Universidad del Valle de Guatemala

Facultad de ingeniería

Data Science

Catedrático: Luís Furlán

Laboratorio 8. Detección de Anomalías

Joaquín André Puente Grajeda 22296

José Antonio Mérida Castejón 201105

Guatemala, 5 de octubre de 2025

1. Conjunto de Datos y Preparación

Covtype (Covertype) reúne mediciones topográficas y ambientales de parcelas forestales en Estados Unidos para predecir el tipo de cobertura forestal. Es un dataset grande y mixto (variables continuas y binarias), con 581,012 observaciones, 54 características y un objetivo multi–clase (Cover_Type, 7 clases). En este caso, buscamos definir "Lodgepole Pine" cómo valores "normales" y datos diferentes cómo "anomalías"

Preparación

Escala

Antes de iniciar con los modelos, escalamos los datos utilizando el StandardScaler de sklearn. Este fue aplicado únicamente a las variables numéricas, ignorando las variables binarias resultantes de columnas categóricas One Hot Encoded.

Train, Test y Validation

Para la división de los datos, utilizamos un conjunto de prueba y validación constando únicamente de los datos "normales" y un conjunto de prueba utilizando un split 50/50 de anomalías y normales.

2. Preguntas Guía

Umbral de AE

En este caso, utilizamos la métrica F1-Score en conjunto con las gráficas F1@Umbral para definir el umbral utilizado. Consideramos la métrica de F1 apropiada, ya que nuestro conjunto de prueba tiene una división 50-50 y esta métrica no se ve afectada por un posible desbalanceo de clases. Exploramos diferentes umbrales basados en percentiles, desde el 70-99.99 percentil en cuanto a puntajes de error utilizando Isolation Forest y el Autocodificador. En el caso de LOF, también trabajamos con percentiles de los puntajes que regresa el modelo.

Comparación de Métricas

Con una división de anomalías de 50/50, la métrica más estable para comparar modelos es ROC-AUC. PR-AUC también es informativa considerando la distribución de clases, pero su valor puede moverse más con pequeños cambios en precisión / recall. Adicionalmente, para decidir operación consideramos que F1@umbral es una métrica bastante adecuada como mencionamos anteriormente. Por último, la matriz de confusión complementa estas métricas y nos permite analizar realmente que fue lo que predijo el modelo. En conclusión, consideramos que todas las métricas tienen su uso y se complementan una a la otra. Sin embargo, la más adecuada para comparar entre diferentes modelos es ROC-AUC.

Falsos Positivos / Negativos

Con una proporción de anomalías 50/50, nuestro mejor modelo mostró comportamiento equilibrado entre clases: la tasa de falsos positivos (FP) sobre normales y la tasa de falsos negativos (FN) sobre anómalos fueron similares, lo que sugiere que el umbral seleccionado mantiene un buen compromiso entre precisión y recall. En términos prácticos, los FP implican revisar casos normales marcados como anómalos, mientras que los FN representan anomalías que pasan

desapercibidas. Si el contexto penaliza más uno de los dos, podemos ajustar el percentil del umbral para desplazar la balanza: bajar el umbral reduce FN (más recall) a costa de más FP; subirlo hace lo contrario.

Si la otra Clase fuera "Normal"

