DESARROLLO DE UN MODELO DE PREDICCIÓN DE MOLÉCULAS QUE ATRAVIESAN LA BARRERA HEMATOENCEFÁLICA CON IA.

Cristhian Camilo Ibáñez Bersinger
Código: 8976108

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Julián Gil González
PhD En Ingeniería

Codirectora
Nerlis Paola Pájaro Castro
PhD En Toxicología Ambiental

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 5 DE 2024

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

1. INTRODUCCIÓN

La barrera hematoencefálica (BHE) desempeña un papel crucial en el Sistema Nervioso Central (SNC), actuando como una barrera física y bioquímica que regula selectivamente el paso de sustancias desde la circulación sanguínea al cerebro. Su función fundamental radica en la protección cerebral contra toxinas y patógenos, aunque presenta un desafío considerable en el desarrollo de terapias efectivas para trastornos cerebrales debido a su naturaleza selectiva y semipermeable. Específicamente, la permeabilidad de los compuestos en la barrera hematoencefálica emerge como un factor crítico a considerar en el desarrollo de fármacos que actúan sobre el SNC [1]. A pesar de la precisión asociada con los experimentos clínicos, su limitado alcance temporal y la demanda de recursos los posicionan como una herramienta valiosa pero no exhaustiva para medir la permeabilidad de la BHE.

El avance en tecnología de Inteligencia Artificial ha sido fundamental en el desarrollo de modelos de predicción en diversas áreas de la medicina, incluyendo el descubrimiento de fármacos y la mejora de tratamientos médicos. Un ejemplo puntual que ilustra este punto es el uso de IA en la predicción de la permeabilidad de la barrera hematoencefálica (BHE) de moléculas.

Tradicionalmente, determinar si una molécula tiene la capacidad de atravesar la BHE es un proceso costoso y laborioso que implica experimentos in vitro e in vivo. Sin embargo, gracias a los avances en IA, ahora es posible desarrollar modelos de predicción computacional que pueden predecir con alta precisión la capacidad de una molécula para cruzar la BHE utilizando datos de estructura molecular y características químicas.

Un estudio reciente publicado en la revista Frontiers in Neuroscience *"DeePred-BHE: A Blood Brain Barrier Permeability Prediction Model With Improved Accuracy*" ilustra claramente cómo los avances en inteligencia artificial están siendo aplicados para abordar un problema relevante en el campo de la medicina, específicamente en el desarrollo de fármacos dirigidos al sistema nervioso central (SNC). Este estudio demuestra cómo se utilizó un enfoque de aprendizaje profundo para predecir la permeabilidad de la barrera hematoencefálica (BHE) de compuestos químicos, una tarea decisiva en la fase inicial de diseño de fármacos, el campo de la IA ha demostrado un gran potencial en el desarrollo de modelos de predicción que pueden acelerar el descubrimiento de fármacos y mejorar la eficacia de los tratamientos [2].

El objetivo de este proyecto es desarrollar un modelo de predicción utilizando técnicas de IA, como algoritmos de aprendizaje automático para identificar y predecir las moléculas que tienen la capacidad de atravesar la barrera hematoencefálica. Este modelo se basa en un enfoque integral que combina datos experimentales y características fisicoquímicas de las moléculas para realizar predicciones precisas y confiables.

Además, se exploraron metodologías estadísticas disponibles para mejorar la precisión del modelo de predicción. Se considerarán enfoques como Random Forest y Support Vector Machine

(SVM), entre otros, para evaluar su aplicabilidad y utilidad en el contexto de este proyecto. Estas metodologías estadísticas son seleccionadas en función de su idoneidad para analizar los datos recopilados y mejorar la capacidad predictiva del modelo.

Este proyecto ha desarrollado una herramienta para la identificación y selección de moléculas con potencial para atravesar la barrera hematoencefálica. La implementación de este modelo de predicción no solo representa un avance significativo en la comprensión de la permeabilidad de la barrera hematoencefálica, sino que también sienta las bases para la innovación en el diseño y desarrollo de nuevas metodologías, así como en la integración de diferentes técnicas que combinan la ciencia de datos, la química computacional y nuevos procedimientos. Estos avances ofrecen una oportunidad para abordar desafíos clínicos en el tratamiento de enfermedades del sistema nervioso central, lo que podría tener un impacto positivo en la salud y el bienestar de los pacientes.

2. DEFINICIÓN DEL PROBLEMA

2.1 Planteamiento del problema

El sistema nervioso central (SNC) requiere un suministro continuo de oxígeno y glucosa, proporcionado por una compleja red de capilares sanguíneos. Sin embargo, este medio interno está aislado de la circulación sanguínea por la barrera hematoencefálica (BHE), una estructura única en el organismo que regula el acceso a nutrientes y protege al SNC [3-6].

La barrera hematoencefálica (BHE) es una estructura compleja y funcional que protege el SNC al regular el paso de sustancias desde la sangre hacia el cerebro. A menudo se malinterpreta como una simple barrera física, pero su comprensión es esencial para la salud cerebral y el desarrollo de tratamientos médicos. La BHE no solo es relevante para la neurología, sino también para la farmacología, biología celular y química computacional, ya que asegura que solo las sustancias necesarias y no dañinas puedan atravesarla. Conocer su composición y funcionamiento es importante para diseñar tratamientos eficaces para enfermedades del SNC [7, 8].

La salud del SNC es crucial para el bienestar humano, ya que cualquier problema en este sistema afecta la calidad de vida. Para tratar enfermedades como Alzheimer, Parkinson y esclerosis múltiple, los medicamentos deben atravesar la BHE, que dificulta el paso de fármacos desde la sangre al cerebro [9]. La capacidad de una molécula para cruzar la BHE es esencial para el desarrollo de tratamientos eficaces. Sin embargo, los métodos tradicionales para medir esta permeabilidad son costosos y lentos. Los avances en ciencia de datos y aprendizaje automático ofrecen alternativas prometedoras, permitiendo el uso de modelos predictivos basados en inteligencia artificial para identificar moléculas capaces de cruzar la BHE, optimizando así el diseño de medicamentos y reduciendo costos y tiempos de desarrollo. Es fundamental considerar la permeabilidad de la BHE dentro de un enfoque más amplio que aborde los desafíos de la entrega de terapias para diversas enfermedades [10, 11].

El desafío principal en el desarrollo de tratamientos para el sistema nervioso central (SNC) es identificar moléculas que puedan cruzar la barrera hematoencefálica (BHE). Esto requiere un enfoque detallado para comprender los mecanismos de transporte y los factores que afectan la permeabilidad, utilizando estudios experimentales in vitro e in vivo que miden la difusión, calculan coeficientes de permeabilidad y evalúan interacciones con proteínas transportadoras. Las propiedades fisicoquímicas de las moléculas, como tamaño, lipofilicidad y carga eléctrica, también son determinantes en su capacidad para atravesar la BHE [12].

Aunque la inteligencia artificial (IA) tiene el potencial de optimizar la predicción de qué moléculas pueden cruzar la BHE al analizar grandes volúmenes de datos y construir modelos precisos, enfrenta retos significativos debido a la complejidad de las interacciones moleculares y la variabilidad en las características fisicoquímicas. Esto requiere un enfoque multidisciplinario que combine ciencia de datos, biología molecular, química computacional y farmacología para superar

las limitaciones actuales y avanzar en la predicción precisa de la permeabilidad de la BHE, lo que podría facilitar el descubrimiento de nuevos fármacos y el diseño de terapias efectivas para enfermedades del SNC [13].

La predicción precisa de la capacidad de una molécula para cruzar la BHE es crítico en el diseño de tratamientos para enfermedades del sistema nervioso central. Sin embargo, la complejidad de esta barrera representa un desafío significativo para el desarrollo de fármacos, dado que los métodos experimentales tradicionales para medir la permeabilidad son lentos y costosos. Ante esto, surge la necesidad de emplear métodos computacionales basados en inteligencia artificial que, mediante el análisis de propiedades fisicoquímicas de las moléculas, permitan identificar de forma eficiente compuestos con potencial de permeabilidad. Estos avances en modelos predictivos no solo pueden optimizar los tiempos y costos de desarrollo, sino que también ofrecen una herramienta poderosa para facilitar el descubrimiento de nuevos tratamientos efectivos para el sistema nervioso central.

2.2 Formulación del problema

La barrera hematoencefálica (BHE) regula el acceso de sustancias al cerebro. Un modelo de predicción de su permeabilidad, basado en inteligencia artificial (IA) y aprendizaje automático, puede identificar moléculas capaces de cruzarla.

- ¿Cómo diseñar un modelo de inteligencia artificial que prediga qué moléculas pueden atravesar la barrera hematoencefálica?

