

Tecnológico de Monterrey

Escuela de Ingeniería y Ciencias

Maestría en Inteligencia Artificial Aplicada

Proyecto Integrador · TC5035.10

Avance 2. Ingeniería de características

Bioactive AI

Predicción y priorización de compuestos bioactivos con potencial neuroprotector a partir de su estructura molecular

Profesora titular

Dra. Grettel Barceló Alonso

Patrocinadora

Dra. Mariana Martínez Ávila

Equipo 7

Ingrid Pamela Ruiz Puga	A01021209
Artemio Santiago Padilla Robles	A01796613
José Antonio Vázquez Martínez	A01797208

11 de octubre de 2026

Índice

1	Introducción	3
1.1	Alcance y continuidad del análisis exploratorio	3
2	Datos de entrada y preparación	4
3	Representación molecular	4
3.1	Descriptores calculados desde SMILES	4
3.2	Fingerprints de subestructuras	4
3.3	Características derivadas del dominio	5
3.4	Codificación de variables categóricas	6
4	Partición y control de fuga estructural	6
5	Transformación y escalamiento	7
6	Selección y extracción de características	8
6.1	Filtros de varianza y correlación	8
6.2	Extracción por componentes principales	8
7	Salidas de la preparación	9
8	Conclusiones	9
	Bibliografía	10

1 Introducción

La ingeniería de características transforma una estructura molecular en variables que un algoritmo puede procesar. En este avance se completa la preparación de la etapa de permeabilidad de la barrera hematoencefálica (BBB); el notebook ajusta modelos auxiliares para comparar transformaciones, pero no se presenta un modelo predictivo final. La unidad de entrada es el SMILES estandarizado y la fuente primaria sigue siendo el archivo curado por la patrocinadora.

El proyecto conserva su planteamiento de dos etapas: permeabilidad BBB y actividad neuroprotectora. Una predicción BBB—no descarta un compuesto; ambas etapas aportan evidencia paralela para su priorización. Este reporte se limita a la primera etapa porque es donde ya existen datos curados y un análisis de características reproducible. La matriz construida no demuestra por sí misma actividad neuroprotectora ni exposición cerebral.

1.1 Alcance y continuidad del análisis exploratorio

La fase de análisis exploratorio de datos (EDA) se extenderá una semana más, en paralelo con el cierre documental de ingeniería de características. Esto no significa que el notebook de ingeniería esté incompleto: ya fue ejecutado y las transformaciones fueron medidas; lo pendiente es incorporar al análisis las observaciones del EDA y completar el reporte con las decisiones que siguen abiertas. El margen corresponde al periodo entre el Avance 2 (11 de octubre) y el Avance 3: es trabajo del equipo previo al baseline¹, no una pausa ni un aplazamiento de la entrega formal del notebook, y no fija ningún plazo a la patrocinadora. Qué entradas del registro necesitan su definición se consulta en la Ruta, que es donde se mantiene.

En ese periodo falta completar la lectura de la diversidad estructural—incluida la concentración de clases en andamios grandes—, precisar el alcance de la estratificación por `group` ante el posible riesgo de fuga de información² y aclarar la pertinencia del conjunto externo. También se incorpora la cuestión abierta de los 131 esqueletos de conectividad (429 filas) con estereoisómeros³ de etiquetas opuestas: los datos muestran la discordancia, pero no permiten distinguir por sí solos una diferencia estereoquímica real de una inconsistencia entre ensayos. Esta pregunta requiere criterio farmacológico y se mantiene abierta en la decisión 1.8 del registro.

La ingeniería de características se documenta con lo que ya está medido en el notebook 02. Los valores derivados del EDA—por ejemplo, umbrales exploratorios— se identifican como dependientes de los datos y no como reglas biológicas universales. Cuáles de las decisiones dependen del criterio de la patrocinadora está marcado en el registro, que es la lista única.

Este documento acompaña el notebook ejecutado de principio a fin. Las cifras provienen de sus celdas y de la actualización metodológica documentada en el registro; no se generaron datos o experimentos nuevos para escribir el reporte.

Ninguna entrada del registro detiene este reporte: las que siguen como propuestas se trabajan sobre su versión vigente y se señalan como no ratificadas donde aparecen. La rúbrica solicita la liga al notebook y que se ejecute secuencialmente, por lo que el PDF es complementario y no sustituye el entregable formal.

