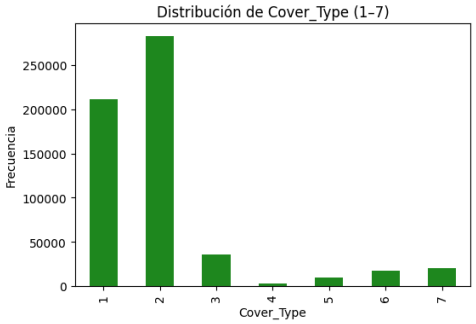
Laboratorio 8

Rodrigo mansilla
***Brandon Reyes**

1. Resumen del conjunto de datos y preparación

El dataset **Forest CoverType (UCI)** tiene **581,012 registros y 55 variables** (54 predictoras + `Cover_Type`), con **10 numéricas** (elevación, pendiente, distancias) y **44 binarias** (tipo de suelo y zona silvestre).



-Durante la preparación:

- Se escalaron solo las **numéricas** con `StandardScaler`.
- Las **binarias** se dejaron sin cambio (0/1).
- No se detectaron valores nulos.
- Se creó la etiqueta `is_normal`:
  - **1:** `Cover_Type = 2` (Lodgepole Pine, normal)
  - **0:** `Cover_Type ≠ 2` (anómalo)

El dataset quedó casi balanceado (Normal=48.8%, Anómalo=51.2%).
Se dividió en:

- **Train/Valid:** solo normales
- **Test:** mezcla de normales y anómalos

¿Por qué entrenar el Autoencoder solo con datos normales?

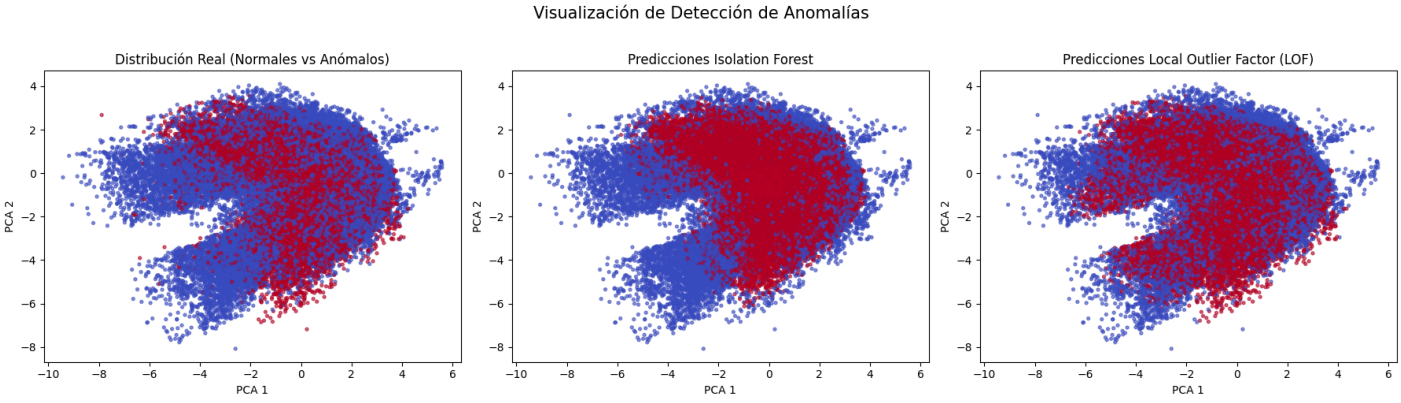
El Autoencoder debe aprender solo el patrón normal; incluir anómalos haría que los reconstruya y deje de distinguirlos.

2. Modelos y metodología

Se entrenaron tres detectores de anomalías:

1. **Autoencoder (AE)** con arquitectura simétrica `128–64–32–16–32–64–128`, activación *ReLU* y pérdida **MSE**.
   Entrenamiento con **EarlyStopping (paciencia=5)** para evitar sobreajuste.
2. **Isolation Forest (IF)** con 300 árboles y entrenamiento solo sobre normales.
3. **Local Outlier Factor (LOF)** en modo *novelty detection*, también con normales.

