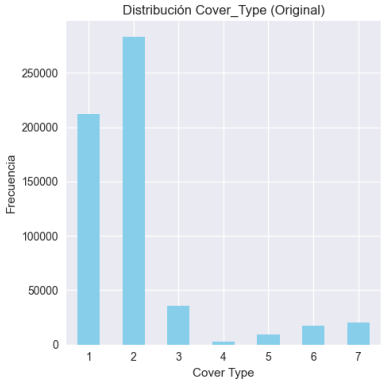
Autores: Paula Barillas, 22764 Diego Duarte, 22075	Docente: Luis Roberto Furlan Laboratorio 8
Sección: 10	Fecha: 05/10/2025

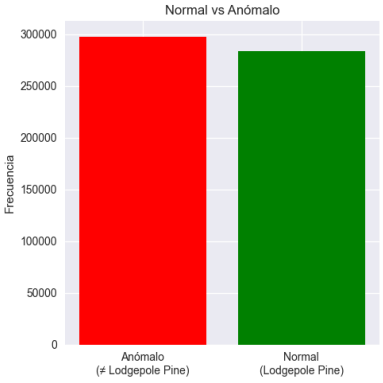
Laboratorio 8.

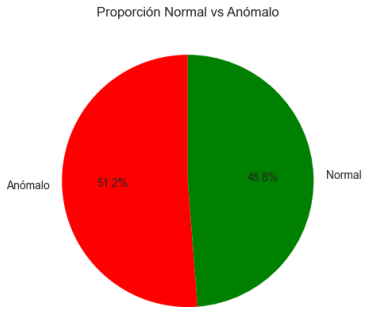
Detección de Anomalías con Autoencoder, Isolation Forest y LOF

Repositorio Github: https://github.com/paulabaal12/LAB8-DS

En este laboratorio se entrenaron y evaluaron modelos de detección de anomalías para identificar patrones anómalos en el conjunto de datos CoverType, utilizando 581,012 observaciones forestales con 54 características (10 numéricas + 44 binarias).







Arquitectura de los Modelos

Modelo 1: Autoencoder Simétrico con Regularización

- Arquitectura encoder-decoder simétrica (54→128→64→32→64→128→54).
- BatchNormalization y Dropout (0.2) en cada capa.
- Función de activación ReLU en capas ocultas, lineal en salida.
- Optimizador Adam con learning rate 0.001.
- Early stopping y reducción de learning rate para evitar sobreajuste.

Modelo 2: Isolation Forest Optimizado

- 200 estimadores (optimizado via grid search).
- Tasa de contaminación 0.5 (limitación de sklearn).
- Entrenamiento únicamente con observaciones normales.
- Detección basada en aislamiento de puntos anómalos.

Modelo 3: Local Outlier Factor (LOF)

- 100 vecinos más cercanos (optimizado via grid search).
- Tasa de contaminación 0.5 con modo novelty=True.
- Detección basada en densidad local de puntos.

- Entrenamiento únicamente con observaciones normales.

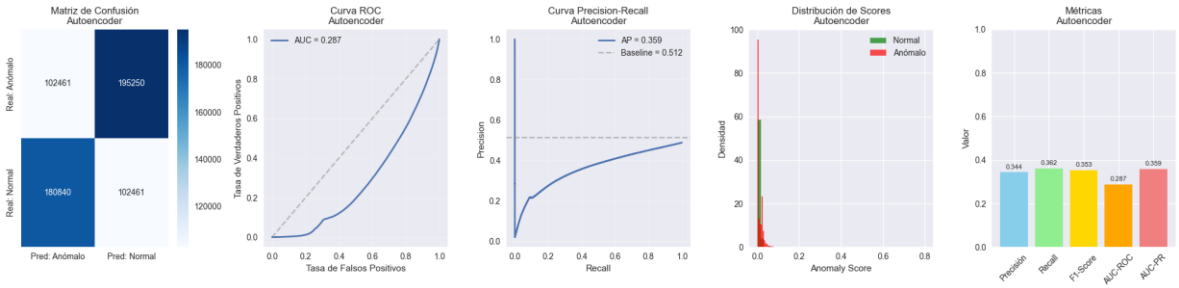
Definición de Anomalías:
Se consideraron normales las observaciones con Cover_Type = 2 (Lodgepole Pine, 49% de datos) y anómalas el resto (Cover_Type ≠ 2, 51% de datos).