¹Modelo de referencia entrenado con una configuración sencilla, que sirve como punto de comparación para modelos posteriores.

²Uso indirecto, durante entrenamiento o selección, de información que no estaría disponible al predecir o que deriva de la etiqueta objetivo; puede inflar artificialmente la evaluación.

³Molécula con los mismos átomos y conectividad que otra, pero con distinta disposición espacial; esa diferencia puede modificar su actividad biológica.

2 Datos de entrada y preparación

La fuente es el archivo de permeabilidad BBB curado por la patrocinadora, no las descargas independientes de B3DB. Tiene 7,811 filas y 7,805 etiquetas BBB integradas. Los 7,811 SMILES parsean en RDKit; las seis filas sin clase integrada no se incorporan a la clasificación ni se les imputa una etiqueta a partir de su conectividad, porque la clase puede depender de la estereoquímica.

Las 7,805 observaciones etiquetadas contienen 4,956 BBB+ (63.5%) y 2,849 BBB- (36.5%). Comparten solo 4,018 esqueletos de conectividad, por lo que las filas no representan observaciones estructuralmente independientes. La preparación mantiene todas las filas etiquetadas; no filtra por la columna `group`.

La columna `group` tiene asociación con la etiqueta y su criterio de asignación no está completamente documentado. Se conserva para describir y estratificar, pero no se interpreta como escala de calidad ni se usa para descartar filas. La preocupación de que pueda contener información relacionada con la variable objetivo no queda resuelta por estratificar: su uso como predictor requiere consulta con la patrocinadora. En consecuencia, la matriz molecular principal se construye con estructura y descriptores, no con un supuesto de que `group` sea una característica válida. Esta cautela atiende la observación recibida en el Avance 1: estratificar permite estudiar heterogeneidad, pero no descarta una posible fuga.

3 Representación molecular

3.1 Descriptores calculados desde SMILES

RDKit calcula 217 descriptores moleculares 2D por molécula: resúmenes de tamaño, carga, polaridad, lipofiliidad, flexibilidad y topología (Landrum & RDKit contributors, 2026). Cinco descriptores constantes no aportan variación y ocho contienen valores NaN en esta muestra, así que no pueden entrar sin tratamiento adicional. Además se excluye `Ipc`: aunque sus valores son finitos en precisión de 64 bits, puede alcanzar aproximadamente 10^{41} , por encima del máximo representable en `float32`; scikit-learn convierte internamente a ese formato y el valor puede hacer fallar el ajuste. El chequeo de valores finitos por sí solo no detectaría este problema.

Quedan 203 descriptores fisicoquímicos utilizables. Se conservan las columnas de tamaño, polaridad e hidrofobicidad que el EDA identificó como relevantes, entre ellas TPSA, HBD, HBA, logP y peso molecular. Esto no afirma que cada descriptor aporte señal independiente: las correlaciones altas que se miden después motivan un filtrado explícito.

3.2 Fingerprints de subestructuras

Además de los resúmenes fisicoquímicos se generan dos fingerprints, representaciones binarias que registran la presencia de fragmentos moleculares. Morgan/ECFP4 usa radio 2 —el 4 de ECFP4 corresponde al diámetro— y 2,048 bits plegados. La matriz tiene densidad de bits activos de 2.33%; cada compuesto activa una mediana de 47 bits (rango 1–142). El plegado reduce almacenamiento, pero distintos fragmentos pueden terminar en el mismo bit.

Como alternativa interpretable se aplica Sort & Slice: ordena las subestructuras por frecuencia y conserva un vocabulario de 2,048, aprendido exclusivamente en entrenamiento. Cada columna corresponde a una sola subestructura, por lo que se puede rastrear un bit hasta el fragmento que representa. La matriz tiene densidad de 1.91%. Con ese vocabulario se cubre el 84.8% de las subestructuras del entrenamiento, pero el 68.0% de las del conjunto de prueba. La diferencia de 16.8 puntos muestra que los andamios menos frecuentes de la prueba incluyen fragmentos que el vocabulario aprendido no vio.

