Informe técnico: Predicción de riesgo de litiasis renal con un pipeline de Machine Learning reproducible

EDA y limpieza

La variable objetivo **stone_risk** muestra un balance de clases con un desequilibrio moderado (60,82%/39,17%). Respecto a los rangos clínicos **serum_calcium** muestra 1200 pacientes por debajo del mínimo clínico (valores distópicos). Y **gfr** muestra una distribución (esperada) donde predomina la disfunción renal.

El df está formado por 15 columnas numéricas correctamente tipadas y 9 categóricas con codificación clara y niveles bien distribuidos. Variables binarias sin ambigüedad.

Anti-leakage y definición de variables

Se han excluido las columnas **ckd_pred**, **ckd_stage**, **cluster** y **months** por contener información derivada de la variable objetivo.

y es la variable objetivo (**stone_risk**). Las variables predictoras (X) se han dividido en numéricas y categóricas, todas disponibles antes del evento a predecir, evitando un sobreajuste.

```
Variables numéricas para el preprocesado: ['serum_creatinine', 'gfr', 'bun', 'serum_calcium', 'ana', 'c3_c4', 'hematuria', 'oxalate_levels', 'urine_ph', 'blood_pressure', 'water_intake']

Variables categóricas para el preprocesado: ['alcohol', 'diet', 'family_history', 'painkiller_usage', 'physical_activity', 'smoking', 'stress_level', 'weight_changes']
```

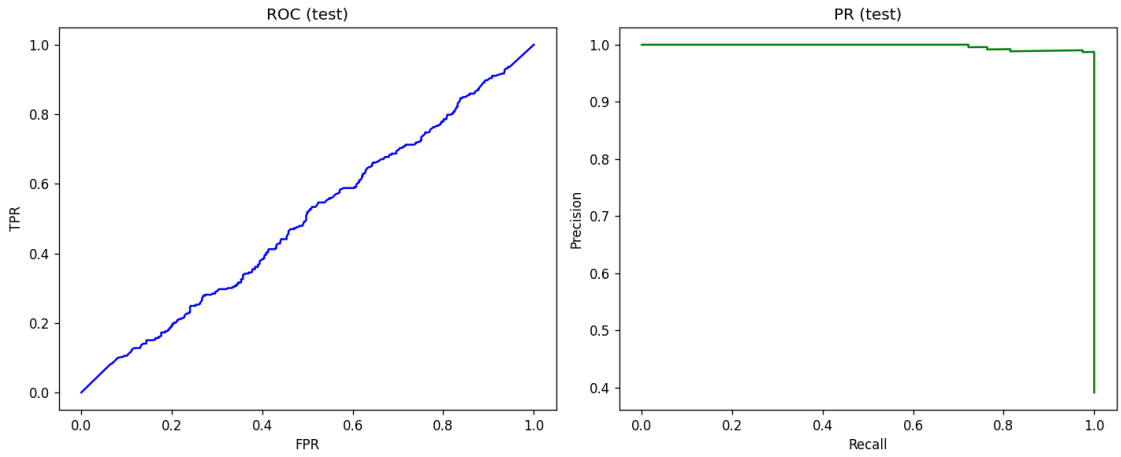
Preprocesado reproducible

Para la validación cruzada estratificada se ha aplicado **StratifiedKFold(n_splits=5)** con shuffle=True y random_state=42 ya que garantiza reproducibilidad, un balance de clases en cada fold y una prevención de sobreajuste por particiones sesgadas. Todo el preprocesado se aplica dentro del ciclo de validación cruzada, evitando cualquier fuga entre folds.

El flujo se ha estructurado en funciones independientes para EDA, preprocesado, validación y calibración.