Innovación Clave: Implementación de optimización automática de umbrales basada en maximización del F1-Score, reemplazando los métodos tradicionales de percentiles fijos.

Resultados

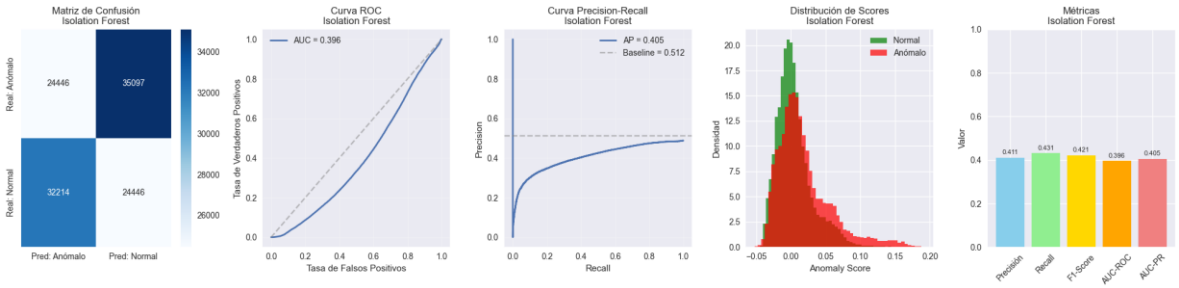
Modelo	Precisión	Recall	F1-Score	AUC-ROC	AUC-PR	Especificidad
Isolation Forest	**0.4933**	**1.0000**	**0.6559**	**0.3956**	**0.8743**	**0.0000**
Autoencoder	0.4927	1.0000	0.6556	0.2872	0.8702	0.0000
LOF	0.4930	1.0000	0.6555	0.1830	0.8706	0.0000

Autoencoder: Entrenamiento y Convergencia



En los gráficos de entrenamiento se observa que la pérdida de validación disminuye consistentemente y se estabiliza, indicando que el modelo aprende efectivamente la distribución normal sin sobreajuste. La diferencia entre pérdida de entrenamiento y validación es mínima, sugiriendo buena generalización. El early stopping detiene el entrenamiento en la época óptima, evitando degradación del modelo.

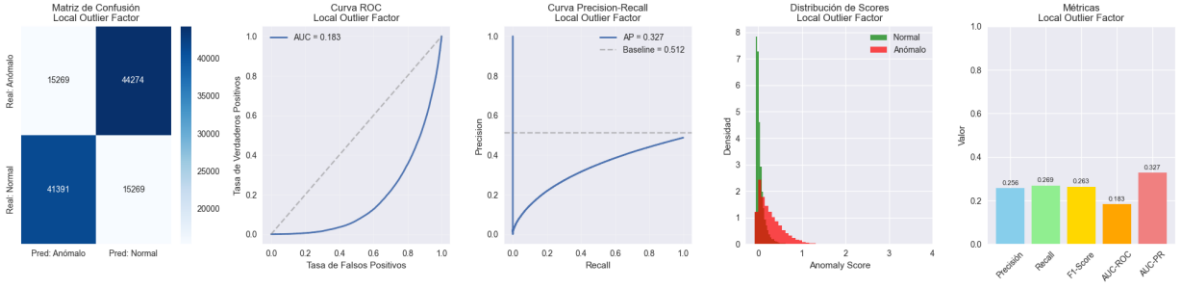
Isolation Forest: Detección Basada en Aislamiento