El tamaño del vocabulario no queda ratificado. El entrenamiento contiene 11,475 subestructuras distintas; con 2,048 columnas se retiene el 18% de ese vocabulario y una parte de la química novedosa de prueba queda sin representar. La cobertura observada es una medición de representación y no basta por sí sola para elegir un tamaño final. Lo que sí está medido es el desempeño: con el error estándar corregido por pliegue, **los tamaños de vocabulario son indistinguibles entre sí desde aproximadamente 128 columnas**, y 2,048 es el menor tamaño dentro del ruido en dos de los tres modelos (1.7). El tamaño se elige por modelo en el Avance 3, con los modelos ya afinados.

3.3 Características derivadas del dominio

El notebook evalúa tres variables adicionales motivadas por lo observado en el EDA. Se reportan como características candidatas, no como reglas de clasificación validadas:

Tabla 1: Características derivadas y evaluación univariante reproducida en el notebook 02

Característica	Construcción	AUC	Comparación y resultado del filtrado
<code>regla_2p</code>	Indicador de $\text{TPSA} < 67 \text{ \AA}^2$ y $\text{HBD} \leq 1$	0.684	TPSA: 0.844; el indicador no la supera
<code>tpsa_por_mw</code>	TPSA dividida entre peso molecular	0.810	Correlación máxima: 0.755 con logP; se conserva
<code>lipo_neta</code>	$\log P - \text{TPSA} / 100$	0.795	Correlación 0.975 con logP, que está protegido del filtro; sale de la matriz

La AUC individual y la correlación no dicen si una característica aporta. Dicen si tiene señal y si es redundante, que es otra pregunta: lo que importa es si añade algo que el modelo no tenga ya. Medido así —`scripts/experimento_derivadas.py`, con la misma validación cruzada agrupada, con y sin ellas, como diferencia pareada de MCC sobre 15 evaluaciones— **ninguna de las tres aporta**: con regresión logística la diferencia es de $+0.0001 \pm 0.0014$, y con Random Forest las tres juntas quedan 0.0062 ± 0.0039 **por debajo** de no usarlas. `tpsa_por_mw`, la única con asomo de señal, gana en 10 de 15 pliegues por nueve diezmilésimas.

`regla_2p` y `tpsa_por_mw` se conservan en la matriz porque tampoco estorban, y retirarlas es un cambio de una línea. `lipo_neta` ya no entra: correlaciona 0.975 con logP, que está protegido del filtro de correlación, de modo que ahora sale la derivada y no el original. La versión anterior de esta sección hacía lo contrario —conservaba `lipo_neta` por su estadístico F y descartaba logP—, y la razón para invertirlo es de explicabilidad: logP es el descriptor que se puede explicar, y además el más frágil del conjunto, con una caída de AUC de 0.744 a 0.594 entre entrenamiento y validación.

La regla `regla_2p` se conserva como característica evaluada, pero su corte $\text{TPSA} < 67 \text{ \AA}^2$ procede del análisis exploratorio y no tiene intervalo de confianza. El umbral que maximiza la separación cambia al cambiar el subconjunto: el notebook lo vuelve a calcular usando únicamente entrenamiento y obtiene 107 $\text{\AA}^2$. Por esa variación, los umbrales no se presentan como constantes universales ni se usan para filtrar compuestos.

Las AUC de la tabla son comparaciones descriptivas del notebook, sin intervalos de confianza y no estimaciones de rendimiento fuera de muestra. El reporte 1 había recibido la observación de que la derivación de reglas y la evaluación discriminativa pertenecen a una fase posterior al EDA; aquí se ubican como parte de la selección de características, con sus límites explícitos.

3.4 Codificación de variables categóricas

El notebook convierte a indicadores *one-hot* las categorías auxiliares `group` y el indicador de coincidencia exacta con COCONUT (`natural`). Esta representación permite tratar categorías sin imponerles un orden numérico. Las columnas codificadas se guardan en una matriz auxiliar, separada de las características moleculares. El indicador `natural` significa coincidencia estructural exacta: el conjunto distingue 3,366 productos naturales exactos, 1,005 análogos que solo difieren en estereoquímica y 3,440 compuestos sin equivalente conocido. Cómo tratar esos análogos sigue abierto; por eso la variable no se incorpora automáticamente como predictor.

