Máster en Inteligencia Artificial UMU

PRÁCTICA 1

Machine Learning

Ignacio Ruiz Chicano
Juan Jesús Torralba Mateos
Ana Gil Molina

TAREA DE DESBALANCEO

DATASET: CocheRadar.csv

Distinguir a un coche de cualquier otro vehículo en función de ciertas características que se describen en estos 18 atributos todos numéricos:

- **Compacto**
- **Circularidad**
- **Distancia_circular**
- **Relación_radio**
- **Relación_aspecto_praxis**
- **Relación_aspecto_longitud_máx**
- **Relación_dispersión**
- **Alargamiento**
- **Praxis_rectangular**
- **Longitud_rectangular**
- **Varianza_mayor**
- **Varianza_menor**
- **Rotación_radio**
- **Asimetría_mayor**
- **Asimetría_menor**
- **Curtosis_menor**
- **Curtosis_mayor**
- **Huecos**

OBJETIVO: Clasificar según esos atributos el tipo del vehículo entre un coche y otros vehículos.

Análisis y preprocesamiento

tipo	count
otros	629
coche	217

dtype: int64

Compacto	int64
Circularidad	int64
Distancia_circular	int64
Relación_radio	int64
Relación_aspecto_praxis	int64
Relación_aspecto_longitud_máx	int64
Relación_dispersión	int64
Alargamiento	int64
Praxis_rectangular	int64
Longitud_rectangular	int64
Varianza_mayor	int64
Varianza_menor	int64
Rotación_radio	int64
Asimetría_mayor	int64
Asimetría_menor	int64
Curtosis_menor	int64
Curtosis_mayor	int64
huecos	int64
tipo	object

dtype: object

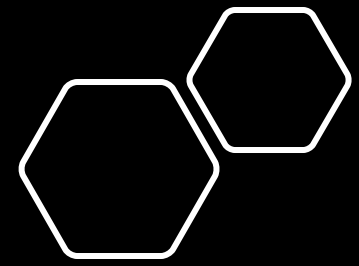
Compacto	0
Circularidad	0
Distancia_circular	0
Relación_radio	0
Relación_aspecto_praxis	0
Relación_aspecto_longitud_máx	0
Relación_dispersión	0
Alargamiento	0
Praxis_rectangular	0
Longitud_rectangular	0
Varianza_mayor	0
Varianza_menor	0
Rotación_radio	0
Asimetría_mayor	0
Asimetría_menor	0
Curtosis_menor	0
Curtosis_mayor	0
huecos	0
tipo	0

dtype: int64

Técnicas, modelos y métricas

- Técnicas de desbalanceo:
  - SMOTE
  - Borderline-SMOTE
  - ADASYN
  - Borderline-SMOTE SVM
- Modelos de validación:
  - SVM
  - Árbol de decisión
  - Red neuronal
  - K-Vecinos
- Métricas:
  - F1-Score
  - Accuracy
  - Recall
  - Precision

Modelo	F1_Train	F1_Test	Accuracy	Precision	Recall
SVM	0.88	0.85	0.78	0.84	0.86
Árbol de decisión	0.87	0.86	0.78	0.79	0.93
Red neuronal	0.99	0.88	0.83	0.86	0.90
K-Vecinos	0.88	0.85	0.78	0.83	0.88
SMOTE: SVM	0.84	0.79	0.74	0.96	0.67
SMOTE: Árbol de decisión	0.94	0.81	0.75	0.90	0.73
SMOTE: Red neuronal	0.94	0.87	0.83	0.94	0.81
SMOTE: K-Vecinos	1.00	0.81	0.76	0.95	0.70
Borderline_SMOTE: SVM	0.84	0.80	0.75	0.97	0.68
Borderline_SMOTE: Árbol de decisión	0.96	0.80	0.75	0.95	0.69
Borderline_SMOTE: Red neuronal	0.99	0.89	0.84	0.95	0.83
Borderline_SMOTE: K-Vecinos	1.00	0.78	0.73	0.95	0.66
ADASYN: SVM	0.83	0.79	0.75	0.96	0.68
ADASYN: Árbol de decisión	0.94	0.79	0.74	0.94	0.68
ADASYN: Red neuronal	0.92	0.82	0.76	0.92	0.73
ADASYN: K-Vecinos	1.00	0.77	0.72	0.93	0.66
Borderline_SMOTE SVM: SVM	0.82	0.78	0.73	0.96	0.65
Borderline_SMOTE SVM: Árbol de decisión	0.97	0.84	0.78	0.88	0.79
Borderline_SMOTE SVM: Red neuronal	0.99	0.86	0.81	0.93	0.80
Borderline_SMOTE SVM: K-Vecinos	1.00	0.76	0.71	0.91	0.65