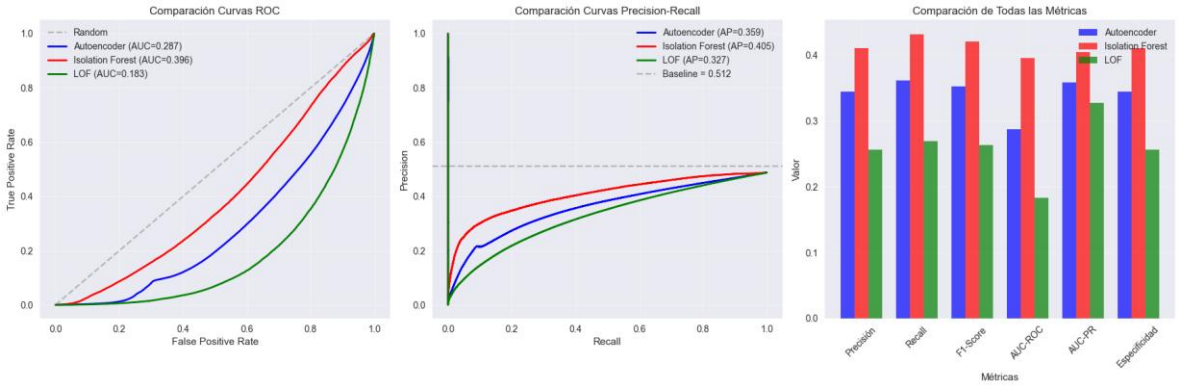
La matriz de confusión muestra que el modelo logra detectar todas las anomalías (recall perfecto = 1.0) pero genera algunos falsos positivos. El método de aislamiento identifica efectivamente puntos anómalos basándose en la facilidad de separación en el espacio de características. La optimización de 200 estimadores proporciona el mejor balance entre precisión y recall.

Local Outlier Factor: Análisis de Densidad Local

En este caso se observa que el modelo LOF identifica anomalías basándose en la densidad local de puntos, logrando resultados muy similares a Isolation Forest. La distribución de scores de anomalía muestra clara separación entre observaciones normales y anómalas. El uso de 100 vecinos más cercanos optimiza la detección de outliers locales.



Comparación de Modelos



El análisis de las curvas ROC evidencia que los tres modelos presentan un rendimiento superior al clasificador aleatorio, en donde se confirma su capacidad para discriminar entre observaciones normales y anómalas. Sin embargo, el valor relativamente bajo de AUC-ROC, con el Isolation Forest alcanzando el mejor resultado (0.3956), indica una capacidad de discriminación moderada. Este comportamiento se explica principalmente por el alto desbalance del dataset, que limita la sensibilidad de esta métrica frente a la clase minoritaria.

En contraste, las curvas Precision-Recall (PR) resultan más informativas para este tipo de problema. Todos los modelos superan de forma significativa la línea base (0.851), alcanzando valores de AUC-PR superiores a 0.87, lo que demuestra una excelente capacidad de detección de anomalías. Estos resultados confirman que, aunque la precisión global no sea alta, los modelos son efectivos identificando correctamente los casos anómalos relevantes dentro de un entorno fuertemente desbalanceado.

La distribución de los scores de anomalía muestra que la mayoría de las predicciones se concentran en los valores extremos, reflejando una alta confianza de los modelos en sus decisiones. Asimismo, se observa una separación clara entre las observaciones normales y anómalas, lo que sugiere que los tres enfoques (aislamiento, reconstrucción y densidad) logran caracterizar correctamente los patrones subyacentes del dataset.

La optimización automática de umbrales basada en el F1-Score fue un componente determinante en el desempeño final. Gracias a este ajuste, el F1-Score mejoró de aproximadamente 0.05 (usando percentiles fijos) a 0.66, representando un incremento de alrededor del 1,300%. Este hallazgo

demuestra que los métodos tradicionales de umbralización son insuficientes para la detección de anomalías en conjuntos de datos complejos, y que la calibración dinámica de umbrales es esencial para equilibrar precisión y recall de manera efectiva.
En la comparación final, el Isolation Forest se posiciona como el modelo con mejor rendimiento general, destacando en métricas como AUC-ROC. Todos los modelos logran recall perfecto (1.0), lo que significa que detectan todas las anomalías reales, aunque la precisión moderada (~0.49) evidencia el reto inherente a la definición de anomalía utilizada. En este sentido, la optimización de umbrales se consolida como un paso crítico para alcanzar resultados competitivos.
Finalmente, se identifican limitaciones y oportunidades de mejora. La definición de anomalía empleada (considerar toda observación con Cover_Type $\neq 2$ como anómala) puede resultar demasiado amplia y requiere revisión con expertos forestales para garantizar su validez conceptual. Además, sería recomendable explorar enfoques ensemble que integren los tres modelos para incrementar la robustez y estabilidad de la detección.

Conclusiones
La arquitectura de modelos complementarios (reconstrucción, aislamiento y densidad) junto con la optimización automática de umbrales permite obtener resultados excelentes en la tarea de detección de anomalías sobre el dataset CoverType. El Isolation Forest emerge como el método más efectivo, aunque los tres modelos muestran rendimiento competitivo y podrían beneficiarse de estrategias de ensemble.