No se discretizan de forma general los descriptores continuos. La transformación de valores a intervalos descartaría diferencias dentro de cada intervalo y requeriría cortes que no están justificados como decisiones del proyecto. El indicador `regla_2p` sí es una característica binaria construida a partir de una regla ya analizada; no equivale a discretizar todo el espacio de descriptores.

La codificación no resuelve si esas variables deben entrar al modelo. En particular, `group` podría filtrar información asociada a la etiqueta BBB, y el tratamiento del origen natural depende de cómo se definan los análogos que difieren en estereoquímica. Ambas decisiones quedan pendientes de consulta; no se asume aquí que el *one-hot encoding* convierta una variable en un predictor apropiado.

4 Partición y control de fuga estructural

El conjunto se separa con un scaffold split⁴ por andamio de Bemis–Murcko, que representa los anillos y enlaces que los conectan, y se agrupa también por esqueleto de conectividad para mantener juntas las variantes estereoquímicas (Bemis & Murcko, 1996). Los andamios frecuentes se asignan a entrenamiento y los raros a prueba, una evaluación más exigente que repartir filas al azar.

Tabla 2: Partición por andamio y proporción de clase BBB+

Subconjunto	Compuestos	Andamios	BBB+
Entrenamiento	5,557	1,741	64.0%
Validación	749	498	72.0%
Prueba	1,499	1,390	57.3%

El procedimiento produce 3,629 andamios distintos; 312 compuestos acíclicos, que no tienen andamio de anillo, se asignan a entrenamiento. La verificación del notebook encuentra cero andamios y cero esqueletos de conectividad compartidos entre particiones. Por tanto, no hay solapamiento estructural según esos dos identificadores.

La composición de prueba difiere de entrenamiento: contiene 57.3% BBB+ frente a 64.0% en entrenamiento. No se corrige esa diferencia porque refleja la asignación por andamios raros; se reporta para interpretar las métricas futuras. La línea base de AUPRC para BBB– en prueba es 0.427, igual a la prevalencia de esa clase en dicho subconjunto. La prueba se mantiene intacta durante el ajuste de vocabularios, umbrales, escaladores y selectores.

⁴Partición que asigna andamios moleculares completos a una sola partición, en lugar de repartir filas al azar, para medir generalización a estructuras no vistas.

5 Transformación y escalamiento

Los 203 descriptores tienen escalas muy diferentes: la amplitud observada va de aproximadamente 0.51 a 18,796. Además, muchas distribuciones son asimétricas. El notebook compara ausencia de transformación, escalado, min-max y Yeo-Johnson. En 193 de 203 descriptores hay valores menores o iguales a cero; por ello, el logaritmo y Box-Cox no son transformaciones aplicables de manera directa a todas las columnas. No se aplican transformaciones exponenciales ni de raíz cuadrada, pues el notebook no presenta una necesidad medida ni una comparación de esas opciones. En el conjunto de columnas y filas de entrenamiento usado en esa comparación, Yeo-Johnson reduce el número con asimetría absoluta mayor que 2 de 106 a 66.

La transformación se compara en validación cruzada agrupada por andamio y conectividad, con cinco pliegues y tres semillas; cada ajuste aprende sus parámetros solo del pliegue de entrenamiento y la prueba final no se toca. Los resultados medios de las 15 evaluaciones son:

Tabla 3: Promedios de validación cruzada agrupada documentados en la decisión 1.5; AUPRC para BBB-

Transformación	Logística, MCC	Logística, AUPRC	Random Forest, MCC	Random Forest, AUPRC
Sin escalar	0.687	0.867	0.704	0.901
Estandarización	0.690	0.864	0.705	0.901
Min-max	0.695	0.878	0.704	0.901
Yeo-Johnson	0.665	0.862	0.705	0.901

Sin escalar, la regresión logística agota las 3,000 iteraciones de `lbfgs` y no converge⁵ en ninguno de los 15 pliegues. Con las tres transformaciones converge: en 53 iteraciones con min-max y entre 137 y 144 con estandarización o Yeo-Johnson. Por ello, el MCC de 0.687 de la fila sin escalar refleja un ajuste incompleto y no es comparable con los demás. En la ejecución documentada, el ajuste mediano tomó 0.399 s sin escalar y 0.007 s con min-max; el tiempo depende del equipo, mientras que la falta de convergencia y las iteraciones observadas motivan el escalado.

