MCH – INSEA, Rapporteur
Pr. Mohamed EL MEROUANI, Université Abdelmalek Essadi FST Tétouan
Pr. Tarek ZARI, Université Hassan II Casablanca

INSEA
Institut National de Statistique et d'Économie Appliquée
CEDOC-SIDD

Les Résumés de la thèse (F1)

Réservé à l'administration

N°de thèse : .....

Date : 15 /09/2025.....

Nom : ...**BENHLIMA**.....

Prénom : LAMYAE .....

Résumé

Le développement de modèles prédictifs robustes et généralisables repose sur la qualité structurelle des données d'apprentissage, et notamment sur leur représentativité. Or, cette exigence est souvent mise à mal dans les contextes contemporains d'apprentissage automatique, en raison de biais de sélection, de données massives non contrôlées ou de déséquilibres marqués entre classes. Ces situations compromettent la validité inférentielle des modèles et accentuent les risques de généralisation erronée.

Tandis que la statistique s'attache à garantir la validité inférentielle à partir d'échantillons représentatifs, les modèles d'apprentissage automatique privilégient l'adaptation aux données et la performance prédictive, souvent au détriment de la représentativité. Ce contraste soulève une question centrale : comment intégrer les principes de représentativité statistique dans la construction de modèles d'apprentissage capables de généraliser de manière fiable et équitable ?

Dans ce cadre, cette thèse propose une approche fondée sur la pondération différenciée des individus, inspirée des méthodes d'enquête statistique, afin de renforcer la représentativité des échantillons dans l'apprentissage automatique. Les poids sont dérivés de plans de sondage, puis ajustés par les techniques de redressement. L'objectif est d'évaluer, de manière théorique et empirique, dans quelles conditions cette pondération améliore la généralisabilité des modèles.

Deux axes structurent ce travail. Le premier consiste à intégrer la pondération différenciée des individus dans le processus d'entrainement des modèles d'apprentissage statistique, et à comparer les versions pondérée et non pondérée à l'aide de la technique bootstrap et de tests non paramétriques. L'analyse montre que l'usage de la pondération est particulièrement justifié lorsque les différences entre les deux versions sont significatives, notamment au niveau des coefficients estimés. Le second axe aborde le déséquilibre des classes non comme une difficulté algorithmique isolée, mais comme une forme structurelle de non-représentativité. Dans ce contexte, deux approches sont développées : une première intégrant les poids ajustés dans des modèles confrontés à des données

INSEA *B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
Tél. *: +212 (0)5 37 77 48 59/60* ***Fax*** *: + 212 (0) 537 77 94 57*

déséquilibrées ; une seconde, hybride, combinant pondération ajustée et rééchantillonnage.

L'ensemble des approches a fait l'objet d'une évaluation empirique rigoureuse sur plusieurs jeux de données réelles. Les résultats confirment que, dans les contextes présentant des déséquilibres ou des biais de couverture, la prise en compte de la structure de l'échantillon permet non seulement d'améliorer la représentativité, mais aussi les performances des modèles. Ce travail propose ainsi une voie d'intégration des méthodes statistiques dans l'apprentissage automatique, en plaçant la représentativité au cœur de la robustesse, de l'équité et de la généralisabilité.

Mots clés : Représentativité, Échantillonnage, Pondération des individus, Apprentissage Automatique, Données massives, Déséquilibre des classes

INSEA *B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
Tél. *: +212 (0)5 37 77 48 59/60* ***Fax*** *: + 212 (0) 537 77 94 57*

INSEA
Institut National de Statistique et d'Économie Appliquée
CEDOC-SIDD

Abstract

The development of robust and generalizable predictive models critically depends on the structural quality of training data, and particularly on their representativeness. However, this requirement is often challenged in contemporary machine learning settings due to selection biases, uncontrolled large-scale data sources, or pronounced class imbalances. Such conditions compromise the inferential validity of models and increase the risk of unreliable generalization.

While statistical methods aim to ensure inferential validity through representative sampling designs, machine learning models primarily focus on adapting to data and optimizing predictive performance, often at the expense of representativeness. This methodological tension raises a central question: how can statistical principles of representativeness be incorporated into the construction of learning models capable of generalizing reliably and equitably?

To address this issue, this thesis introduces an approach based on differentiated individual weighting, inspired by survey sampling techniques, to enhance the representativeness of training data in supervised learning. The weights are derived from complex sampling designs and adjusted using calibration-based post-stratification methods. The objective is to assess, both theoretically and empirically, under which conditions such weighting improves model generalizability.

The thesis is organized around two main axes. The first involves integrating differentiated individual weights into the estimation process of statistical learning models and comparing the weighted and unweighted versions using bootstrap procedures and non-parametric statistical tests. The analysis shows that weighting is particularly relevant when the differences between the two versions are statistically significant, especially with respect to estimated parameters. The second axis addresses class imbalance not as an isolated algorithmic challenge, but as a structural form of non-representativeness. In this context, two approaches are proposed: the first incorporates adjusted weights into models trained on imbalanced datasets; the second is a hybrid strategy that combines adjusted weighting with resampling techniques.

