□

Línea de Trabajo fin de Máster 2024-2025

(Fecha última actualización:04/10/2024)

Máster Universitario en Estadística Aplicada.	
Título	Análisis de diferentes métricas en matrices de confusión para problemas de clasificación con clases no balanceadas
Tipo	INVESTIGACIÓN ☒ ORIENTACIÓN PRÁCTICA ☐
Número de alumnos admitidos	1
Profesor(es)/ email	Pedro Femia (pfemia@ugr.es),Pedro Carmona Sáez (pcarmona@ugr.es)
Descripción	En el ámbito sanitario, existen numerosos problemas de clasificación binaria, como la predicción de qué pacientes desarrollarán un tumor a partir de datos genómicos o cuáles responderán positivamente a un tratamiento. Para abordar estas cuestiones, es posible aplicar técnicas clásicas de clasificación, como k-vecinos más cercanos, random forest o máquinas de soporte vectorial (SVM). En estos análisis, es fundamental utilizar métricas adecuadas para evaluar el rendimiento de los modelos. Entre las métricas más comunes se encuentran la sensibilidad, especificidad, F1 y el coeficiente kappa de Cohen. Un desafío particular en algunas aplicaciones es el desbalanceo de clases, es decir, cuando una clase está sobrerrepresentada en comparación con la otra. En estos casos, algunas métricas, como la precisión, pueden no ser apropiadas. Este Trabajo Fin de Máster (TFM) propone un análisis comparativo de diferentes métricas de evaluación de clasificadores en un contexto de diagnóstico de enfermedades a partir de datos de transcriptoma, donde el número de individuos sanos es significativamente menor que el de los pacientes. El trabajo incluirá una revisión bibliográfica de las principales métricas de evaluación y su aplicación a un caso real de clasificación de datos genómicos.
Objetivos particulares	- Desarrollar tareas de investigación en un entorno multidisciplinar - Conocer algunas de las principales técnicas estadísticas de clasificación y medidas de acuerdo - Implementar metodologías en R para análisis de datos reales
Prerrequisitos y recomendaciones	• Conocimientos avanzado de lenguaje R
Plan de trabajo	- Revisión bibliográfica - Implementación de funciones en R - Aplicación a datos reales
Competencias generales y específicas	CB: 6, 7, 8, 9, 10 CG: 1, 2, 3, 6, 9 CE: 10, 13, 15, 16, 17, 18, 19, 20, 22, 23, 26, 28, 29
Bibliografía	- D Chicco, N Tötsch, G Jurman. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation BioData mining 14, 1-22, 2021 - Wang et al. Cancer Diagnosis by Gene-Environment Interactions via Combination of SMOTE Tomek and Overlapped Group Screening Approaches with Application to Imbalanced TCGA Clinical and Genomic Data. Mathematics, 2024, № 14, p. 2209 - D Chicco, G Jurman The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation BMC genomics 21, 1-13, 2020