TAREA DE MULTI-ETIQUETA

DATASET: agua.csv

- Datos sobre calidad del agua en ríos.
- Incluye mediciones de 16 parámetros:
  - Temperatura del agua (temp)
  - Alcalinidad (pH)
  - Conductividad eléctrica (conductividad)
  - Oxígeno disuelto (O2)
  - Saturación de oxígeno (o2sat)
  - Concentración de CO2 (co2)
  - Dureza del agua (dureza)
  - Dinitrógeno (no2)
  - Amoniaco (no3)
  - Amonio (nh4)
  - Ortofosfato (po4)
  - Concentración de cloro (cl)
  - Cantidad de sílice (sio2)
  - Permanganato de Potasio (mno4)
  - Dicromato de potasio (k2cr2o7)
  - Demanda biológica de oxígeno (bod)
- Problema multi-etiqueta: 14 etiquetas binarias que indican la presencia (1) o ausencia (0) de 14 taxones.
- **OBJETIVO:** predecir la presencia de taxones en el agua según las características de la muestra.

DESARROLLO Y SELECCIÓN DEL MEJOR MODELO

MODELO INICIAL: ExtraTreesClassifier

- Ensamble de árboles de decisión entrenados sobre subconjuntos aleatorios de los datos, que selecciona aleatoriamente las divisiones en cada árbol y combina las predicciones de todos ellos, reduciendo el overfitting.
- Parámetros óptimos obtenidos mediante validación cruzada:
  - n_estimators = 70
  - min_samples_split = 7
  - min_samples_leaf = 2
  - max_depth = 6
  - criterion = 'gini'
- Se añadió class_weight = 'balanced' para abordar el problema de desbalanceo en los datos.
- Este modelo muestra un buen equilibrio entre rendimiento y generalización en los datos de entrenamiento y prueba, con métricas como el F1-score, la pérdida de Hamming y el Mean Accuracy destacándose en comparación con otros modelos probados.

DESARROLLO Y SELECCIÓN DEL MEJOR MODELO

MEJORA DEL MODELO: BinaryRelevance

- Técnica de transformación multi-etiqueta que trata cada etiqueta como un problema binario de clasificación independiente, entrenando un clasificador separado para cada una.
- Partiendo del modelo anterior ExtraTreesClassifier, modificando algunos parámetros para controlar el overfitting:
  - n_estimators = 100
  - min_samples_split = 10
  - min_samples_leaf = 5
  - max_depth = 4
  - criterion = 'gini'
  - class_weight = 'balanced
- Aunque ExtraTreesClassifier puede manejar problemas multi-etiqueta directamente, al combinarlo con BinaryRelevance mejora su desempeño gracias al tratamiento independiente de cada etiqueta.

COMPARACIÓN DE MÉTRICAS

	ExtraTreesClassifier		BinaryRelevance	
	Train	Test	Train	Test
Hamming Loss	0.331	0.389	**0.298**	**0.343**
Mean Accuracy	0.669	0.611	**0.702**	**0.657**
F1 Micro	**0.655**	0.598	0.651	**0.600**
F1 Macro	**0.641**	0.576	0.635	**0.582**
F1 Weighted	**0.665**	**0.610**	0.658	0.608
F1 Samples	**0.625**	**0.572**	0.617	0.564

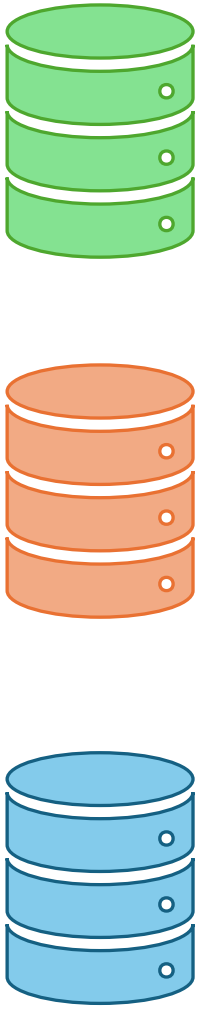
TAREA DE ENSAMBLES

- **DATASET:** TarjetaCredito.csv
- Datos de los usuarios de una entidad bancaria.
- Incluye 16 columnas:
  - D1 nominal
  - D2 numérica
  - D3 numérica
  - D4 nominal
  - D5 nominal
  - D6 nominal
  - D7 nominal
  - D8 numérica
  - D9 nominal
  - D10 nominal
  - D11 numérica
  - D12 nominal
  - D13 nominal
  - D14 numérica
  - D15 numérica
  - Objetivo nominal

- **OBJETIVO:** determinar si se les concede o no una tarjeta de crédito a usuarios de un banco.