Para abordarlo, se plantean las siguientes preguntas clave:

1. ¿Cómo preparar los datos necesarios para un modelo de predicción de la permeabilidad de la BHE?
2. ¿Qué pasos son necesarios para crear un modelo que capture y procese datos de manera precisa y confiable?
3. ¿Cuáles son las técnicas y algoritmos de IA más adecuados para analizar los datos y estimar la permeabilidad de las moléculas?
4. ¿Qué metodología de validación y evaluación se debe aplicar para medir el rendimiento y la capacidad de generalización del modelo usando conjuntos de datos externos?

Responder a estas preguntas facilitará el desarrollo de modelos más precisos y eficientes para predecir la permeabilidad de la BHE con técnicas avanzadas de IA.

3. OBJETIVOS DE PROYECTO

3.1 Objetivo General.

- Desarrollar un modelo de predicción basado en Inteligencia Artificial (IA) que permita detectar las moléculas que tienen la capacidad de atravesar la Barrera Hematoencefálica.

3.2 Objetivos Específicos

- Preparar, procesar y estandarizar las estructuras moleculares de datos para asegurar su compatibilidad en el modelo.
- Implementar un sistema integral para la adquisición y procesamiento de datos, facilitando la estimación precisa de la permeabilidad de la Barrera Hematoencefálica (BHE).
- Diseñar un modelo predictivo para identificar moléculas que atraviesan la barrera hematoencefálica con IA.
- Ejecutar una validación integral del modelo de predicción mediante la utilización de conjuntos de datos externos y métricas de evaluación pertinentes.

4. MARCO TEÓRICO Y ANTECEDENTES

4. Marco teórico

4.4.1 Barrera Hematoencefálica (BHE)

La BHE se puede definir como una propiedad funcional de los vasos sanguíneos del SNC que impide el intercambio libre de iones y moléculas orgánicas entre el plasma sanguíneo y el tejido nervioso. El concepto de la BHE surgió a finales del siglo XIX, cuando varios investigadores notaron que la inyección intravenosa de un colorante teñía todo el cuerpo excepto el cerebro y la médula espinal [3, 4, 15]. Sin embargo, al infundir el colorante en los ventrículos cerebrales, éste se difundía por el parénquima cerebral, tiñendo todo el cerebro. Lewandowsky acuñó el término "barrera hematoencefálica" en 1900, después de descubrir que un producto neurotóxico solo causaba daño cuando se inyectaba directamente en el parénquima cerebral, mientras que la inyección intravenosa del mismo producto no tenía efectos nocivos [16-18].

La propiedad de BHE se basa en la existencia de una permeabilidad muy restringida del endotelio vascular del SNC al paso de solutos plasmáticos, de modo que, excepto el agua, gases como el oxígeno y el CO2 y determinadas moléculas liposolubles muy pequeñas –menores de 400-600 Da de peso molecular–, las moléculas orgánicas no pueden atravesar libremente dicho endotelio, sino que deben hacerlo a través de sistemas de transporte específicos y finamente regulados [19].

La BHE es una estructura compleja constituida por células endoteliales de la red capilar del Sistema Nervioso Central SNC. Además, participan funcionalmente los pericitos, la lámina basal abluminal, los astrocitos perivasculares y la microglía como se ve en la figura 1. El endotelio de los capilares cerebrales se caracteriza porque cada borde celular está íntimamente unido a la célula adyacente que hace impermeable a la pared interna del capilar [20].

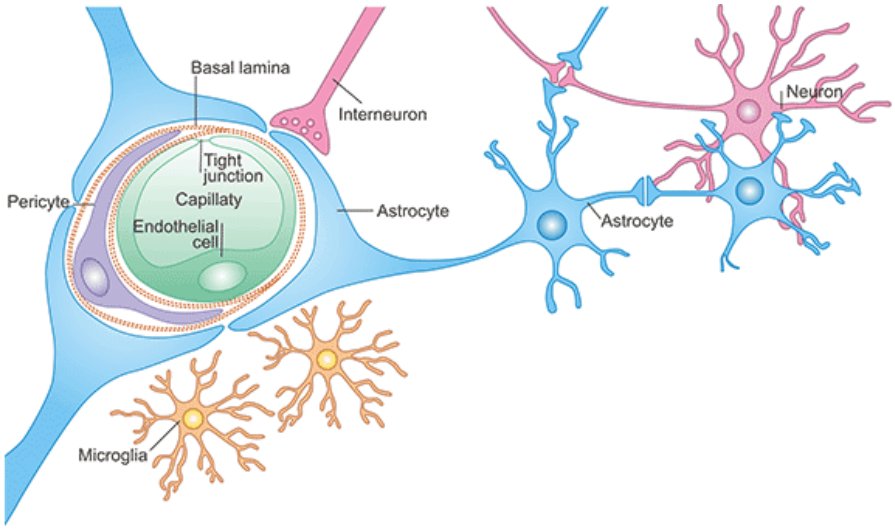


Figura 1, Barrera Hematoencefálica, creada for N. Joan Abbott, L. Rönnbäck, E. Hansson [23]

La BHE se caracteriza por la confluencia de tres componentes principales:

1. Uniones celulares endoteliales: Las células endoteliales presentan uniones estrechas con proteínas específicas intramembranales y citoplasmáticas, que están estrechamente unidas al citoesqueleto. Esta disposición única restringe la difusión de compuestos a través de las células endoteliales. Adicionalmente a las células endoteliales, la barrera presenta una membrana basal, en la cual se localizan pericitos y astrocitos, que conforman una capa que refuerza las propiedades de la barrera [21].
2. Transportadores: La BHE exhibe una alta densidad de transportadores de absorción y sobreexpresión de transportadores de secreción, junto con un transporte vesicular limitado y la ausencia de fenestraciones [22].
3. Metabolismo: La presencia de enzimas específicas en la BHE desempeña un papel transcendental en la protección del cerebro al metabolizar sustancias potencialmente dañinas [3, 4].

Todos los componentes de la BHE son indispensables para preservar su integridad estructural, funcionalidad y estabilidad. Estos elementos, al actuar de manera coordinada, no solo garantizan la cohesión arquitectónica y el correcto funcionamiento de la BHE, sino que también desempeñan un papel clave en los mecanismos de transporte selectivo que regulan el intercambio de sustancias entre la sangre y el cerebro. En conjunto, detallan la anatomía y fisiología de la BHE, así como los procesos dinámicos que sustentan su función como barrera protectora y reguladora en el sistema nervioso central [23].

Dentro del campo de la investigación sobre la BHE y el acceso al SNC, comprender y predecir la permeabilidad de esta barrera es un desafío de gran importancia. Los modelos predictivos basados en IA y aprendizaje automático han surgido como herramientas fundamentales para identificar moléculas con potencial para atravesar la BHE. Estos modelos aprovechan un amplio conjunto de descriptores moleculares y características estructurales, lo que permite una predicción precisa y eficiente de la permeabilidad de la BHE. Al hacer uso de estos modelos, es posible agilizar el proceso de investigación al seleccionar candidatos con la capacidad de atravesar la BHE y llegar al SNC. Este avance en la predicción de la permeabilidad de la BHE impulsa la investigación en neurofarmacología y permite superar las barreras biológicas que limitan la eficacia de los tratamientos para enfermedades del SNC, así como también promueve la integración de avances en ciencia de datos, química computacional y farmacología. La inclusión de técnicas in silico en este ámbito amplía nuestra capacidad para predecir con precisión qué moléculas tienen el potencial de atravesar la BHE y, Por lo tanto, contribuye significativamente al desarrollo de tratamientos más efectivos, al aprovechar los avances de la ciencia de datos y la química computacional en la predicción y selección de compuestos que puedan atravesar la barrera hematoencefálica con mayor precisión y eficacia [24-27].

4.4.2 Enfoques de aprendizaje automático para la predicción de la permeabilidad BHE

En el ámbito del estudio de la BHE, uno de los desafíos clave ha sido comprender y predecir su permeabilidad a diversas moléculas. Para abordar esta cuestión, se han desarrollado y empleado una variedad de modelos de predicción basados en técnicas de aprendizaje automático. Estos modelos, que van desde enfoques clásicos hasta técnicas más avanzadas, aprovechan el poder del análisis computacional de datos para identificar patrones y relaciones entre las características moleculares y la capacidad de las sustancias para atravesar la BHE. Desde el ampliamente utilizado Aprendizaje Automático hasta modelos más específicos como las Redes Neuronales Convolucionales y los Bosques Aleatorios, estos enfoques ofrecen una comprensión cada vez más detallada y precisa de la permeabilidad de la BHE [28-30]. Este trabajo se centra en explorar y comprender cómo estos enfoques de aprendizaje automático se aplican para predecir la permeabilidad de la BHE, proporcionando así una visión más completa de este importante fenómeno biológico.

1. Aprendizaje Automático (Machine Learning**):** Es un enfoque computacional que permite a las máquinas aprender automáticamente a través del análisis de datos y la identificación de patrones. Este método se basa en algoritmos que reconocen y utilizan la información subyacente en los datos para hacer predicciones y tomar decisiones. En el contexto de la predicción de la permeabilidad de la barrera hematoencefálica (BHE), el Aprendizaje Automático es fundamental. Permite construir modelos que analizan grandes conjuntos de datos moleculares para predecir la capacidad de las sustancias para atravesar la BHE [31].