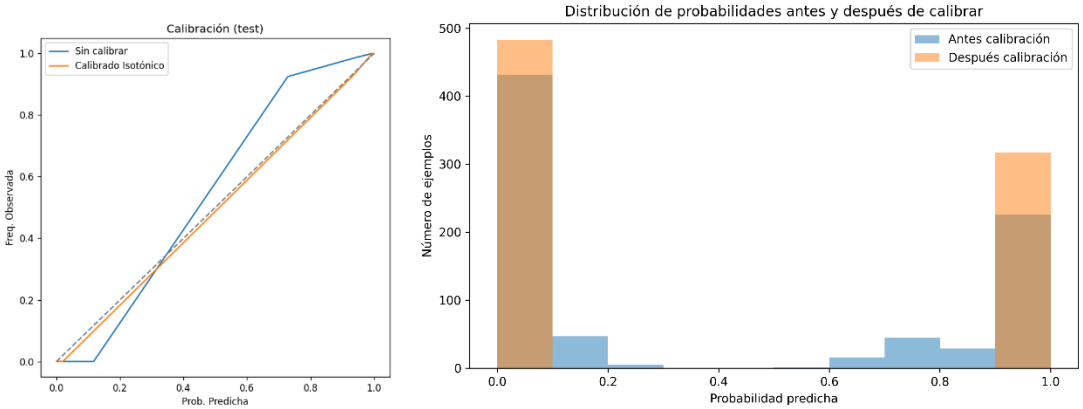
Métricas y validación principal

Se aplica un **RandomizedSearchCV** sobre un **tree_pipe** con un RandomForestClassifier. Los resultados en CV muestran unos valores de `CV ROC-AUC: 0.998 ± 0.001, CV PR-AUC : 0.997 ± 0.001`, con una cm `confusion (thr=0.5): TN 483 FP 4 FN 0 TP 313`



Calibración: método y Brier score

Se ha utilizado calibración isotónica que ha reducido el error cuadrático en más del 50% como se puede ver en el Brier score: `Brier sin calibrado: 0,012, Brier con calibrado: 0,005`. No hay cambios significativos en las probabilidades calibradas (0,391 *vs* 0,393).



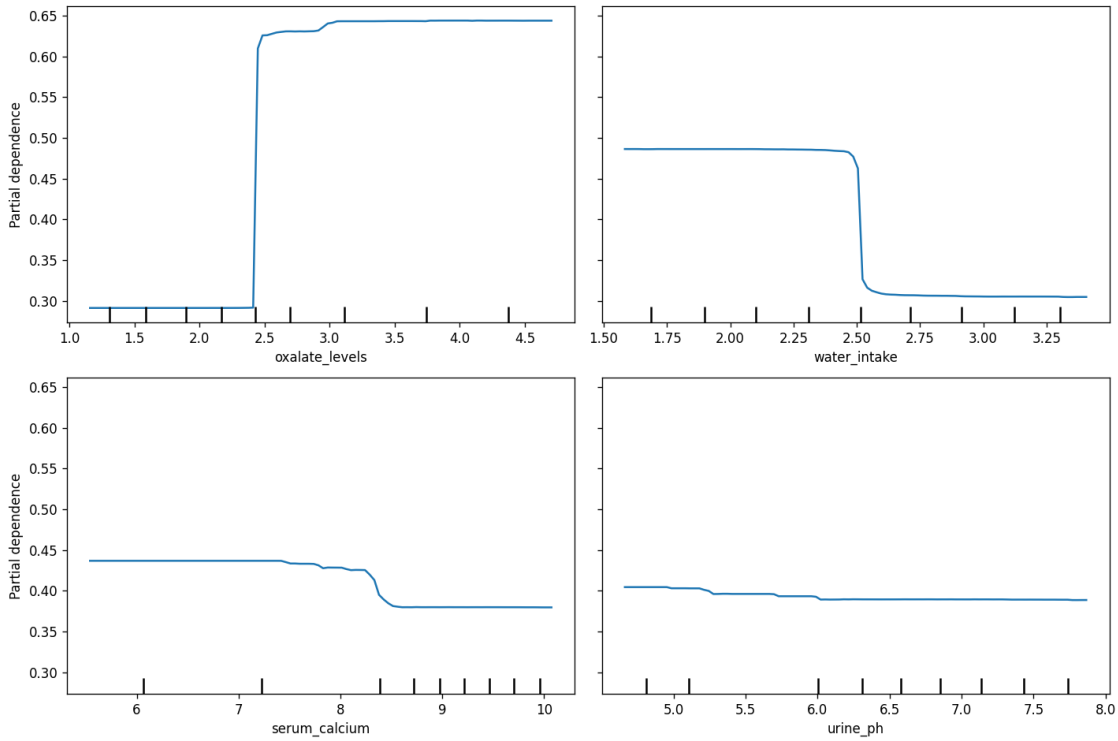
Selección de umbral: criterio y motivación

Se ha seleccionado el umbral que maximiza la sensibilidad con especificidad mínima aceptable (≥ 0,6). Dando un `Umbral aplicado: 0.184, Precisión: 0.981, Recall: 1.000, F1: 0.991`. El objetivo es priorizar la detección de positivos sin comprometer la utilidad clínica.

Interpretabilidad: PI y PDP/ALE sobre variables clave

La permutación nos indica que 4 variables destacan por encima del resto: `oxalate_levels: 0.095, water_intake: 0.025, serum_calcium: 0.003, urine_ph: 0.001`.

PDP muestra como las variables **oxalate_levels** y **water_intake** llevan el 95% del AUPRC.