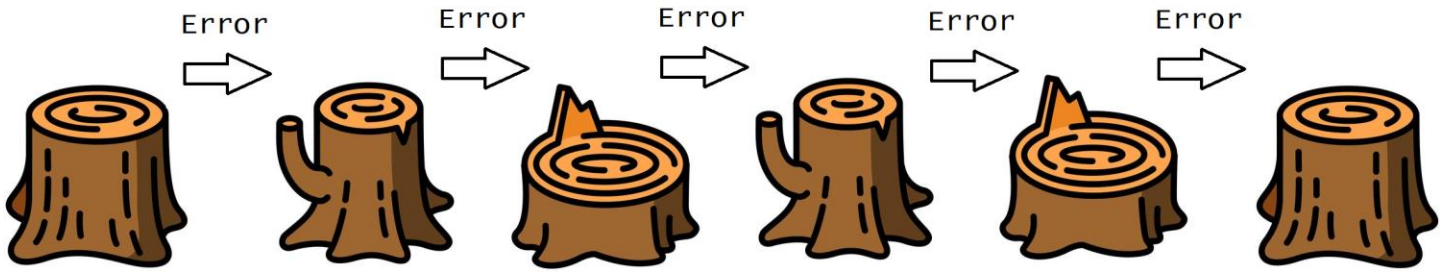
Datasets

- Sin las instancias con NA
- Imputando con la media y la moda
- Imputando usando la técnica de KNN



