Dynamic Selection of Machine Learning Models For The Ambiental Analysis of Agroecosystems

[1] Tomás Ferraz[0000-1111-2222-3333], [2] Mario González, [1] Gastón Notte, [3] Silvina Niell and [1] Parag Chatterjee [0000-0001-6760-4704]

[1] Department of Biological Engineering, University of the Republic, Paysandu, Uruguay
[2] Department of Mathematics and Statistics, University of the Republic, Salto, Uruguay
[3] Department of Chemistry, University of the Republic, Paysandu, Uruguay
tomasnfh1@gmail.com

Abstract. Artificial Intelligence (AI), and in particular Machine Learning (ML), provides powerful tools for extracting knowledge from complex data and supporting well-founded, evidence-based decision-making. In this context, while Decision Support Systems (DSS) are essential for analyzing large volumes of data, a persistent challenge lies in making them sufficiently flexible and adaptable to a wide range of problems. This work presents the design and implementation of a web-based DSS, focused on biological datasets, which leverages dynamic model selection in ML. The system automatically adapts to different datasets by selecting the most appropriate model based on the structure and quality of the available data. The methodology follows a modular approach, comprising several stages—from database upload and target variable selection to the prediction of new entries for decision support. The work concludes with a case study using a dataset involving chemical and biological aspects related to pesticide residue levels in honeybee hives and their environmental impact.

Keywords: User friendly, adaptable to different databases, exploratory data analysis, data cleaning and preprocessing, dynamic model selection, prediction of new entries.

Selección Dinámica de Modelos de Aprendizaje Automático para el Análisis Ambiental en Agroecosistemas.

Resumen. La inteligencia artificial (IA), y en particular el aprendizaje automático (AA), ofrece herramientas potentes para extraer conocimiento a partir de datos complejos y respaldar decisiones con bases sólidas y fundamentadas. En este contexto, aunque los sistemas de apoyo a la toma de decisiones (SSD) son esenciales para asistir en el análisis de grandes volúmenes de datos, aún persiste el desafío de lograr que sean lo suficientemente flexibles y adaptables a diversos problemas. En este trabajo, se diseñó e implementó un SSD en una página web, basado en la selección dinámica de modelos de AA,

orientado a conjuntos de datos biológicos. El sistema es capaz de adaptarse a diferentes bases de datos mediante la selección automática del modelo más adecuado, según la estructura y calidad de los datos disponibles. La metodología empleada fue modular, compuesta por varias etapas que comienzan con la carga de la base de datos y la selección de la variable a predecir, y finalizan en la predicción de nuevas entradas para la toma de decisiones. El trabajo concluye con un caso de estudio, se trabajó con una base de datos que involucra aspectos químicos y biológicos sobre los niveles de residuo de pesticida sobre colmenas melíferas y su impacto en el ambiente.

Palabras clave: Fácil de utilizar para usuarios principiantes, Adaptable a diferentes bases de datos, Análisis exploratorio de los datos, Limpieza y procesamiento de datos, Selección dinámica de modelos, Predicción de nuevas entradas.

1 Introducción

El presente trabajo busca desarrollar un sistema de apoyo a la toma de decisiones con selección dinámica de modelos de aprendizaje automático sobre datos biológicos. El trabajo es relevante debido a que dado el contexto actual de biología y medicina, la toma de decisiones clínicas basada en datos se ha vuelto crucial. La implementación de sistemas de apoyo a la toma de decisiones (SSD) permite optimizar la eficacia y mejorar los resultados de las decisiones. La propuesta busca abordar la necesidad de un sistema dinámico que se adapte a diversas bases de datos biológicas y seleccione el mejor modelo de aprendizaje automático según las características de los datos, por ejemplo, datos clínicos o agrícolas.

2 Métodos

Para el desarrollo del sistema se adopta una metodología modular, en la cual cada etapa del proceso se implementa como una función autónoma y especializada, responsable de una única tarea específica. Estas funciones se integran dentro de un bloque principal denominado *dyn_model_selection()*, el cual integra las distintas fases del flujo de aprendizaje automático. Dicho bloque principal estructura y ejecuta las etapas necesarias para la toma de decisiones basada en modelos predictivos, garantizando así flexibilidad, escalabilidad y claridad en el diseño del sistema.