Para las comparaciones de desempeño entre transformaciones convergentes, ningún intervalo del 95% excluye cero. Con el mismo parámetro de regularización `C` para las cuatro, la diferencia mayor de MCC —min-max contra Yeo-Johnson en regresión logística— es 0.030, con intervalo $[-0.010, 0.071]$. Esa diferencia, sin embargo, no es de la transformación: cuando cada una elige su propio `C` por validación cruzada interna, min-max y Yeo-Johnson empatan en 0.695 y la mayor diferencia baja a 0.004. Los valores elegidos están a dos órdenes de magnitud de distancia —0.01 con estandarización y Yeo-Johnson, 1.0 con min-max—, de modo que comparar a `C` fijo mide la transformación y la regularización a la vez. En Random Forest, las diferencias de MCC son de 0.000 a 0.001, y las de AUPRC son indistinguibles al redondeo; su tiempo de ajuste tampoco cambia de forma apreciable. Esto respalda no escalar para los árboles y escalar para la regresión logística por convergencia, no por una mejora demostrada de sus métricas. Min-max queda como opción inicial por su convergencia más rápida y el mayor AUPRC observado, no como superioridad estadística establecida. La decisión 1.5 sigue abierta en cuanto al balance entre explicabilidad

⁵En un algoritmo iterativo, alcanzar una solución estable antes de agotar el límite de iteraciones; si no converge, las métricas corresponden a un ajuste incompleto.

y PCA. El notebook conserva las matrices transformada y sin transformar para mantener las alternativas reproducibles.

6 Selección y extracción de características

6.1 Filtros de varianza y correlación

La selección se aplica a 206 columnas de descriptores y características derivadas, **sobre los valores sin transformar**. El primer filtro descarta las columnas casi constantes por valor modal —un valor cubre el 99% o más del entrenamiento—: se van 24 y quedan 182. El segundo busca pares con correlación de Spearman absoluta mayor que 0.95; se retiran 31 variables redundantes y quedan 151. El filtrado reduce almacenamiento y evita que el modelo reciba varias columnas casi equivalentes.

Una versión anterior de esta sección aplicaba un umbral de varianza (> 0.01) sobre la matriz ya transformada con Yeo–Johnson y estandarizada, donde por construcción toda columna tiene varianza 1: de las 206 solo eliminaba una. Medir ambos criterios mostró que el umbral de varianza, puesto en ese orden, no filtraba. Con el criterio por valor modal y la correlación de rangos, la selección **ya no depende de la transformación**: el notebook comprueba que las mismas 151 columnas resultan con estandarización, min–max y Yeo–Johnson.

Cinco descriptores quedan protegidos de ambos filtros —TPSA, HBD, HBA, logP y peso molecular— porque son los que la literatura usa para explicar permeabilidad y los que el proyecto exige para verificar cualquier afirmación estructural. Con esa protección, `lipo_neta` deja de desplazar a logP y es la derivada la que sale por correlación (0.975 con logP).

Los parámetros de las transformaciones y de selección se aprenden solo con entrenamiento. De ese modo, ni el valor modal, ni las correlaciones, ni el umbral para características derivadas usan la distribución de prueba. El corte de correlación es una propuesta operativa medida en esta muestra y no un valor ratificado por la patrocinadora.

6.2 Extracción por componentes principales

El análisis de componentes principales (PCA) comprime las columnas seleccionadas en 85 componentes que explican el 95% de la varianza, cerca de la mitad. Cada componente mezcla múltiples descriptores y deja de corresponder a una propiedad o fragmento químico identificable, y la explicabilidad a nivel de subestructura es un requisito del proyecto: por eso PCA **no es una representación candidata**, y el notebook lo conserva solo como vista del espacio químico (dos componentes reúnen el 28.8% de la varianza y muestran que la partición de prueba cae en una región más dispersa que el entrenamiento).