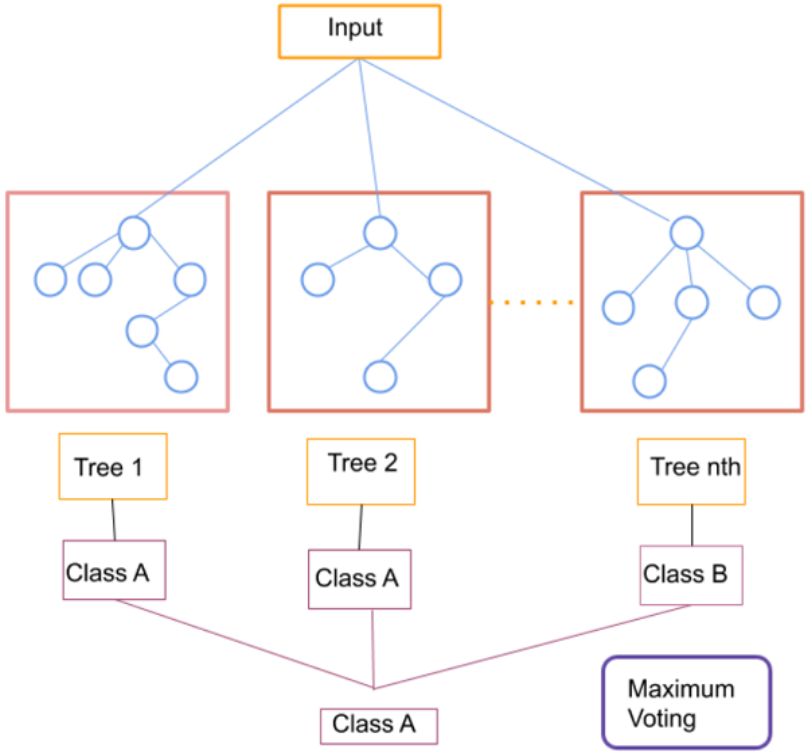
Técnicas empleadas

Adaboost



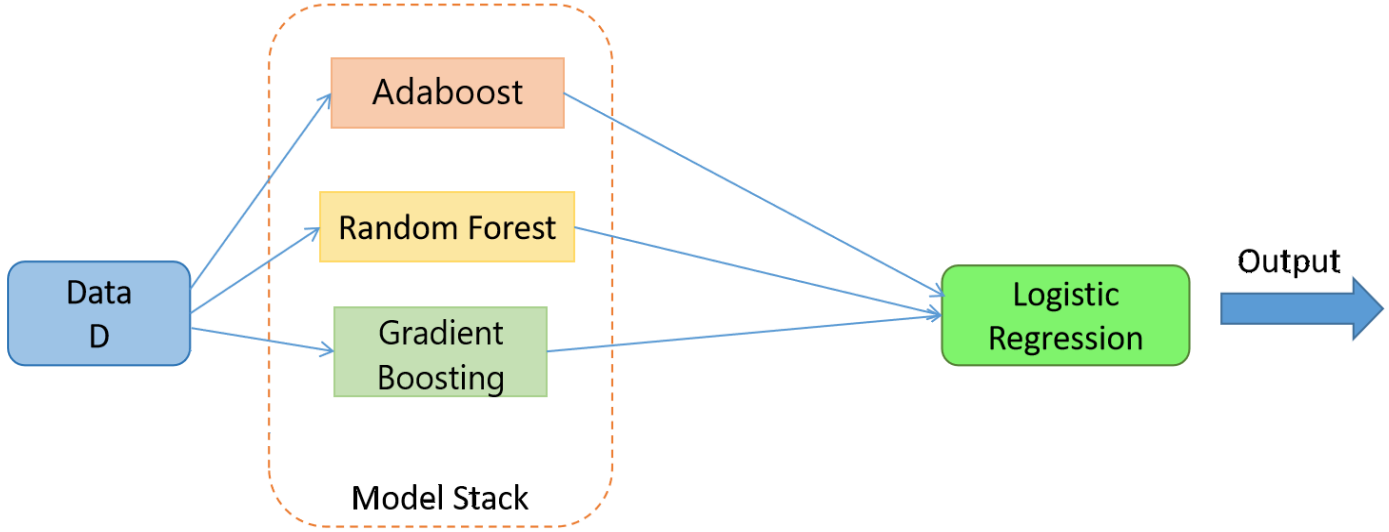
Técnicas empleadas

Random Forest



Técnicas empleadas

Stacking



Comparación de métricas

Modelo	Precisión SI	Precisión NO	Recall SI	Recall NO	F1-Score SI	F1-Score NO	Accuracy Test
ADA-NA	0'83	0'91	0'9	0'85	0'86	0'88	0'87
ADA-MM	0'84	0'86	0'82	0'87	0'83	0'86	0'85
ADA-KNN	0'83	0'85	0'81	0'87	0'82	0'86	0'84
RF-NA	0'91	0'87	0'83	0'93	0'87	0'9	0'89
RF-MM	0'91	0'86	0'81	0'93	0'85	0'89	0'88
RF-KNN	0'91	0'84	0'77	0'93	0'83	0'88	0'86
STA-NA	0'89	0'87	0'83	0'92	0'86	0'89	0'88
STA-MM	0'89	0'84	0'79	0'92	0'84	0'88	0'86
STA-KNN	0'89	0'83	0'77	0'92	0'83	0'88	0'86

Leyenda

ADA --> Adaboost
RF --> Random Forest
STA --> Stacking

NA --> Dataset sin NA
MM --> Dataset con imputación de la media y la moda
KNN --> Dataset con imputación con la técnica KNN

¿Cuál es el modelo óptimo?

Si queremos evitar las pérdidas al banco por impagos del crédito...

Modelo	Precision SI	Precision NO	Recall SI	Recall NO	F1-Score SI	F1-Score NO	Accuracy Test
ADA-NA	0'83	0'91	0'9	0'85	0'86	0'88	0'87
ADA-MM	0'84	0'86	0'82	0'87	0'83	0'86	0'85
ADA-KNN	0'83	0'85	0'81	0'87	0'82	0'86	0'84
RF-NA	**0'91**	0'87	**0'83**	0'93	**0'87**	0'9	0'89
RF-MM	0'91	0'86	0'81	0'93	0'85	0'89	0'88
RF-KNN	0'91	0'84	0'77	0'93	0'83	0'88	0'86
STA-NA	0'89	0'87	0'83	0'92	0'86	0'89	0'88
STA-MM	0'89	0'84	0'79	0'92	0'84	0'88	0'86
STA-KNN	0'89	0'83	0'77	0'92	0'83	0'88	0'86

Leyenda

ADA --> Adaboost
RF --> Random Forest
STA --> Stacking

NA --> Dataset sin NA
MM --> Dataset con imputación de la media y la moda
KNN --> Dataset con imputación con la técnica KNN

Si queremos evitar rechazar a demasiados clientes...

Modelo	Precision SI	Precision NO	Recall SI	Recall NO	F1-Score SI	F1-Score NO	Accuracy Test
ADA-NA	0'83	0'91	**0'9**	0'85	**0'86**	0'88	0'87
ADA-MM	0'84	0'86	0'82	0'87	0'83	0'86	0'85
ADA-KNN	0'83	0'85	0'81	0'87	0'82	0'86	0'84
RF-NA	0'91	0'87	0'83	0'93	0'87	0'9	0'89
RF-MM	0'91	0'86	0'81	0'93	0'85	0'89	0'88
RF-KNN	0'91	0'84	0'77	0'93	0'83	0'88	0'86
STA-NA	0'89	0'87	0'83	0'92	0'86	0'89	0'88
STA-MM	0'89	0'84	0'79	0'92	0'84	0'88	0'86
STA-KNN	0'89	0'83	0'77	0'92	0'83	0'88	0'86

Leyenda

ADA --> Adaboost
RF --> Random Forest
STA --> Stacking

NA --> Dataset sin NA
MM --> Dataset con imputación de la media y la moda
KNN --> Dataset con imputación con la técnica KNN