2. Máquinas de Vectores de Soporte (SVM): Las Máquinas de Vectores de Soporte (SVM, por sus siglas en inglés, Support Vector Machines) son un poderoso modelo de aprendizaje automático supervisado que se utiliza principalmente para la clasificación y el análisis de datos. Su objetivo principal es encontrar un hiperplano óptimo en un espacio dimensionalmente elevado que pueda separar de manera efectiva los datos en diferentes categorías o clases. Este hiperplano se define como $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, dónde: **w** es el vector de pesos. **x** es el vector de características y b es el término de sesgo. El margen (M) es la distancia entre el hiperplano y los puntos más cercanos de cada clase (vectores de soporte), y se maximiza resolviendo:

$$\frac{\min}{\mathbf{w},\mathbf{b}}, \frac{1}{2} \parallel \mathbf{w} \parallel 2 \text{ sujeto a } yi(\mathbf{w} \cdot \mathbf{x}i + b) \geq 1 \forall i$$

Para datos no linealmente separables, se utiliza una función de kernel.

$$K(\mathrm{x}i, \mathrm{x}j) = \exp(-\gamma \parallel \mathrm{x}i - \mathrm{x}j \parallel 2)$$

SVM es ideal para problemas de clasificación binaria, como la predicción de moléculas que atraviesan la barrera hematoencefálica (BHE), debido a su capacidad para manejar datos no lineales y su robustez frente a overfitting [32].

3. Árboles de Decisión: Los Árboles de Decisión son modelos de aprendizaje automático utilizados en clasificación y regresión, que dividen el conjunto de datos en subconjuntos más homogéneos en términos de la variable objetivo. Su estructura jerárquica se compone de nodos internos, que representan atributos, y hojas, que corresponden a una decisión final. A medida que se desciende en el árbol, se aplican reglas de decisión basadas en valores de las características, segmentando los datos en función de su relevancia para la predicción.

El criterio para dividir los datos en los nodos se basa en la maximización de la ganancia de información:

$$IG(S, A) = H(S) - \sum_{v \in V} \frac{|S_v|}{|S|} H(S_v)$$

donde $H(S)$es la entropía del conjunto original y Sv son los subconjuntos generados tras la partición por el atributo A. [33].

4. Modelos de Regresión: Los Modelos de Regresión son herramientas esenciales en el análisis de datos, especialmente cuando se trata de predecir valores numéricos continuos en función de variables independientes. En el ámbito de la predicción de la permeabilidad de la barrera hematoencefálica (BHE), estos modelos desempeñan un papel crucial al estimar el grado de permeabilidad de una molécula a partir de sus características moleculares específicas. En este contexto, el modelo de regresión puede expresarse de manera general como:

$$\mathrm{y} = \mathrm{f}(\mathrm{X1}, \mathrm{X2}, \ldots, \mathrm{Xn}) + \epsilon$$

donde y representa la variable dependiente (permeabilidad de la BHE), $X1, X2, \ldots, Xn$ son las variables independientes (características moleculares), y ϵ es el término de error. Existen diferentes tipos de Modelos de Regresión, como la Regresión Lineal, Regresión Logística, y Regresión Polinómica, cada uno con aplicaciones específicas dependiendo del tipo de datos y la naturaleza del problema. La versatilidad de estos modelos, junto con su capacidad para revelar relaciones subyacentes entre variables, los convierte en una herramienta indispensable en la investigación biomédica y farmacéutica. Estos enfoques facilitan la optimización del diseño de fármacos con una mayor probabilidad de atravesar la barrera hematoencefálica y llegar al cerebro de manera eficaz, mejorando así el desarrollo de terapias para enfermedades del sistema nervioso central [34].

5. Redes Neuronales Artificiales (RNA): Son un modelo de inteligencia artificial inspirado en el funcionamiento del cerebro humano. Estas redes consisten en una estructura de nodos interconectados, llamados neuronas, que imitan el proceso de procesamiento y transmisión de información del cerebro. En el contexto de la predicción de la permeabilidad de la barrera hematoencefálica (BHE), las RNA son herramientas poderosas. Permiten analizar de manera eficiente las complejas relaciones entre las características moleculares y la capacidad de las sustancias para atravesar la BHE, Esta capacidad se puede modelar matemáticamente de la siguiente forma:

$$y = f(W1X1 + W2X2 + \cdots + WnXn + b)$$

donde y representa la salida (predicción de la permeabilidad de la BHE), $X1, X2, \ldots, Xn$ son las características moleculares, $W1, W2, \ldots, Wn$ son los pesos de las conexiones entre neuronas, y b es el sesgo. La función de activación f permite que el modelo aprenda relaciones no lineales entre las variables, facilitando así la identificación de moléculas potencialmente eficaces para el tratamiento de enfermedades del sistema nervioso central.lo que facilita la identificación de potenciales candidatos farmacológicos para tratar enfermedades del sistema nervioso central [35].

6. Redes Neuronales Convolucionales (CNN): Son un tipo de redes neuronales diseñadas específicamente para el procesamiento de datos estructurados en forma de cuadrícula, como imágenes. Estas redes son altamente efectivas en la extracción de características espaciales, lo que las hace ideales para analizar datos con una estructura espacial compleja. La operación básica en una CNN se describe como:

$$y = f(\sum_{ij} x_{ij} w_{ij} + b)$$

donde y es la salida de la red (predicción de la permeabilidad de la BHE), Xi, j son los valores de las características moleculares en la cuadrícula, Wi,j, son los filtros de convolución (pesos) aplicados a los datos, y b es el sesgo. La función f es una función de activación que introduce no linealidades, permitiendo a la CNN identificar patrones complejos y relaciones espaciales entre las características moleculares que afectan la permeabilidad de la BHE. Esto facilita la identificación de candidatos farmacológicos potenciales para tratar enfermedades del sistema nervioso central. En el contexto de la predicción de la permeabilidad de la BHE, las CNN pueden ser utilizadas para analizar imágenes o datos moleculares tridimensionales, ayudando así a identificar patrones y relaciones entre las características moleculares y la capacidad de las sustancias para atravesar la BHE [36].

7. K-Nearest Neighbor: Es un algoritmo de aprendizaje automático utilizado para clasificar o predecir muestras desconocidas basándose en la similitud con las muestras de entrenamiento más cercanas en un espacio de características. El parámetro "k" se refiere al número de vecinos más cercanos que se consideran al realizar la clasificación o predicción. La idea central es que la clase o el valor de la muestra desconocida se determina por la mayoría de los vecinos más cercanos en el espacio de características. Matemáticamente, esto se puede expresar como:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

En esta fórmula, y es la predicción para la muestra desconocida, calculada a partir del promedio de las etiquetas yi de los k vecinos más cercanos. El valor de k define cuántos vecinos se consideran, lo que permite hacer predicciones basadas en la similitud local de las muestras, siendo útil en clasificación y regresión. [37].

8. Redes Neuronales Recurrentes (RNN): son un tipo especializado de redes neuronales que se destacan por su capacidad para procesar secuencias de datos, como series temporales o texto. Lo que las distingue de otros tipos de redes neuronales es su capacidad para mantener y utilizar la información de entrada anterior a medida que procesan nuevas entradas. Esto se logra mediante conexiones retroalimentadas que les permiten mantener una especie de "memoria" interna. Esta característica les permite capturar dependencias temporales en los datos, lo que las hace especialmente útiles en problemas donde el orden y el contexto de la información son importantes para la predicción o clasificación, como el reconocimiento de voz, la traducción automática y la generación de texto, entre otros Matemáticamente, el funcionamiento de una RNN se expresa como:

$$h_t = f(W_h h_{t-1} + W_x x_t + b_h)$$

Dónde: ht es el estado oculto en el tiempo t, xt es la entrada en el tiempo $t, ht - 1$ es el estado oculto del tiempo anterior, Wh y Wx son matrices de pesos, bh es el sesgo y f es una función de activación, como tanh o ReLU.

9. Bosques Aleatorios (Random Forests): Los Bosques Aleatorios (Random Forests) son una técnica de aprendizaje automático que consiste en construir múltiples árboles de decisión, cada uno entrenado de forma independiente con una muestra aleatoria del conjunto de datos original. Las predicciones de cada árbol se combinan para obtener una predicción final. Este enfoque de re muestreo reduce el sobreajuste y mejora la precisión del modelo. Matemáticamente, la predicción de un bosque aleatorio se expresa como:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

Donde:

- $\hat{y}$ es la predicción final de regresión,
- $fi(x)$ es la predicción del i-ésimo árbol para la entrada x,
- N es el número total de árboles en el bosque.