La comparación de desempeño con PCA se midió en la versión anterior del notebook —con Random Forest el MCC cambiaba de 0.566 a 0.558 y el AUROC de 0.916 a 0.872, con el intervalo bootstrap abarcando cero — y ya no se reejecuta, porque la decisión se sostiene por explicabilidad y no por esa diferencia. Lo que el notebook sí mide es el costo de reducir: de 206 a 151 columnas el MCC del Random Forest sobre validación baja de 0.589 a 0.562 y el de la regresión logística sube de 0.484 a 0.497.

El ANOVA univariante (prueba F) se calcula con las filas de entrenamiento para ordenar las características candidatas dentro de los grupos correlacionados; no se usa para afirmar diferencias estadísticamente significativas en todos los descriptores. No se aplicó chi-cuadrado: requiere valores no negativos y no corresponde directamente a la mayoría de los descriptores continuos firmados. No se aplicó análisis factorial porque el objetivo no es inferir factores latentes de un constructo, sino conservar correspondencia

entre columnas y propiedades/subestructuras químicas. La selección y PCA cubren los métodos de filtrado y extracción ejecutados; no se aplicaron todos los métodos ilustrativos de la rúbrica por el solo hecho de que aparezcan como opciones.

7 Salidas de la preparación

La fase entrega representaciones trazables desde SMILES: descriptores 2D, Morgan/ECFP4, Sort & Slice, características derivadas y codificaciones auxiliares, además de una partición que no comparte andamios ni esqueletos entre entrenamiento, validación y prueba. Los artefactos intermedios se guardan en formato Parquet y se excluyen del control de versiones; los parámetros de escalado, selección y vocabulario se guardan aparte para aplicarlos a moléculas nuevas.

Las matrices transformadas y sin transformar se mantienen para comparar y construir canalizaciones específicas por modelo: escalamiento para la regresión logística y matriz original para los árboles. El equipo definirá la transformación operativa; cambiarla **ya no obliga a repetir la selección de variables**, porque los dos filtros operan sobre los valores sin transformar. La consulta a la patrocinadora de la decisión 1.5 se limita a priorizar interpretabilidad frente a PCA. El vocabulario de Sort & Slice también se conserva junto con su correspondencia a subestructuras, necesaria para transformar moléculas futuras e interpretar las columnas. La prueba de humo confirma que el flujo puede ejecutarse, pero no sustituye la evaluación comparativa de modelos ni constituye una afirmación de rendimiento final; se evalúa sobre validación y no sobre la partición de prueba, que queda reservada.

Las integraciones recientes agregan registro reproducible de corridas, verificación SHA-256⁶ de los archivos de datos y pruebas automatizadas de la partición, Sort & Slice y limpieza de descriptores. También incorporan el registro retrospectivo de las mediciones de escalamiento y del vocabulario de tamaño L, con sus parámetros, métricas y entorno. Estos controles permiten rastrear y reproducir mejor la evidencia reportada, sin cambiar las salidas científicas del notebook.

8 Conclusiones

En CRISP-ML(Q), este Avance corresponde a la preparación de datos previa a la ingeniería y evaluación del modelo (Visengeriyeva et al., 2023). La estructura molecular ya se convierte en matrices numéricas reutilizables; se eliminaron columnas constantes, no utilizables o incompatibles con la precisión interna; se construyeron descriptores derivados motivados por el dominio; se redujo redundancia; y se dejó una prueba estructuralmente disjunta. Las transformaciones y los selectores se ajustan solo en entrenamiento, lo que limita la fuga de información.

La medición incorporada sobre escalamiento aclara que la regresión logística sin transformar no converge y respalda escalarla, mientras que el Random Forest no requiere escalamiento. Sin embargo, medir el comportamiento del algoritmo no resuelve las decisiones que requieren contexto de la patrocinadora. Siguen sin ratificar la elección entre la matriz interpretable y PCA (1.5) y la definición y evaluación de los compuestos naturales y sus análogos estereoquímicos (1.6). El tamaño del vocabulario de Sort & Slice (1.7) dejó de ser una pregunta para ella: está medido y se elige por modelo con los modelos afinados. También falta confirmar cómo puede usarse `group`: su criterio de asignación no está completamente publicado y su asociación con la etiqueta podría filtrar información; estratificar no descarta esa posibilidad (1.1). La decisión 1.8 añade otra pregunta abierta: si las etiquetas BBB opuestas en 131 esqueletos

