

Clasificación probabilística a partir de datos no deterministas

Autor: Daniel Valladolid Moreno Tutor: Jesús S. Aguilar Ruiz

Grado en Ingeniería Informática en Sistemas de Información · Universidad Pablo de Olavide Defensa: 25 de septiembre de 2026

≈ -22 % NLL

Probabilidades más fiables que un Random Forest clásico, con mejora estadísticamente significativa

= Accuracy

Mismo nivel de acierto: sin diferencias significativas en clasificación

20 datasets

Benchmark UCI, validación cruzada estratificada y contrastes estadísticos formales

EL PROBLEMA

La clasificación supervisada convencional trata cada etiqueta de entrenamiento como una **certeza absoluta**: cada ejemplo pertenece por completo a una única clase. En datos reales esto rara vez es así: hay **clases solapadas**, **ruido de anotación** y **fronteras graduales**. Además, dos modelos con la misma accuracy pueden diferir mucho en la calidad de sus probabilidades, algo crítico cuando la probabilidad predicha condiciona una decisión.

¿Es posible obtener probabilidades más fiables de un modelo basado en árboles de decisión, sin sacrificar su acierto, teniendo en cuenta la incertidumbre presente en los datos?

ENFOQUE

El trabajo propone un **método original** que incorpora la incertidumbre de los datos al **entrenamiento** de modelos basados en árboles de decisión, en lugar de corregir las probabilidades a posteriori. El método se ha **diseñado e implementado desde cero en Python** y se ha comparado de forma homogénea con sus equivalentes clásicos: un árbol de decisión y un Random Forest.



DISEÑO EXPERIMENTAL

- 20 datasets UCI de referencia, binarios y multiclase.
- Validación cruzada **estratificada de 10 folds** con predicciones out-of-fold.
- Pipeline reproducible y **sin fuga de datos**: preprocesado ajustado solo con el entrenamiento.
- Métricas de clasificación: accuracy, F1-macro, G-mean y H-mean.
- Métricas de calidad probabilística: **NLL** y calibración.
- Contrastes estadísticos: **Wilcoxon pareado**, Friedman y Nemenyi.

RESULTADOS PRINCIPALES

- Probabilidades más fiables**: el método reduce la pérdida logarítmica (NLL) en torno a un **22 %** respecto a un Random Forest clásico, con significación estadística.
- Sin coste en acierto**: las métricas de clasificación se mantienen al mismo nivel, sin diferencias significativas.
- Consistencia**: la mejora en NLL se observa en la mayoría de los datasets evaluados, y la capacidad discriminativa (AUC) es equivalente.

CONCLUSIONES Y APLICACIONES

Tener en cuenta la incertidumbre de los datos durante el entrenamiento **no penaliza el acierto y mejora la fiabilidad de las probabilidades**. Es especialmente útil cuando la probabilidad predicha guía decisiones: **evaluación de riesgo**, **priorización de casos** o **diagnóstico asistido**.

Tecnologías: Python · NumPy · pandas · scikit-learn · SciPy · joblib · Matplotlib · Seaborn · LaTeX · **Palabras clave**: clasificación probabilística, incertidumbre, árboles de decisión, Random Forest, calibración.

Investigación en curso, en preparación para su publicación. Este resumen describe el problema, el alcance y los resultados principales del trabajo. Los detalles del método, la implementación y los resultados completos se reservan para su publicación científica y no se distribuyen públicamente.