La característica distintiva de los Bosques Aleatorios radica en su enfoque de re muestreo, donde cada árbol se construye utilizando una muestra aleatoria del conjunto de datos, lo que ayuda a mejorar la precisión de las estimaciones al reducir el sobreajuste y aumentar la robustez del modelo. Esto los convierte en una de las técnicas más populares en el

aprendizaje automático, especialmente cuando se requiere alta precisión en las predicciones [39].

10. Máquinas de Aprendizaje Extremo (Extreme Learning Machines, ELM): Las Máquinas de Aprendizaje Extremo (Extreme Learning Machines, ELM) son un enfoque de aprendizaje automático basado en redes neuronales de una sola capa oculta, utilizado para problemas de clasificación y regresión. A diferencia de las redes neuronales tradicionales, en ELM los pesos entre la capa de entrada y la capa oculta se asignan aleatoriamente y no se ajustan durante el entrenamiento; solo los pesos de la capa de salida se optimizan mediante un algoritmo de mínimos cuadrados. Matemáticamente, el modelo se representa como:

$$\mathrm{H}\beta = \mathrm{Y}$$

Donde:

- H es la matriz de activaciones de la capa oculta,
- β es el vector de pesos optimizados de la capa de salida,
- Y es la matriz de las etiquetas o valores esperados.

Para obtener los pesos de salida β, se utiliza la pseudo-inversa de Moore-Penrose:

$$\beta = \mathrm{H} \dagger \mathrm{Y}$$

donde H† es la pseudo-inversa de H. Este enfoque permite un entrenamiento extremadamente rápido en comparación con redes neuronales tradicionales, haciéndolo eficiente para grandes volúmenes de datos [40].

11. Redes Neuronales Profundas (Deep Neural Networks, DNN): Son modelos de aprendizaje automático formados por múltiples capas de neuronas interconectadas, lo que les permite aprender representaciones abstractas de los datos. Su capacidad de modelar relaciones complejas las hace útiles en la predicción de la permeabilidad de la barrera hematoencefálica al analizar información molecular compleja. Matemáticamente, una DNN se representa como:

$$h^{(l)} = f(W^{(l)}h^{(l-1)} + b^{(l)})$$

Dónde: $h^{(l)}$ es la activación de la capa l, $W^{(l)}$ y $b^{(l)}$ son los pesos y sesgos de la capa l, $f(\cdot)$ es la función de activación (como ReLU, sigmoide o softmax) y $H^{(0)}$es la entrada del modelo.

El modelo se entrena ajustando los pesos W y sesgos BHE mediante descenso de gradiente y retropropagación, optimizando una función de pérdida para mejorar la precisión de las predicciones [41].

Con base en la amplia gama de modelos de aprendizaje automático descritos todos estos ofrecen

herramientas clave para abordar el complejo problema de la predicción de la permeabilidad de la barrera hematoencefálica (BHE). Desde métodos clásicos como la regresión hasta enfoques avanzados como redes neuronales profundas y bosques aleatorios, estos modelos han sido fundamentales para analizar características moleculares y comprender los mecanismos de la BHE [42].

4.4.3 Fundamentos y principios de la inteligencia artificial en el análisis biomédico

La inteligencia artificial (IA) comprende un conjunto de teorías y metodologías computacionales fundamentadas en el desarrollo de sistemas que emulan procesos cognitivos humanos, como el aprendizaje automático, el reconocimiento de patrones y la toma de decisiones basada en datos. En el contexto biomédico, específicamente en el estudio de la barrera hematoencefálica (BHE), los principios teóricos de la IA se aplican mediante algoritmos de aprendizaje profundo y redes neuronales artificiales, que procesan y analizan datos moleculares complejos. La base teórica de estos modelos predictivos se sustenta en la identificación sistemática de descriptores moleculares y sus correlaciones, permitiendo establecer marcos matemáticos robustos para la predicción de la permeabilidad de nuevos compuestos. Este fundamento teórico ha transformado el paradigma tradicional del descubrimiento de fármacos, estableciendo metodologías computacionales validadas para la optimización de tratamientos dirigidos al sistema nervioso central [43].

4.4.4 Validación y métricas de desempeño en modelos de predicción de permeabilidad de la BHE

La validación de modelos de predicción de la permeabilidad de la BHE es esencial para garantizar su precisión y aplicabilidad en nuevas moléculas. Esto requiere conjuntos de datos independientes y métodos de evaluación rigurosos. Por ejemplo, Doe et al. (2020) desarrollaron modelos predictivos basados en árboles de decisión utilizando una base de datos de 153 compuestos, logrando una tasa de clasificación corregida (CCR) del 90%. Además, implementaron un análisis binario basado en Ant Colony Optimization (ACO) para aportar información mecanicista sobre el transporte a través de la BHE, subrayando la importancia de una validación sólida para la confiabilidad de estos modelos [44].

Por otra parte, los descriptores moleculares son fundamentales para estos modelos, ya que capturan características estructurales y fisicoquímicas críticas de las moléculas. Su selección cuidadosa permite identificar propiedades clave que influyen en la permeabilidad de la BHE. Mediante técnicas avanzadas de ciencia de datos, como el aprendizaje automático, se aprovecha esta información para mejorar la precisión de las predicciones y profundizar en los mecanismos subyacentes de la permeabilidad [45].

Además de su aplicación en el descubrimiento de fármacos dirigidos al SNC, los modelos de predicción de permeabilidad de la BHE tienen otras aplicaciones en el desarrollo farmacéutico. Estos modelos pueden ayudar en la optimización de la administración de fármacos existentes, así

como en el diseño de nuevas estrategias de administración para mejorar la entrega de fármacos al cerebro [46].

A pesar de los avances en ciencia de datos y química computacional, todavía existen limitaciones en la capacidad de modelar con precisión la permeabilidad de la barrera hematoencefálica. La transición de la realidad a modelos computacionales presenta desafíos significativos, lo que dificulta la predicción precisa de qué moléculas pueden atravesar la barrera hematoencefálica y cuáles no. Estos desafíos representan oportunidades para futuras investigaciones en este campo [47].

Al considerar las métricas de desempeño, es esencial evaluar de igual forma la efectividad y el rendimiento del modelo de predicción. Algunas métricas comunes utilizadas en el contexto de modelos de aprendizaje supervisado incluyen:

- **Precisión (Accuracy):** Es la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones. Se calcula dividiendo el número de predicciones correctas entre el número total de muestras. Es una métrica general que se utiliza para evaluar el rendimiento global del modelo [48].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

  Donde:

  1. TP: Verdaderos positivos
  2. TN: Verdaderos negativos
  3. FP: Falsos positivos
  4. FN: Falsos negativos

- **Precisión (Precision):** También conocida como valor predictivo positivo, es la proporción de verdaderos positivos (muestras clasificadas correctamente como positivas) en relación con el total de muestras clasificadas como positivas. Se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos positivos. Esta métrica es útil cuando el costo de los falsos positivos es alto [49].

$$\text{Precision} = \frac{TP}{TP+FP}$$

  Donde:

  1. TP: Verdaderos positivos
  2. FP: Falsos positivos

- **Sensibilidad (Recall o True Positive Rate):** Es la proporción de verdaderos positivos en relación con el total de muestras positivas. Se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos. Esta métrica es útil cuando el costo de los falsos negativos es alto [50].

$$\text{Recall} = \frac{TP}{TP+FN}$$

Donde:

1. TP: Verdaderos positivos
2. FN: Falsos negativos

- **Especificidad (Specificity):** También conocida como True Negative Rate, es la proporción de verdaderos negativos (muestras clasificadas correctamente como negativas) en relación con el total de muestras clasificadas como negativas. Se calcula dividiendo el número de verdaderos negativos entre la suma de verdaderos negativos y falsos positivos. Esta métrica es útil cuando el costo de los falsos positivos es alto [51].

$$\text{Specificity} = \frac{TN}{TN+FP}$$

Donde:

1. TN: Verdaderos negativos
2. FN: Falsos positivo

- **Valor F1 (F1 Score):** Es una métrica que combina precisión y sensibilidad, calculando la media armónica entre ambas. El valor F1 es útil cuando se desea encontrar un equilibrio entre la precisión y la sensibilidad [52].

$$\text{F1 Score} = 2 \text{ X} \frac{\text{Precision x Recall}}{\text{Precision + Recall}}$$

Donde:

1. Precisión: Precisión
2. Recall: Sensibilidad

Además, es clave considerar métricas como el área bajo la curva ROC (AUC-ROC), que evalúa la capacidad discriminativa en distintos umbrales, y el error cuadrático medio (MSE), útil para medir la precisión de predicciones probabilísticas. Estas métricas complementan el análisis en problemas de clasificación y regresión, ofreciendo una evaluación más detallada y completa

4.4.5 Enfoques Estadísticos y Farmacocinéticas para la Evaluación de Candidatos a Fármacos y su Potencial de Penetración de la Barrera Hematoencefálica

En el ámbito de la predicción de la permeabilidad molecular, resulta fundamental emplear diversos enfoques basados en modelos estadísticos para examinar la relación entre las variables predictoras y la permeabilidad. Estos modelos brindan una poderosa herramienta para analizar y comprender los factores que influyen en la capacidad de una molécula para atravesar la barrera.