INSEA *B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
Tél. *: +212 (0)5 37 77 48 59/60* ***Fax*** *: + 212 (0) 537 77 94 57*

All proposed approaches were rigorously evaluated on multiple real-world datasets. The results confirm that in settings characterized by coverage bias or class imbalance, accounting for the sampling structure not only enhances representativeness, but also improves the generalization performance of predictive models. This work thus contributes to bridging statistical and machine learning methodologies by positioning representativeness as a cornerstone of robustness, fairness, and generalizability in model design.

Keywords: Representativeness, Sampling, Individual Weighting, Machine Learning, Big Data, Class Imbalance

INSEA *B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
Tél. *: +212 (0)5 37 77 48 59/60* ***Fax*** *: + 212 (0) 537 77 94 57*

ملخص

يعتمد بناء نماذج تنبؤية قوية وقابلة للتعميم على جودة البيانات التعليمية، ولا سيّما على مدى تمثيليتها للواقع المدروس. غير أن هذه الخاصية الأساسية تُقوَّض في كثير من الأحيان في سياقات التعلم الآلي المعاصر، نتيجة لتحيّزات في اختيار البيانات، أو لاعتماد بيانات ضخمة غير مُحكمة، أو العوامل إلى إضعاف الصلاحية الاستدلالية لوجود اختلالات واضحة في توزيع الفئات. وتؤدي هذه للنماذج وزيادة احتمالية التعميم الخاطئ.

في حين تركز المقاربات الإحصائية التقليدية على ضمان الصلاحية الاستدلالية انطلاقًا من عينات ممثلة، تُعطي النماذج المستندة إلى التعلم الآلي الأولوية للتكيف مع البيانات وتحقيق الأداء التنبؤي، ولو على حساب التمثيلية. يطرح هذا التباين إشكالية جوهرية: كيف يمكن دمج مبدأ التمثيلية الإحصائية ضمن عملية بناء نماذج تعلم تُعمّم بشكل موثوق ومنصف؟

في هذا الإطار، تقترح هذه الأطروحة مقاربة تقوم على إسناد أوزان تفاضلية للوحدات الإحصائية، مستلهمة من تقنيات المعاينة المعتمدة في البحوث الإحصائية، وذلك بهدف تعزيز تمثيلية العينات في سياق التعلم الآلي. ويتم استخراج هذه الأوزان من مخططات للمعاينة، ثم تُعدّل باستخدام تقنيات المعايرة لضبطها وفقًا لخصائص الهيكل العام للبيانات. وتهدف الدراسة إلى تقييم فاعلية هذه الأوزان نظريًا وتطبيقيًا، والوقوف على مدى إسهامها في تحسين قابلية التعميم للنماذج.

ينقسم هذا العمل إلى محورين متكاملين يتناول المحور الأول إدماج الأوزان التفاضلية ضمن عملية تقدير نماذج التعلم الإحصائي، مع إجراء مقارنة دقيقة بين النسخ الموزونة وغير الموزونة من هذه النماذج باستخدام تقنيات إعادة سحب العينات واختبارات لا معلمية. وقد أظهر التحليل أن اعتماد الأوزان مبرَّر بشكل خاص عندما تكون الفروقات بين النسختين ذات دلالة إحصائية، لا سيما على مستوى المعاملات المقدرة أما المحور الثاني، فيسلّط الضوء على اختلال توازن الفئات، ليس فقط باعتباره تحديًا خوارزميًا، بل بوصفه انعكاسًا لغياب التمثيلية البنيوية داخل البيانات.

وفي هذا السياق، تم تطوير مقاربتين: الأولى تدمج الأوزان المعدّلة في نماذج تتعامل مع بيانات غير متوازنة؛ والثانية هجينة، تجمع بين المعايرة وإعادة المعاينة وقد خضعت هذه المقاربات لتقييم تجريبي دقيق عبر تطبيقها على عدة قواعد بيانات واقعية. وأكدت النتائج أن أخذ بنية العينة بعين الاعتبار، خاصة في السياقات التي تشهد اختلالات أو تحيزات في التغطية، لا يعزز فقط التمثيلية، بل يساهم أيضًا في تحسين الأداء التعميمي للنماذج. وعليه، تقترح هذه الأطروحة مسارًا لدمج المبادئ الإحصائية ضمن منظومة التعلم الآلي، من خلال اعتبار التمثيلية ركيزة أساسية لتحقيق الاستدلال السليم، والعدالة، والموثوقية في التنبؤ.

INSEA *B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
Tél. *: +212 (0)5 37 77 48 59/60* ***Fax*** *: + 212 (0) 537 77 94 57*

الكلمات المفتاحية: التمثيلية الإحصائية، المعاينة، ترجيح المشاهدات، التعلّم الآلي، البيانات الضخمة، عدم توازن الفئات

INSEA *B.P. 6217, Rabat-Instituts, Rabat 10100, Maroc*
Tél. *: +212 (0)5 37 77 48 59/60* ***Fax*** *: + 212 (0) 537 77 94 57*