جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes

Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THÈSE DE DOCTORAT

Madame Yousra CHERIF

Soutiendra publiquement sa thèse de Doctorat en Informatique

Le 18 Juillet 2026 à 11h00 à l'Amphi 4 à l'ENSIAS

Intitulé de la thèse

Enhancing the performance and Interpretability of Species Distribution Models using Feature Selection Techniques

Président :

Pr. Mohammed ESSAAIDI, PES, ENSIAS, Université Mohammed V

Directeur de thèse :

Pr. Ali IDRI, PES, ENSIAS, Université Mohammed V

Rapporteurs :

Pr. Azeddine ZAHI, PES, FST-Fes, Université Sidi Mohammed Ben Abdellah

Pr. Saloua BENSIALI, MCH, Département Statistique et Informatique, Institut Agronomique et Vétérinaire Hassan II

Pr. Aziza CHAKIR, MCH, Faculté des Sciences juridiques, économiques et sociales, Université Hassan II

Examinateur(s) :

Pr. Laila CHEIKHI, PES, ENSIAS, Université Mohammed V

جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

Résumé :

Cette thèse examine le rôle de la sélection des caractéristiques dans l'amélioration des performances et de l'interprétabilité des modèles de distribution des espèces (SDM) basés sur l'apprentissage automatique. Les données environnementales à haute dimension et les algorithmes d'apprentissage opaques entravent souvent la compréhension écologique et la précision des prévisions, ce qui fait de la sélection des caractéristiques une étape cruciale dans l'identification de prédicteurs informatifs et écologiquement significatifs. L'étude évalue d'abord cinq méthodes de sélection de caractéristiques univariées basées sur des filtres à travers cinq seuils de sélection à l'aide de quatre classificateurs (SVM, LGBM, DT et RF) sur trois espèces de Phoenicurus.

Les résultats montrent que la sélection de 40% et 50% des caractéristiques surpasse systématique- ment l'ensemble complet de caractéristiques et les autres seuils, tandis que SVM affiche les performances les plus faibles et est par conséquent remplacé par XGBoost. Sur la base de ces résultats, des techniques de sélection de caractéristiques univariées et multivariées, individuelles et basées sur des ensembles, y compris des ensembles basés sur des wrappers, sont évaluées à l'aide des seuils optimaux sur sept es- pèces d'oiseaux des genres Oenanthe et Phoenicurus. En parallèle, quatre techniques d'interprétabilité sont appliquées pour analyser le comportement du modèle et évaluer le compromis entre les performances prédictives et l'interprétabilité.

L'évaluation des modèles repose sur la précision, le score F1 et le coefficient Kappa de Cohen, la méthode Borda Count permettant un classement multicritère et le test SK identifiant les différences statistiquement significatives entre les modèles. Dans l'ensemble, la thèse démontre que la suppression des prédicteurs redondants et non pertinents améliore à la fois les performances et l'interprétabilité du SDM. Les caractéristiques sélectionnées correspondent systématiquement aux principaux facteurs éco- logiques tels que la température, les précipitations et les caractéristiques de l'habitat, ce qui confirme la validité écologique des modèles. Ces résultats ont une valeur pratique pour la planification de la conservation, l'évaluation de l'adéquation des habitats et la compréhension des impacts du changement climatique sur la répartition des espèces.

جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

Mots-clés :

Conservation des oiseaux, classification, apprentissage d'ensemble, données environnementales, sélection des caractéristiques, apprentissage automatique, modèles de répartition des es- pèces, interprétabilité globale, interprétabilité locale

Abstract:

This thesis examines the role of feature selection in improving the performance and interpretability of machine learning–based Species Distribution Models (SDMs). High-dimensional environmental data and opaque learning algorithms often hinder ecological insight and predictive accuracy, making feature selection a crucial step in identifying informative and ecologically meaningful predictors. The study first evaluates five univariate filter-based feature selection methods across five selection thresholds using four classifiers (SVM, LGBM, DT, and RF) on three Phoenicurus species. Results show that selecting 40% and 50% of features consistently outperforms both the full feature set and other thresholds, while SVM exhibits the weakest performance and is subsequently replaced by XGBoost. Building on these findings, single and ensemble-based univariate and multivariate feature selection techniques, including wrapper-based ensembles, are assessed using the optimal thresholds across seven bird species from the Oenanthe and Phoenicurus genera. In parallel, four interpretability techniques are applied to analyze model behavior and evaluate the trade-off between predictive performance and interpretability.

Model evaluation relies on accuracy, F1-score, and Cohen's Kappa, with the Borda Count method enabling multi-criteria ranking and the SK test identifying statistically significant differences among models. Overall, the thesis demonstrates that removing redundant and irrelevant predictors enhances both SDM performance and interpretability. The selected features consistently align with key ecological drivers such as temperature, precipitation, and habitat characteristics, supporting the ecological valid- ity of the models. These findings provide practical value for conservation planning, habitat suitability assessment, and understanding the impacts of climate change on species distributions.

Keywords:

Bird conservation, Classification, Ensemble Learning, Environmental Data, Feature Se- lection, Machine Learning, Species Distribution Models, Global interpretability, Local interpretability.