Abordando la complejidad inherente de los datos y realizando una evaluación rigurosa, los modelos estadísticos proporcionan una perspectiva valiosa para la comprensión y predicción de la permeabilidad molecular, permitiendo explorar una amplia gama de opciones sin limitarse a un único modelo en el presente estudio. A continuación, se presentan algunos de estos modelos junto con sus breves conceptos:

- **Regresión Lineal:** Modelo estadístico que busca establecer una relación lineal entre una variable dependiente y una o más variables independientes, utilizado para predecir valores numéricos continuos [53].

- **Regresión Logística:** Modelo estadístico utilizado cuando la variable dependiente es categórica o binaria, estima la probabilidad de pertenecer a una categoría en función de las variables independientes [54].

- **Análisis de Componentes Principales (ACP):** Técnica utilizada para reducir la dimensionalidad de un conjunto de variables correlacionadas, creando nuevas variables no correlacionadas llamadas componentes principales que capturan la mayor parte de la variabilidad de los datos [55].

- **QED (Quantitative Estimate of Drug-likeness):** es una herramienta fundamental que evalúa cuán similar es una molécula a un fármaco conocido, mediante el análisis de múltiples propiedades moleculares como peso molecular, logP, características de enlaces de hidrógeno y grupos funcionales específicos. Su principal fortaleza radica en la capacidad de integrar todas estas propiedades en un único valor entre 0 y 1, donde valores cercanos a 1 indican mayor similitud con fármacos existentes, permitiendo así filtrar eficientemente bibliotecas de compuestos y priorizar moléculas con mayor potencial de éxito en el desarrollo farmacéutico, optimizando recursos y tiempo en el proceso de descubrimiento de nuevos fármacos. [56].

- **Screening Molecular:** El screening o cribado constituye un proceso metodológico de evaluación y selección sistemática que permite filtrar y clasificar elementos según criterios específicos predefinidos. Esta técnica se fundamenta en principios de análisis secuencial, donde se establecen parámetros y puntos de corte que permiten discriminar entre elementos deseables y no deseables dentro de una población objetivo [57].

- **Lipinski:** La "regla de 5" de Lipinski establece criterios clave para predecir la absorción oral y permeabilidad de fármacos potenciales: no más de 5 donantes y 10 aceptores de enlaces de hidrógeno, peso molecular menor a 500 Da, y un coeficiente de partición (LogP) menor a 5. Esta regla empírica se ha convertido en una herramienta fundamental en el descubrimiento temprano de fármacos para evaluar rápidamente la "drug-likeness" de nuevos compuestos candidatos [58].

- **Bayes Ingenua (Naive Bayes):** Modelo de clasificación basado en el teorema de Bayes, que asume independencia condicional entre las variables predictoras, simplificando el cálculo de las probabilidades condicionales y facilitando la clasificación de nuevas observaciones [59].

- **Rules Veber:** Constituyen un enfoque complementario a la regla de Lipinski que se centra en la flexibilidad molecular y la superficie polar de los compuestos como predictores claves de la biodisponibilidad oral [60].

- **BBB Score" (Blood-Brain Barrier Score):** Es una métrica establecida para evaluar la probabilidad de penetración cerebral de compuestos a través de la barrera hematoencefálica mediante el análisis de parámetros fisicoquímicos específicos [61].

En el campo de la ciencia de datos y la inteligencia artificial, el análisis estadístico es clave para predecir la permeabilidad molecular. La exploración de diversos enfoques nos permite comprender y predecir con precisión la capacidad de las moléculas para atravesar barreras. Cada uno de estos modelos desempeña un papel fundamental en el análisis riguroso de los datos de permeabilidad, proporcionando una perspectiva integral que facilita la selección del modelo más adecuado para cada caso, lo que permite obtener resultados significativos en el emocionante campo de la predicción de la permeabilidad molecular.

4.5 Antecedentes

La predicción precisa de la permeabilidad de la barrera hematoencefálica (BHE) es esencial para el desarrollo de terapias efectivas para enfermedades del sistema nervioso central. En este sentido, los avances en ciencia de datos y aprendizaje automático han permitido la creación de modelos más sofisticados. El empleo de técnicas como redes neuronales, algoritmos genéticos y máquinas de vectores de soporte ha sido fundamental. Integrar estos enfoques de manera inteligente promete mejorar significativamente las predicciones y abrir nuevas posibilidades en la investigación de la BHE.

Recent Studies of Artificial Intelligence on In Silico Drug Distribution Prediction trabajo el cual se basa en demostrar que La barrera hematoencefálica (BHE) es fundamental para que los medicamentos lleguen al sistema nervioso central (SNC). Los fármacos dirigidos al SNC deben atravesar la BHE, mientras que los que actúan en objetivos periféricos pueden evitarla para evitar efectos adversos. Se utiliza la ratio logarítmica de las concentraciones de un fármaco en el cerebro y en la sangre para medir la permeabilidad de la BHE. Se han utilizado métodos de inteligencia artificial (IA) y aprendizaje automático (ML) para predecir la permeabilidad de la BHE, logrando una precisión superior al 80% en algunos algoritmos. Recientemente, los enfoques basados en IA se han centrado en clasificar si un compuesto es permeable (BHE+) o no (BHE-) en lugar de utilizar el ratio log BB. También se han desarrollado algoritmos de aprendizaje profundo (DL) con excelentes resultados en la predicción de la permeabilidad de la BHE. Sin embargo, estos modelos

carecen de interpretabilidad, por lo que se han propuesto métodos híbridos que combinan ML y DL para obtener reglas simples y precisas. Es importante utilizar conjuntos de datos confiables y lo suficientemente grandes para construir modelos precisos de predicción de permeabilidad de la BHE [62].

A Recurrent Neural Network model to predict blood–brain barrier permeability," Computational Biology and Chemistry muestra que los modelos eficientes de predicción de la permeabilidad de la BHE son cruciales en el desarrollo temprano de fármacos. Se han realizado esfuerzos persistentes para identificar la permeabilidad de la BHE mediante métodos de aprendizaje automático, con el objetivo de reducir la eliminación de candidatos a fármacos en ensayos preclínicos y clínicos. Sin embargo, es necesario revisar urgentemente el progreso de estos modelos de predicción basados en el aprendizaje automático para la permeabilidad de la BHE [63]

DeePred-BHE: A Blood Brain Barrier Permeability Prediction Model With Improved Accuracy, en este artículo se aborda la importancia de predecir la permeabilidad de la barrera hematoencefálica (BHE) en el desarrollo de fármacos para el sistema nervioso central. Se ha utilizado aprendizaje automático y aprendizaje profundo en un conjunto de datos de 3,605 compuestos para mejorar la precisión de la predicción. El modelo seleccionado, denominado "DeePred-BHE", utiliza redes neuronales profundas y notaciones moleculares simplificadas (SMILES) para predecir la permeabilidad de la BHE. Este modelo puede ser útil en la selección de compuestos en las etapas iniciales del desarrollo de fármacos [64].

El rápido desarrollo de métodos computacionales y la creciente disponibilidad de grandes volúmenes de datos han fomentado la expansión de la investigación en química computacional. Dentro del campo de la quimioinformática, las técnicas de aprendizaje automático se aplican para analizar y predecir propiedades de estructuras químicas. Un área de estudio relevante en este ámbito es la permeabilidad de la barrera hematoencefálica (BHE), la cual se centra en evaluar la capacidad de diversas moléculas para acceder al sistema nervioso central (SNC). Estudios previos, como el desarrollado en "A Recurrent Neural Network model to predict blood–brain barrier permeability," Computational Biology and Chemistry, han demostrado el potencial de las Redes Neuronales Recurrentes (RNN) para mejorar la precisión en la predicción de la permeabilidad de la BHE. Este modelo en particular alcanzó una precisión del 96.53% y una especificidad del 98.08%, superando modelos tradicionales. Estos avances resaltan la efectividad de los métodos de aprendizaje profundo y subrayan la importancia de continuar desarrollando herramientas predictivas para optimizar el diseño de fármacos orientados al SNC [65].

5 DESARROLLO DEL PROYECTO

5.1 Metodología

En esta sección se presenta la metodología aplicada en cuatro etapas clave para lograr los objetivos propuestos en este proyecto. El primero de estos objetivos es la preparación, procesamiento y estandarización de datos, incluyendo la validación de las estructuras moleculares que formarán la base del análisis. A continuación, se procedió al desarrollo de un sistema integral de captura y procesamiento de datos, diseñado para estimar la permeabilidad de la barrera hematoencefálica. Posteriormente, se diseñó un modelo predictivo, empleando técnicas de inteligencia artificial, con el fin de identificar moléculas capaces de atravesar dicha barrera. Finalmente, se llevó a cabo una validación del modelo predictivo utilizando conjuntos de datos externos y métricas de evaluación apropiadas, garantizando la robustez y precisión del sistema.

