

Ensemble Híbrido

En esta sección del trabajo se implementa un método de ensamblaje híbrido.

El ensamblaje híbrido implica la combinación de diversas técnicas de ensamblaje en un único marco de trabajo con el fin de mejorar el rendimiento predictivo del modelo. Este enfoque se fundamenta en la premisa de que la combinación de múltiples modelos puede ofrecer predicciones más precisas y robustas que cualquier modelo individual.

El método de ensamblaje utilizado emplea un modelo lineal para combinar las predicciones de varios modelos, asignándoles pesos según su rendimiento, con el objetivo de obtener una predicción final ponderada.

Selección de los modelos para el ensamblaje

En la estrategia de ensamblaje híbrido que hemos implementado, se han seleccionado los tres mejores modelos de entre todos los evaluados a lo largo de nuestro estudio.

Nos hemos enfocado especialmente en tres aspectos clave para nuestra investigación: la precisión general del modelo (accuracy, en distintos conjuntos de prueba) y la sensibilidad. Estos tres modelos seleccionados como los mejores, en orden de desempeño, son Random Forest, SVM (Support Vector Machine) y XGBoost.

El proceso de selección se ha basado en la capacidad de estos modelos para generalizar patrones relevantes sin caer en el sobreajuste a los datos de entrenamiento. Además, nos hemos centrado en la sensibilidad, ya que es crucial para nuestro trabajo identificar correctamente los casos positivos, en este caso, los morosos.

Una vez finalizado el apartado dedicado al ensamblaje híbrido, se realizará una exposición más detallada sobre la selección de los mejores modelos. Esta sección se centrará en brindar una explicación exhaustiva de los modelos seleccionados y los motivos que respaldan su elección como los mejores dentro del estudio.

Resultados del ensamblaje En la implementación de los modelos en nuestro ensamblaje obtenemos los siguientes pesos para cada componente con su precisión (Accuracy) en el conjunto de entrenamiento:

Intercept: 2.3433 Random Forest: Peso de -1.8752. Support Vector Machine con kernel radial: Peso de 0.3761. Extreme Gradient Boosting Tree: Peso de -3.1035.

La precisión (Accuracy) obtenida en el conjunto de entrenamiento para este modelo ensamblado fue del 72.55 %.

Estos pesos nos indican de la contribución relativa de cada modelo al rendimiento general del ensamblaje, reflejando su influencia en la predicción del resultado final.

Interpretación de los coeficientes

Intercept en 2.34: Indica el valor base del modelo. En el contexto de un ensamblaje de modelos, este podría ser un término de ajuste o el valor inicial alrededor del cual se hacen las predicciones.

Random Forest con -1.8752: El peso negativo sugiere que el modelo de Random Forest está contribuyendo a disminuir la predicción final del ensamblaje. Puede ayudar en la corrección de la sobreestimación, y esto puede deberse a la interacción con el valor elevado del intercept.

Support Vector Machine con kernel radial en 0.3761: Este peso positivo indica que el modelo SVM está contribuyendo a aumentar la predicción final del ensamblaje, aunque su contribución es cercana a 0, lo que sugiere que este modelo tiene el menor impacto en el ensamblaje.

Extreme Gradient Boosting Tree con -3.1035: Este peso negativo, similar al de Random Forest, sugiere que el modelo XGBoost está contribuyendo a disminuir la predicción final del ensamblaje. Dado que este peso es más grande en magnitud que el de Random Forest, el modelo XGBoost tiene un mayor impacto en la corrección de la sobreestimación.

Validación

Ahora evaluaremos el modelo con un conjunto de prueba balanceado para evaluar si hay overfitting:

Cuadro 84: Métricas de validación

	Accuracy	Specificity	Sensitivity
Accuracy	0.7839286	0.940625	0.575

Observamos que los valores de precisión (accuracy) en los conjuntos de datos de entrenamiento y prueba son bastante cercanos, lo que sugiere que no hay evidencia de sobreajuste (overfitting) por parte del modelo.

Asimismo, se observa una disparidad entre la sensibilidad y la especificidad del modelo, donde la última es considerablemente más alta que la sensibilidad. Esto puede indicar un desequilibrio en el rendimiento del modelo, con una mayor capacidad para predecir correctamente los casos negativos en comparación con los casos positivos.

Ahora hacemos la prueba ácida, con datos de prueba desbalanceados:

Cuadro 85: Métricas de validación

	Accuracy	Specificity	Sensitivity
Accuracy	0.9219219	0.9564744	0.525

Vemos que el modelo mejora considerablemente su precisión en el conjunto de datos desbalanceado, lo cual podría deberse al desequilibrio entre sensibilidad y especificidad. En nuestro caso, la sensibilidad mide los verdaderos positivos, es decir, cuando un individuo moroso es clasificado correctamente como tal. Dado que nuestros datos están desbalanceados con muchos más no morosos que morosos, una especificidad tan alta hace que la mayoría de los datos, que son no morosos, se clasifiquen correctamente.

Estos resultados no son óptimos para nuestro estudio, ya que nuestro interés principal radica en clasificar de manera más eficiente a los individuos morosos que a los no morosos. La alta especificidad, si bien puede ser útil para identificar correctamente los casos negativos, no responde de manera satisfactoria a nuestro objetivo prioritario de identificar con precisión a los morosos.

El ensamblaje híbrido, en base a nuestros objetivos, no parece mejorar el rendimiento de nuestros modelos.