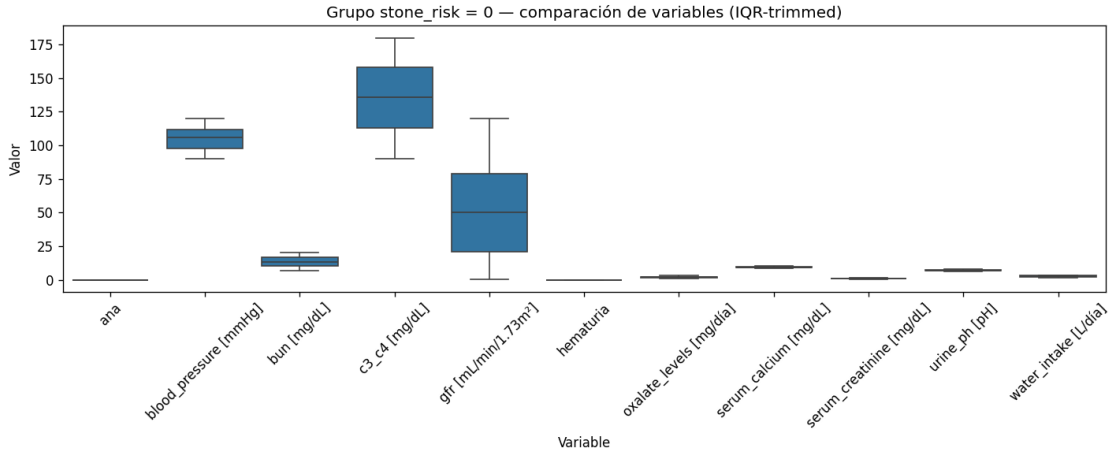
ALE muestra correlaciones muy fuertes (≥ 0,7) donde **oxalate_levels** está inversamente correlacionada con **serum_calcium** y **urine_ph** mientras que **water_intake** puede interpretarse de forma independiente.

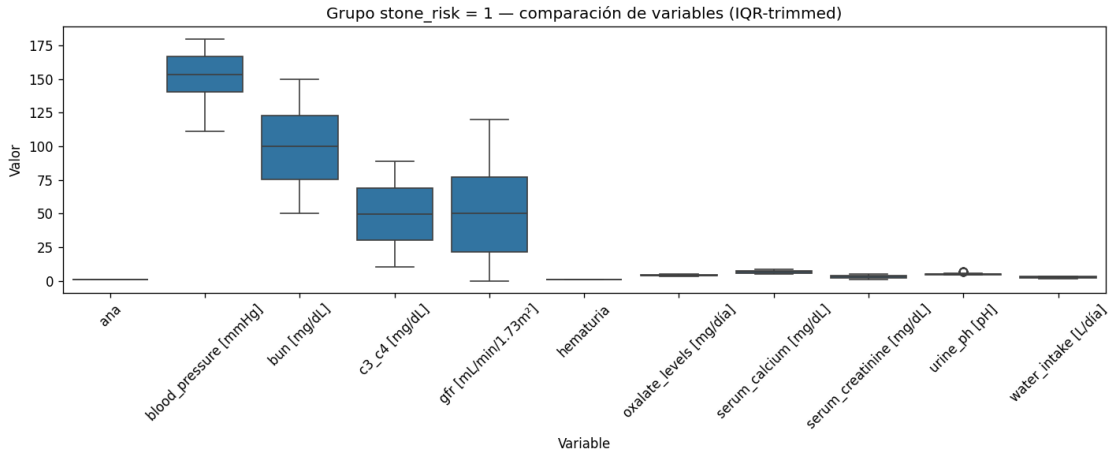
```
=== Correlación entre variables clave ===
                oxalate_levels  water_intake  serum_calcium  urine_ph
oxalate_levels           1.000        -0.006         -0.746    -0.746
water_intake            -0.006         1.000          0.023     0.014
serum_calcium           -0.746         0.023          1.000     0.767
urine_ph                -0.746         0.014          0.767     1.000

=== Variables con correlación significativa (usables en ALE) ===
['oxalate_levels', 'serum_calcium', 'urine_ph']
```

NOTA FINAL

Al analizar los datos, llama la atención que la variable **oxalate_levels** muestra un máximo de 5 mg/día, lo que se considera que todos los pacientes padecen de hipooxaluria, lo que es extraño porque el riesgo del litiasis renal está en la hiperoxaluria.

Los niveles de oxalato han resultado ser una variable predominante porqué es la variable que más variabilidad explica dentro del conjunto, aunque esté por debajo del rango clínicamente relevante, pero esto no quiere decir que clínicamente los niveles de oxalato sean un factor de riesgo en estos casos, como demuestra la visualización de los factores dominantes para padecer riesgo de litiasis renal: pH urinario ácido, bajo volumen urinario, calcemia elevada y PAS alta. Los niveles de oxalato tienen valor predictivo relativo, pero no etiológico.





Por tanto, antes de seguir con el mejor modelo predictivo, revisaría otros df con variables similares.