1. Preparación y procesamiento de datos: En el primer objetivo, se realizó la preparación y procesamiento de los datos moleculares para construir una base sólida que permita el análisis de la permeabilidad de la barrera hematoencefálica (BHE). se utilizó la base de datos de código abierto Blood-Brain Barrier Database (B3DB), que contiene 7,807 moléculas con 1,058 valores de coeficiente de partición octanol-agua (logBB) y etiquetas de permeabilidad BHE+/BHE-, se llevó a cabo una verificación de la distribución de las clases, la validación de las estructuras moleculares representadas en formato SMILES, y la identificación de datos nulos. Adicionalmente, se aplicaron técnicas de estandarización, limpieza y enriquecimiento de los datos moleculares, como la transformación a SMILES isoméricos, para garantizar la calidad de los datos utilizados en los análisis predictivos. Como parte del proceso de preparación y procesamiento de datos, se realizó una doble validación de las estructuras moleculares representadas por las cadenas SMILES (Simplified Molecular Input Line Entry System) utilizando dos enfoques complementarios. Esto permitió minimizar posibles errores en la representación de las estructuras químicas y asegurar la coherencia de los datos antes de su análisis. Además, para enriquecer la base de datos y maximizar la cantidad de información disponible para los modelos predictivos, se llevó a cabo un proceso de generación y cálculo de descriptores moleculares. Para ello, se empleó la calculadora de descriptores moleculares Mordred en conjunto con bibliotecas especializadas como RDKit**,** ChEMBL Structure y otras herramientas de química computacional. Gracias a este enfoque, se logró ampliar la base de datos a más de 700 descriptores moleculares, abarcando un amplio espectro de propiedades fisicoquímicas, topológicas, estructurales y geométricas. Entre los descriptores calculados se incluyeron aquellos relacionados con reglas de filtrado para bioactividad y farmacocinética, permitiendo evaluar características esenciales para el desarrollo de fármacos, como la solubilidad, la lipofilicidad, la presencia de fragmentos estructurales relevantes y otras propiedades clave en la predicción de la permeabilidad de la BHE. De esta manera, el conjunto de datos se enriqueció significativamente, proporcionando una base robusta para la construcción de modelos de aprendizaje automático capaces de discriminar entre

moléculas permeables e impermeables a la barrera hematoencefálica con mayor precisión.

2. Implementación un sistema integral para la adquisición y preprocesamiento de datos: Este proceso se estructuró en tres etapas secuenciales, diseñadas para explorar la relación entre la estructura molecular, las propiedades fisicoquímicas y la permeabilidad de la Barrera Hematoencefálica (BHE) desde múltiples perspectivas analíticas. La división en fases responde a la necesidad de abordar el problema con un enfoque jerárquico: desde la descripción estructural básica (SMILES) hasta la evaluación farmacológica de vanguardia, permitiendo identificar patrones para la predicción de la permeabilidad. La primera se enfocó en el análisis de las estructuras SMILES para identificar atributos estructurales distintivos entre moléculas permeables y no permeables. Mediante bibliotecas de química computacional (RDKit), se cuantificaron variables categóricas clave, Estos datos, aunque descriptivos y no determinantes por sí solos, ofrecieron una base estadística para diferenciar moléculas que atraviesan y no la BHE. Seguido de la etapa de Evaluación de descriptores moleculares, esta fase integró técnicas computacionales y estadísticas para analizar descriptores fisicoquímicos para evaluar su relación con la permeabilidad. Se aplicaron reglas de drug-likeness (Lipinski, Ghose), coeficientes de similitud (Tanimoto) y análisis de correlación entre propiedades clave como el logP, la superficie polar accesible (PSA) y la solubilidad. Modelos de clasificación preliminares (Random Forest, SVM entre otros) en relación a la variable 'BBB+/BBB- que permitieron identificar descriptores moleculares interesantes, finalmente la última etapa de esta de este enfoque permitió priorizar candidatos con equilibrio entre propiedades farmacocinéticas y permeabilidad, identificando moléculas con perfiles similares a fármacos. Se enfatizó que la permeabilidad BHE no depende de un único factor si no de diferentes factores lo que demuestra la complejidad de esta. La triangulación de estas etapas generó un conjunto de datos enriquecido, donde las propiedades categóricas (SMILES) se contextualizaron con descriptores cuantitativos y criterios farmacológicos. Esto permitió superar las limitaciones de enfoques unidimensionales, ofreciendo una base sólida para entrenar modelos predictivos de permeabilidad BHE basados en estructura molecular.

3. Diseño de un modelo predictivo: Luego de la selección de los descriptores moleculares más relevantes para la permeabilidad de la Barrera Hematoencefálica BHE a partir de múltiples análisis y pruebas de predicción de las etapas anteriores, evaluando su impacto en la capacidad predictiva y optimizando el rendimiento del modelo. Finalmente, se implementó un sistema de predicción basado en estos descriptores, permitiendo estimar con precisión la permeabilidad de nuevas moléculas y facilitando su aplicación en el desarrollo de fármacos dirigidos al sistema nervioso central. Para el desarrollo del modelo predictivo, se seleccionaron seis algoritmos de aprendizaje supervisado: cuatro de machine learning (Random Forest, SVM, Árboles de Decisión y Naïve Bayes) y dos de deep learning (Redes Neuronales Profundas - DNN y Redes Neuronales Convolucionales - CNN).

Estos modelos fueron elegidos por su capacidad para capturar patrones complejos en los descriptores moleculares y su robustez en la clasificación de moléculas permeables e impermeables a la Barrera Hematoencefálica BHE. Su desempeño fue evaluado comparativamente para identificar el enfoque más preciso y generalizable en la predicción de permeabilidad, conforme a la cantidad de moléculas que estos pudiesen detectar, siendo el modelo de Random Forest el superior a todos. Como evidencia de la implementación de modelos con diferentes arquitecturas, se muestran tres modelos predictivos: SVM, Random Forest y el clasificador basado en optimización de hiperparámetros. Los parámetros relevantes para cada modelo son los siguientes:

Para el modelo SVM:

- C: Parámetro de regularización, controla el balance entre margen máximo y la clasificación de los puntos de datos, optimizado en un rango logarítmico de $1e-3$ a $1e3$.
- gamma: Parámetro del núcleo que define la influencia de un solo punto de entrenamiento, optimizado en un rango logarítmico de $1e-4$ a 1.
- kernel: Función del núcleo utilizada en SVM para transformar los datos, optimizada entre RBF (Radial Basis Function) y sigmoid.
- random_state: Establecido en 42 para garantizar la reproducibilidad de los resultados.
- probability: Establecido en True para habilitar las probabilidades de predicción, necesarias para calcular la curva ROC.

Estos parámetros fueron optimizados utilizando Optuna, con el objetivo de seleccionar la mejor combinación de hiperparámetros para maximizar el rendimiento del modelo en la tarea de clasificación de permeabilidad BBB.

Para el modelo Random Forest:

- n_estimators: Número de árboles en el bosque, optimizado mediante validación cruzada.
- max_depth: Profundidad máxima de los árboles, ajustada para evitar el sobreajuste.
- min_samples_split: Número mínimo de muestras necesarias para dividir un nodo, optimizado en función del rendimiento.
- random_state: Establecido en 42 para asegurar reproducibilidad.
- max_features: Número máximo de características consideradas para cada división, ajustado para mejorar la precisión del modelo.

Este modelo fue entrenado utilizando Optuna para optimizar hiperparámetros, y se emplearon técnicas de Preprocesamiento como el balanceo de clases, la normalización de las características y la división del conjunto de datos en entrenamiento y prueba.

los parámetros óptimos, seleccionado por su capacidad para manejar relaciones no lineales, reducir overfitting mediante bagging, y cuantificar importancia de características mediante disminución media de impureza Gini.

Para el Modelo 3 - DNN (Deep Neural Network):

1. Preprocesamiento de Datos:
   - Imputación de valores faltantes: Se utiliza SimpleImputer con estrategia 'mean' para rellenar los valores faltantes con la media de cada columna.
   - Escalado de características: Se usa StandardScaler para escalar las características numéricas.
2. Optimización de Hiperparámetros (Optuna):
   - Número de capas (n_layers): Se optimiza entre 1 y 3 capas ocultas.
   - Unidades por capa: Para cada capa oculta, se optimiza el número de neuronas entre 32 y 256.
   - Tasa de deserción (dropout_rate): Se optimiza entre 0.1 y 0.5.
   - Tasa de aprendizaje (learning_rate): Se optimiza con una distribución logarítmica entre 1e-5 y 1e-2.
3. Arquitectura de la Red Neuronal:
   - Capas ocultas: El número de capas y las unidades por capa son determinadas por los resultados de la optimización.
   - Capa de salida: La capa de salida tiene una sola neurona con activación sigmoid para clasificación binaria (si un compuesto atraviesa o no la barrera BBB).
   - Función de pérdida: binary_crossentropy, adecuada para clasificación binaria.
   - Optimización: Adam con la tasa de aprendizaje determinada por la optimización.
4. Entrenamiento de la Red Neuronal:
   - Épocas: El modelo se entrena por 50 épocas.
   - Tamaño de lote: Se usa un tamaño de lote de 32.
5. Balanceo de Clases:
   - Si las clases están desbalanceadas, el modelo realiza un submuestreo de la clase mayoritaria para equilibrar la distribución de clases en el conjunto de datos de entrenamiento.
6. Evaluación y Resultados:
   - Matriz de confusión: Para visualizar la exactitud de las predicciones del modelo.
   - Curva ROC: Para evaluar la capacidad de discriminación del modelo.
   - Reporte de clasificación: Incluye métricas como precisión, recuperación (recall), f1-score y soporte.