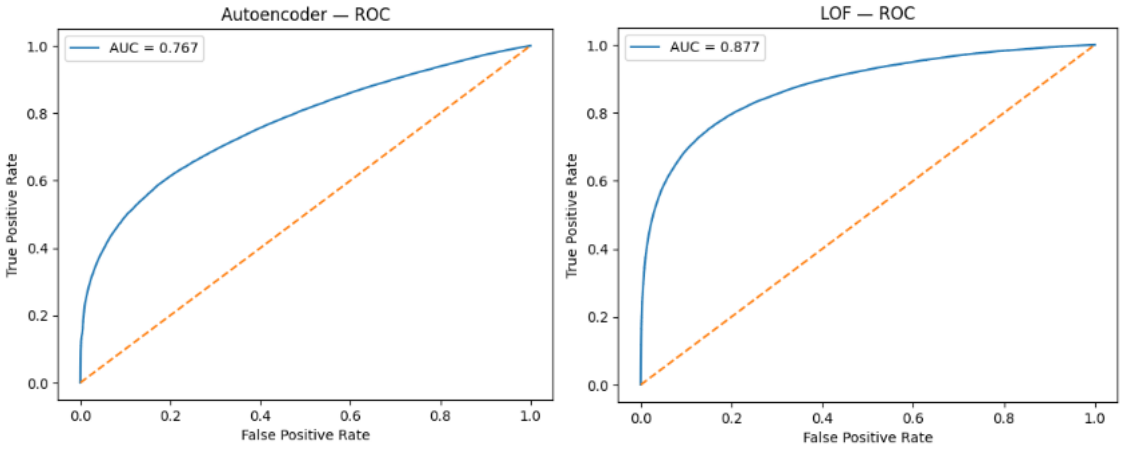
Tomando en cuenta nuestro mejor modelo, consideramos que si la otra clase fuera normal veríamos resultados bastante similares. En este caso, podríamos únicamente "darle la vuelta" y tendríamos resultados bastante buenos ya que se logran predecir valores "normales" de manera bastante precisa.

3. Resultados

Tabla 1. Classification Reports por Modelo

Modelo	Normal (P/R/F1)	Anómalo (P/R/F1)	Exactitud	ROC-AUC	PR-AUC
LOF	0.80 / 0.80 / 0.80	0.80 / 0.80 / 0.80	0.80	0.88	0.89
Isolation Forest	0.65 / 0.70 / 0.67	0.67 / 0.62 / 0.64	0.66	0.71	0.74
Autoencoder	0.69 / 0.70 / 0.70	0.70 / 0.69 / 0.69	0.70	0.77	0.80

Figuras 1 y 2. ROC-AUC AutoEncoder y LOF



4. Discusión

Los resultados muestran una ventaja consistente de LOF frente a Autoencoder e Isolation Forest. En concreto, LOF alcanzó las mejores métricas de ranking (ROC-AUC ≈ 0.88, PR-AUC ≈ 0.89) y un desempeño operativo sólido (P/R/F1 ≈ 0.80 en ambas clases), lo que sugiere que su criterio de densidad local captura bien la estructura del espacio de características (mixto: numéricas escaladas + binarias). El Autoencoder quedó en segundo lugar (ROC-AUC ≈ 0.77, PR-AUC ≈ 0.80, F1 ≈ 0.70), razonable para un modelo de reconstrucción, pero con separación más débil entre normales y anómalos (umbrales más bajos para lograr buen recall). Isolation Forest fue el más flojo (ROC-AUC ≈ 0.71, PR-AUC ≈ 0.74, F1 ≈ 0.66), probablemente por su menor sensibilidad a patrones locales en alta dimensión y por una separación de scores menos marcada.

5. Conclusiones

(1) LOF es el mejor detector en este dataset: ofrece el mejor ranking y un equilibrio FP/FN adecuado al umbral elegido, útil para operación. (2) El Autoencoder funciona como alternativa competitiva, pero requiere más ajuste para cerrar la brecha con LOF; aun así, su enfoque de reconstrucción permite un control fino vía percentiles/umbral. (3) Isolation Forest sirve como *baseline* pero, en esta configuración, no alcanza el desempeño de los otros dos. (4) Fijar umbrales con percentiles sobre validación de normales permitió controlar el FPR y comparar de forma justa; para extrapolar a despliegue, conviene además evaluar con prevalencia real y métricas operativas

5. Recomendaciones

(1) Priorizar LOF como modelo de producción; si el costo computacional es alto, reducir n_neighbors o aplicar PCA (15–30 dims) entrenada en normales antes de LOF. (2) Afinar el Autoencoder con normalización por capas y mayor capacidad con regularización (dropout/L2). (3) Realizar gráficas y tomar medidas adicionales para epxlorar diferentes umbrales, no solo F1@Umbral (3) Mantener Isolation Forest como *baseline*; si se conserva, aumentar n_estimators y el *subsample* para estabilizar el score. (4) Explorar espacios parametrales amplios para cada uno de los modelos.