2.1 Adquisición de datos

El sistema inicia con la adquisición de datos, permitiendo al usuario cargar un archivo (*.csv* o *.xlsx*) mediante una interfaz gráfica. Esta tarea se realiza con la función *data_adq()*, que lee el archivo seleccionado y muestra sus primeras columnas como verificación visual. Posteriormente, con la función *variable_adq()*, el sistema solicita al usuario que seleccione una variable objetivo (diana) del dataset para ser predicha. Esta variable será clasificada como discreta o continua mediante un árbol de decisión, ya que dicha clasificación determina qué modelos serán utilizados en la fase de entrenamiento.

2.2 Pre-procesamiento de datos

Se desarrolló un módulo de control de calidad en Python, encargado de detectar y corregir anomalías, convertir valores y preparar los datos para su análisis. En primer lugar, las variables categóricas presentes en el conjunto de datos deben ser transformadas a un formato numérico que pueda ser interpretado por los algoritmos de aprendizaje automático. Para ello, se implementa la función auxiliar unique_to_int(), la cual asigna un identificador entero único a cada categoría distinta, garantizando una codificación adecuada sin pérdida de información semántica. Respecto al tratamiento de los valores faltantes, y considerando la naturaleza sensible de los datos utilizados, se adopta un enfoque conservador. En primera instancia, se eliminan todos aquellos registros que contengan datos incompletos. Finalmente, se aplican técnicas de normalización para asegurar una escala adecuada de las variables.

2.3 Análisis exploratorio de datos

Esta sección tiene como objetivo facilitar una comprensión más profunda del conjunto de datos cargado mediante técnicas de análisis exploratorio. Para ello, se presenta un resumen estadístico de las variables, acompañado de herramientas que permiten al usuario interactuar directamente con los datos. En particular, se implementa la función *ManualEDAfun()*, la cual genera un resumen estructurado del dataset, brindando información clave para etapas posteriores del flujo. De manera complementaria, se incorpora la herramienta *D-Tale*, que despliega una interfaz web interactiva desde la cual el usuario puede realizar un análisis exploratorio detallado de forma intuitiva y autónoma.

2.4 Entrenamiento de modelos

En esta etapa se utilizan los resultados anteriores para el entrenamiento de los modelos. En la Figura 1 se pueden apreciar los diferentes componentes que conforman a este bloque que representa a la función model_shake().

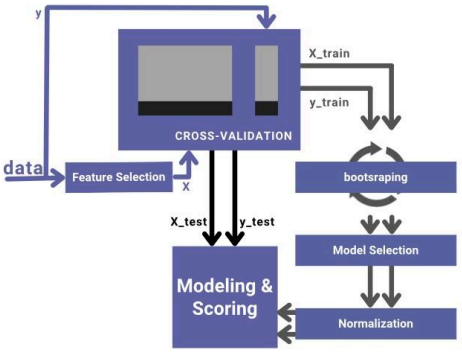

Fig. 1. Representación del flujo de trabajo de la etapa de entrenamiento de los modelos.

La función recibe los datos ya procesados y se encarga de dividir estos en:

- X: el conjunto de variables que se utiliza para el entrenamiento
- y: la variable diana que se busca predecir

En donde X entra en la sección de *Feature Selection*, la cual es la encargada de encontrar las variables con mayor poder predictivo. Posteriormente, estas variables entra en el bloque de *Cross-validation*, en donde se dividen los datos en cinco sub-bloques, donde en cada iteración se toma uno de estos como conjunto de verificación del modelo entrenado y el resto (cuatro sub-bloques del conjunto original) se utiliza para el entrenamiento. En el caso de los datos de entrenamiento entra en la fase de *Bootstrapping*, en esta parte se re-muestrean los datos antes de la normalización y finalmente el entrenamiento. Posteriormente se realiza el

entrenamiento, en este se utilizan conjunto de modelos diferentes dependiendo de la clase de la variable a predecir según la Tabla 1:

Tabla 1. Conjuntos de modelos implementados según el tipo de variable diana a predecir.