⁶Función criptográfica que produce una huella digital del contenido de un archivo; permite detectar cambios accidentales o archivos distintos.

compartidos por estereoisómeros reflejan diferencias farmacológicas reales o ensayos no comparables; esa interpretación puede afectar la representación y el límite alcanzable del modelo. Los datos disponibles permiten cuantificar alternativas, pero no establecen por sí solos las preferencias científicas ni los criterios de uso que faltan. Estos puntos deben llevarse a consulta con la patrocinadora antes de presentarlos como decisiones cerradas. La señal de cobertura más baja en prueba advierte que parte de las subestructuras raras no está representada por el vocabulario actual. Asimismo, el cambio de prevalencia entre particiones exige informar la línea base de BBB– y métricas sensibles a la clase minoritaria junto con MCC.

El conjunto de 154 compuestos atribuido al estudio citado (Spielvogel et al., 2025), con 68 positivos, 42 negativos y 44 casos de eflujo, tampoco se presenta aquí como validación externa. El análisis del Avance 1 identificó el artículo y el archivo, pero no dejó establecido un criterio de selección que permita justificar su representatividad respecto de esta tarea; además, el archivo disponible contiene nombres y descriptores precalculados, no SMILES ni InChI, por lo que no permite una comparación estructural reproducible. Antes de asignarle el papel de conjunto externo se debe aclarar con la patrocinadora su pertinencia para el objetivo del proyecto, documentar el criterio de selección y confirmar si pueden obtenerse identificadores químicos; mientras falte esa información, no se usa para validar el modelo.

El notebook del EDA ya contiene un resultado que debe quedar visible al interpretar la partición: el andamio⁷ mayor reúne 410 compuestos y es 99.5% BBB+, mientras que otros andamios grandes son completamente BBB– (por ejemplo, uno con 59 compuestos); por tanto, en estos grupos la estructura está fuertemente asociada con la clase. Incorporar esa lectura y su efecto en la evaluación es una tarea pendiente del EDA, no una incertidumbre que pueda resolver la patrocinadora. Las entradas que sí dependen de su criterio están marcadas como tales en la Ruta, que es la lista única: este reporte no la duplica, porque dos listas de lo mismo se desincronizan en cuanto una cambia. La fase EDA continúa una semana más; durante ese mismo periodo se cierra la documentación del proceso de ingeniería de características, cuyo notebook ya está ejecutado. No es una semana sin avanzar: el siguiente hito es el baseline. Hasta entonces, este documento es una base reproducible y sus propuestas no deben presentarse como decisiones ratificadas. En modelado se compararán alternativas con particiones agrupadas y se elegirá el preprocesamiento dentro de cada canalización, conservando la interpretabilidad química. Si no se encuentra un patrón generalizable en andamios no vistos, esa ausencia de evidencia predictiva será un resultado válido.

Bibliografía

Bemis, G. W., & Murcko, M. A. (1996). The Properties of Known Drugs. 1. Molecular Frameworks.

Journal of Medicinal Chemistry, 39(15), 2887-2893. <https://doi.org/10.1021/jm9602928>

Landrum, G., & RDKit contributors. (2026,). *RDKit: Open-Source Cheminformatics Software*.

Spielvogel, C. P., Schindler, N., Schröder, C., Stellnberger, S. L., Wadsak, W., Mitterhauser, M., Papp, L., Hacker, M., Pichler, V., & Vranka, C. (2025). Enhancing Blood-Brain Barrier Penetration Prediction by Machine Learning-Based Integration of Novel and Existing, In Silico and Experimental Molecular Parameters from a Standardized Database. *Journal of Chemical Information and Modeling*, 65(6), 2773-2784. <https://doi.org/10.1021/acs.jcim.4c02212>

Visengeriyeva, L., Kammer, A., Bär, I., Kniesz, A., & Plöd, M. (2023). *CRISP-ML(Q). The ML Lifecycle Process*. INNOQ. <https://ml-ops.org/content/crisp-ml>

⁷Núcleo molecular de anillos y enlaces que los conectan, abstraído de los sustituyentes; se usa aquí para agrupar estructuras relacionadas.