   Uso de Optuna: La optimización bayesiana a través de Optuna fue utilizada para realizar la búsqueda de hiperparámetros, evaluando la precisión del modelo para cada combinación de parámetros.

Para garantizar un desarrollo riguroso del modelo predictivo, se realizó un proceso sistemático de ajuste y configuración de los algoritmos seleccionados. Se aplicó optimización de hiperparámetros para cada modelo, ajustando parámetros clave como la profundidad de los árboles en métodos basados en reglas, los valores de regularización en modelos lineales y la arquitectura en redes neuronales profundas. El entrenamiento se llevó a cabo utilizando validación cruzada para evitar sobreajuste y garantizar la generalización del modelo. Además, se emplearon métricas de rendimiento como precisión, recuperación y AUC-ROC para evaluar y comparar el desempeño de cada enfoque. La selección final del modelo óptimo se basó en su capacidad predictiva y estabilidad frente a variaciones en los datos.

Después de evaluar todos modelos predictivos, se concluye que el modelo basado en Random Forest es el más efectivo para predecir las moléculas capaces de atravesar la barrera hematoencefálica BHE. Esto se debe a su robustez y capacidad para manejar grandes cantidades de datos y variables, así como su habilidad para identificar las relaciones complejas entre las características moleculares y la permeabilidad de la BHE. Random Forest demostró ser el más preciso en la predicción de las moléculas que logran atravesar esta barrera, destacándose entre los demás modelos analizados.

4. Ejecutar una validación integral del modelo de predicción mediante la utilización de conjuntos de datos externos y métricas de evaluación pertinentes: Para la validación integral del modelo de predicción, se siguió un proceso que implicó varios pasos. En primer lugar, se utilizaron nuevos conjuntos de datos, distintos a los utilizados durante el entrenamiento y optimización del modelo, con el fin de probar su capacidad de generalización y rendimiento en situaciones externas. Estos conjuntos de datos fueron cuidadosamente seleccionados, asegurando que representaran moléculas relevantes para la predicción de permeabilidad a la BHE, pero que no hubieran sido vistas durante el proceso de entrenamiento. Posteriormente, se replicaron las operaciones previas realizadas en el modelo original, con el objetivo de enriquecer la base de datos y mejorar la calidad de las predicciones. Esto incluyó la limpieza de los datos, la estandarización de las características, además de la aplicación de las mismas transformaciones que fueron exitosas durante la fase de entrenamiento. Una vez enriquecida la base de datos, se procedió a la aplicación del modelo a los nuevos conjuntos de datos. Este paso implicó la predicción de la permeabilidad de las nuevas moléculas a la BHE, para lo cual se evaluó la precisión del modelo en función de la cantidad de moléculas predichas que efectivamente lograron atravesar la BHE. Para evaluar el desempeño del modelo, se utilizaron métricas de evaluación pertinentes, como la precisión, recuperación, F1-score y área bajo la curva ROC, las cuales permitieron medir la efectividad de las predicciones y la capacidad del

modelo para generalizar a datos no vistos. Finalmente, con base en los resultados obtenidos, se realizaron los ajustes y mejoras necesarias al modelo. De esta manera, se garantizó que el modelo mantuviera un alto rendimiento y fuera capaz de predecir con precisión las moléculas con potencial farmacológico que atraviesan la BHE, lo que es crucial para su futura aplicación en el desarrollo de tratamientos dirigidos al sistema nervioso central

Se presenta a continuación un diagrama de flujo que resume las etapas clave de la metodología utilizada en este proyecto.

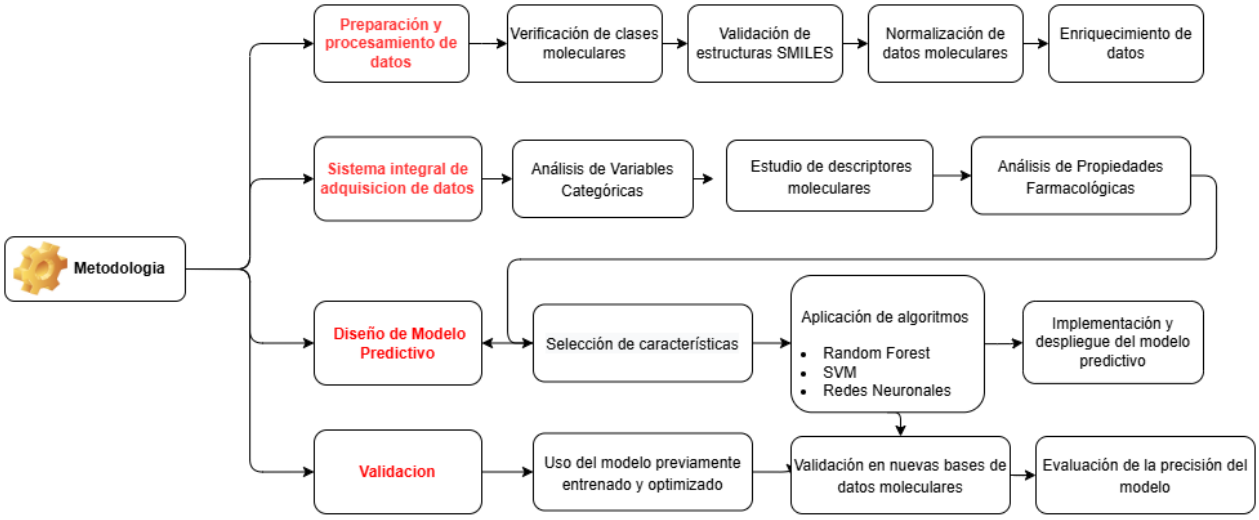


Figura 2, Metodología utilizada. Fuente: Elaboración propia

6. RESULTADOS

6.1 Preparación y procesamiento de datos

Tras una revisión exhaustiva de la literatura para identificar los factores determinantes en la permeabilidad de la barrera hematoencefálica (BHE), se utilizó la base de datos de código abierto Blood-Brain Barrier Database (B3DB), la cual representa uno de los conjuntos de datos más grandes y completos sobre permeabilidad de la BHE. Compilada a partir de 50 fuentes publicadas (véase anexo I), esta base de datos se estructura en función de la consistencia entre diferentes mediciones experimentales, lo que permite una mayor precisión en el análisis. B3DB proporciona valores numéricos de logBB para 1,058 compuestos, mientras que el conjunto de datos completo ofrece etiquetas categóricas de permeabilidad (BBB+, molécula permeable a la BHE o BBB−, molécula no permeable a la BHE) para un total de 7,807 moléculas. Este recurso es una fuente que supera las limitaciones de conjuntos de datos más reducidos y permite un análisis de la permeabilidad en este estudio. Esto posiciona a B3DB como una fuente clave para este análisis, superando las limitaciones de conjuntos de datos más pequeños y con diversidad química reducida utilizados en estudios anteriores. Esta base de datos, fue meticulosamente compilada, abarca una amplia gama de variables clave, como Simplified Molecular Input Line Entry System (SMILES) el cual es un formato de texto utilizado para representar la estructura química de moléculas de manera sencilla y legible, que son fundamentales para comprender la capacidad de las moléculas para atravesar la BHE.

La fase inicial del estudio fue la preparación y procesamiento de los datos, pasos clave para el desarrollo de un modelo de predicción, como parte de este proceso, se elaboró la tabla I, donde se presentan las variables iniciales de la base de datos junto con sus abreviaciones y descripciones. Los descriptores seleccionados proporcionan información esencial para evaluar y comprender la interacción entre las moléculas y la BHE, lo que constituye una base sólida para el análisis de la permeabilidad en el contexto de la investigación sobre fármacos dirigidos al sistema nervioso central. Esta base de datos se deriva del estudi*o A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors*, destacando su importancia como punto de partida para nuestro análisis y proporcionando un marco sólido para nuestro enfoque de investigación.

Tabla I. Variables y descripciones. Fuente: Elaboración propia.