El umbral del AE se definió según el **percentil del error de reconstrucción en validación normal**.
Se compararon percentiles (90–99) y se seleccionó **p95** por su mejor F1.

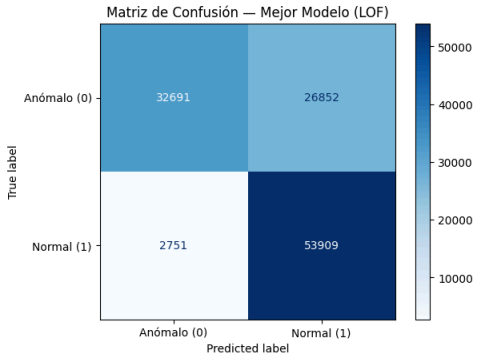


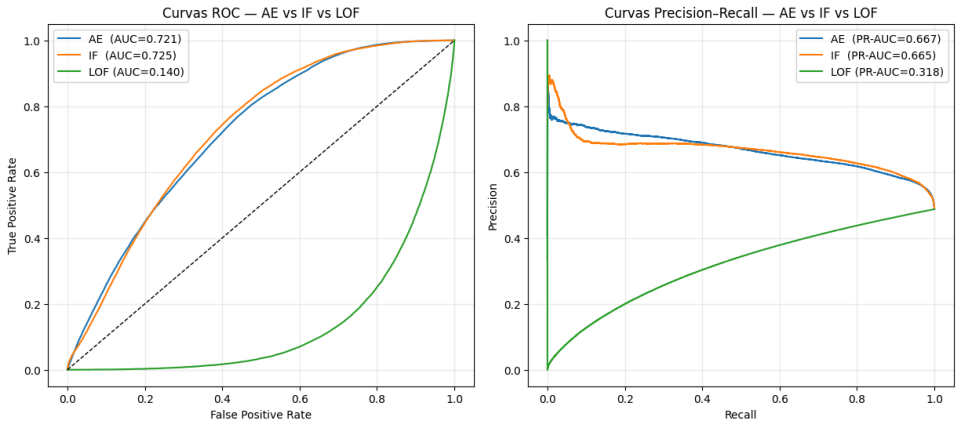
3. Resultados comparados

Modelo	Accuracy	F1	Precision	Recall	ROC-AUC	PR-AUC
AE	0.625	0.712	0.569	0.949	0.721	0.667
IF	0.630	0.715	0.573	0.949	0.725	0.665
LOF	**0.745**	**0.785**	**0.668**	**0.951**	0.140	0.318

Mejor modelo: LOF por F1 (0.785), aunque AE e IF muestran métricas más consistentes en ROC-AUC y PR-AUC.

Mátriz de confusion para el mejor modelo





- ROC-AUC es menos informativa aquí por el desbalance y la naturaleza del LOF.
- La métrica **PR-AUC** resulta más informativa que ROC-AUC para este tipo de problema.

4. Análisis de desempeño

- **Umbral del AE:** p95 equilibra precisión y recall; percentiles mayores aumentan FP.
- **Falsos positivos (FP):** anómalos con baja reconstrucción similares a normales.
- **Falsos negativos (FN):** normales con valores extremos en distancias o elevación.
- **Métricas :** PR-AUC (0.667) refleja mejor la utilidad del modelo que ROC-AUC (0.721), dada la ligera asimetría de clases.

5. Discusión y aplicabilidad

- 0El autoencoder logra reconstruir con precisión el patrón normal del bosque, detectando desviaciones con MSE.
- El entrenamiento solo con normales es clave para mantener un umbral interpretable.
- Los resultados indican que el enfoque basado en reconstrucción (AE) es **efectivo y escalable**, especialmente cuando no hay etiquetas de anomalías. El Isolation Forest puede complementar el AE para reducir FP, y el LOF resulta útil solo en regiones densas bien definidas.

6. Conclusión

El pipeline propuesto permitió detectar patrones anómalos de forma robusta.
El **Autoencoder con umbral p95** ofrece un equilibrio adecuado entre precisión y recall, mientras que **PR-AUC** se confirma como la métrica más representativa para este tipo de problema.