Tipo de variable	Discreta	Continua
	RandomForestClassifier	LinearRegression
	KNeighborsClassifier	SupportVectorMachines
	SupportVectorMachines	RandomForestRegressor
	LogisticRegression (booleana)	

Por lo cual, para hacer un sistema dinámico de selección de modelos de aprendizaje automático, se itera sobre los distintos conjuntos de variables (son tres conjuntos, debido a que tenemos tres métodos), y dentro de cada iteración se vuelve a iterar de forma progresiva (donde se tienen cuatro bucles anidados).

2.5 Prueba y validación

En esta etapa se utiliza un sub-bloque para probar el rendimiento de los modelos entrenados. Las medidas de rendimiento se utilizan para evaluar qué tan bien un modelo realiza sus predicciones, algunas de las medidas utilizadas son: *Sensibilidad, Especificidad, F1 Score, ROC* y *AUC*.

2.6 Sistema de soporte de toma de decisiones

Finalmente se enlaza todo para crear el sistema de apoyo a la toma de decisiones, el cual se implementa utilizando una librería que permite crear una página web dinámica en *Python*. Esta aplicación web guía al usuario a lo largo de todo el flujo del sistema, abarcando desde la carga de datos, el preprocesamiento y el entrenamiento de modelos, hasta la selección del modelo óptimo según métricas de rendimiento y la predicción sobre nuevas entradas, facilitando así un entorno completo para la toma de decisiones basada en datos.

3 Resultados

Como caso de estudio, se trabajó con una base de datos que involucra aspectos químicos y biológicos sobre los niveles de residuo de pesticida sobre colmenas melíferas y su impacto en el ambiente. La variable a predecir seleccionada es ¨Salida¨, la cual es detectada correctamente por el sistema como discreta (clases) y los elementos de la clase son cero, uno y dos.

El sistema ejecutó correctamente los distintos módulos sobre este caso de estudio, determinando que no era necesario eliminar columnas por datos faltantes ni realizar conversiones de variables, quedándonos con el 100% de las entradas. Además, generó un análisis exploratorio que permitió examinar la distribución de la variable a predecir, la cual es insesgada. Las variables con mayor poder predictivo fueron presentadas adecuadamente y, en la etapa de entrenamiento, los modelos de ensamble

(particularmente *Random Forest*) demostraron el mejor desempeño (*Sensibilidad y F1*), donde el mejor modelo tiene una Sensibilidad de 0.91 +/- 0.2 . El sistema final cuenta con una interfaz web que permite cargar datos, automatiza el preprocesamiento y la selección de modelos, y ofrece un panel interactivo que facilita la interpretación y la toma de decisiones.

4 Conclusiones

El desarrollo de un sistema de apoyo a la toma de decisiones con selección dinámica de modelos de aprendizaje automático aplicado a datos biológicos representa un avance significativo hacia soluciones más flexibles, automatizadas y adaptables en contextos clínicos, agrícolas y ambientales. A través de una metodología modular e integrada en una interfaz web interactiva, el sistema permite cargar datos, realizar preprocesamiento automático, entrenar múltiples modelos y seleccionar aquel con mejor desempeño según métricas clave. El caso de estudio sobre residuos de pesticidas en colmenas melíferas demuestra la aplicabilidad y robustez del sistema, que no solo identifica correctamente el tipo de variable objetivo, sino que también optimiza el flujo de trabajo desde el análisis exploratorio hasta la predicción. Este enfoque contribuye al fortalecimiento de decisiones basadas en datos en áreas sensibles como la biología y la salud, ofreciendo una herramienta versátil y eficiente para investigadores y profesionales.

References

Niell, S. (2016.). Desarrollo de metodologías de monitoreo químico y biológico y de modelos implementables en un paquete informático con el fin de evaluar riesgos producidos por pesticidas sobre el ambiente y la agricultura. Tesis de doctorado. Universidad de la República (Uruguay). Facultad de Química.

Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Scientific Reports, 13(1). https://doi.org/10.1038/s41598-023-49962-w

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection. ACM Computing Surveys, 50(6), 1–45. https://doi.org/10.1145/3136625

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A brief primer. Behavior Therapy, 51(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

S. Chowdhury and M. P. Schoen, “Research Paper Classification using Supervised Machine Learning Techniques,” 2020 Intermountain Engineering, Technology and Computing (IETC), Orem, UT, USA, 2020, pp. 1-6, doi: 10.1109/IETC47856.2020.9249211