Variable	**Descripción**
N°	Número de identificación
name	Nombre de la Molécula
IUPAC_name	Nombre según la Unión Internacional de Química Pura y Aplicada (IUPAC)
SMILES	Simplified Molecular Input Line Entry System (SMILES)
CID	Chemical Identifier
logBB	Coeficiente de partición octanol-agua (logBB)
BHE+/BHE-	Barrera hematoencefálica (BHE) Atraviesa/No atraviesa

Inchi	Clave internacional normalizada de la InChI (IUPAC International Chemical Identifier)
threshold	Umbral
reference	Referencia temporal
group	Grupo al que pertenece la molécula

Análisis visual de la distribución de una variable categórica BHE+/BHE-

El análisis de datos comenzó con la evaluación de la distribución de la variable categórica BHE+/BHE-, que identifica si las moléculas atraviesan o no la barrera hematoencefálica. Se calcularon estadísticas descriptivas y la proporción de cada clase, revelando que el 63.48% de las moléculas (4,956) atraviesan la barrera (BHE+), mientras que el 36.52% (2,851) no lo hacen (BHE-). Estos resultados se presentan en la figura 2, ofreciendo una visión clara de la prevalencia de cada categoría en el conjunto de datos.

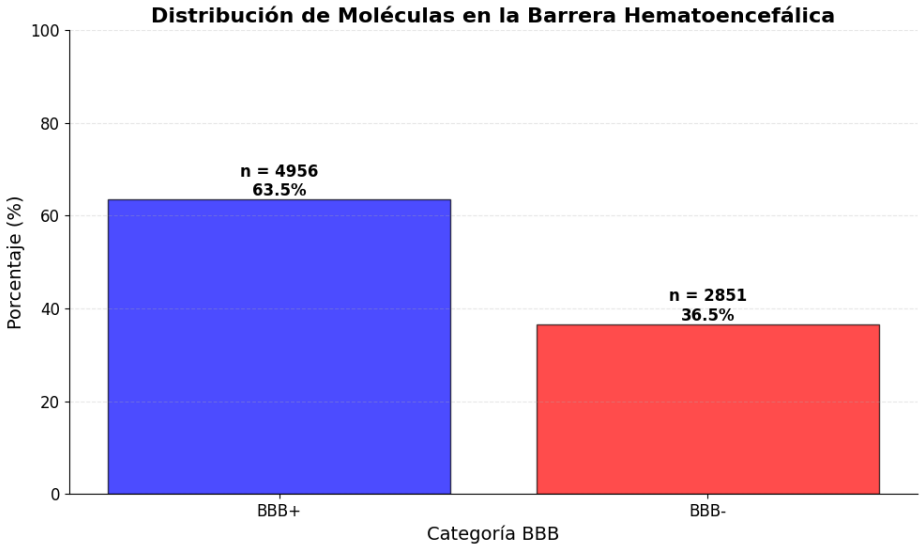


Figura 3 , Distribución de variable BHE+/ BHE-. Fuente: Elaboración propia.

Verificación de datos nulos

El siguiente paso en el proceso de preparación de datos fue la verificación de valores nulos en las distintas variables del conjunto de datos. Se utilizó un análisis visual para identificar la cantidad de datos faltantes por variable. A través de un gráfico de barras como se evidencia en la figura 3, se mostraron las variables junto con el número de valores nulos correspondientes, facilitando la identificación de cualquier problema significativo de datos incompletos que pudiera afectar la calidad del análisis posterior. Esta visualización proporcionó una visión clara y rápida del estado de los datos.

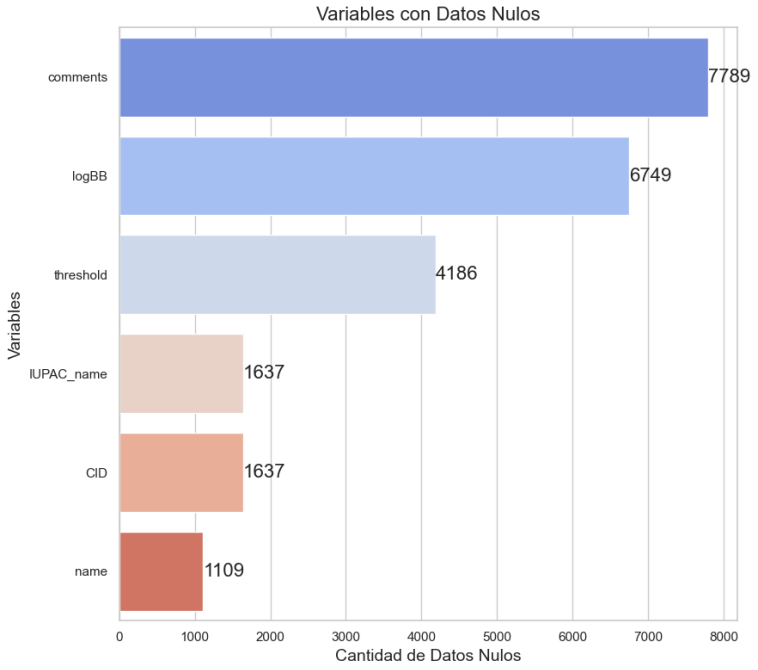


Figura 4, Datos nulos por variable. Fuente: Elaboración propia

Verificación de SMILES (Simplified Molecular Input Line Entry System)

Como parte del proceso de preparación y procesamiento de datos, se realizó una doble validación de las estructuras moleculares representadas por las cadenas SMILES (Simplified Molecular Input Line Entry System) utilizando dos enfoques. Primero, se implementó una función con la biblioteca RDKit para verificar la validez de los SMILES. Luego, se utilizó una combinación de las bibliotecas RDKit y chembl_structure_pipeline para validar las moléculas, asegurando que las representaciones fueran correctas, consistentes y compatibles con las herramientas de química computacional.

De un total de 7807 moléculas, para el segundo enfoque se diseñó una función que estandariza cada molécula en la base de datos, corrigiendo posibles problemas estructurales mediante operaciones como la normalización y la desionización. Durante la iteración sobre los datos, se comprobó si existían diferencias entre los SMILES originales y sus versiones estandarizadas. El resultado de esta validación reveló que no se encontraron moléculas problemáticas, ya que todas las estructuras fueron estandarizadas correctamente. Este proceso garantiza que la base de datos utilizada para el análisis sea precisa y confiable, permitiendo una base sólida para la fase de modelado predictivo, para ambos procesos se encontraron que 4956 eran válidas y atravesaban la barrera hematoencefálica (BHE+), mientras que 2851 eran válidas, pero no la atravesaban (BHE).

Clasificación y análisis de tipos de SMILES

La clasificación de los SMILES en la base de datos reveló una distribución variada entre los

diferentes tipos de representaciones. De un total de 7,807 SMILES procesados, el 52.89% correspondió a SMILES isoméricos, que son representaciones que incluyen información sobre la estereoquímica de las moléculas. El 42.19% de los SMILES fueron clasificados como canónicos, representaciones estándar sin estereoquímica. Además, el 4.60% de las representaciones fueron no estándar correspondiente a 384. Estos resultados proporcionaron una visión detallada sobre la diversidad de SMILES en el conjunto de datos y ayudan a entender mejor la complejidad de las representaciones moleculares utilizadas en el análisis. A su vez la observación de los tipos de SMILES en relación con la capacidad de atravesar la barrera hematoencefálica (BHE) revela diferencias significativas.

Considerando la información proporcionada sobre los registros de CID, nombres IUPAC, y su clasificación como isoméricos o canónicos: El análisis de los 7807 SMILES en la base de datos revela que 384 de ellos son clasificados como no estándar, lo que indica que no pudieron ser reconocidos como isoméricos o canónicos debido a diversos factores estructurales o de representación. Estos SMILES no estándar incluyen registros detallados como el CID, el nombre IUPAC y la clave InChI, lo que proporciona información valiosa sobre su identidad química. Sin embargo, a pesar de contar con esta información, la complejidad estructural y las variaciones en la representación pueden haber impedido su clasificación adecuada. Por ejemplo, se identificaron 4956 moléculas válidas que atraviesan la barrera hematoencefálica (BBB+) y 2851 que no lo hacen (BBB-), sugiriendo que las características estructurales juegan un papel importante en su permeabilidad. Además, el conteo de combinaciones específicas de elementos químicos en los SMILES muestra la presencia de halógenos y grupos funcionales relevantes, lo que podría influir en sus propiedades bioquímicas y farmacológicas. Este estudio destaca la importancia de una representación precisa y estandarizada en la química computacional para facilitar el análisis y la predicción de propiedades moleculares. Este párrafo sintetiza los hallazgos clave y proporciona un contexto sobre por qué ciertos SMILES no se clasifican adecuadamente

Limpieza y estandarización de los datos moleculares:

La operación de limpieza de datos implica la detección, corrección y eliminación de datos erróneos o irrelevantes. En este caso, se ha implementado un pipeline para limpiar y estandarizar estructuras químicas representadas como cadenas SMILES en un DataFrame utilizando la biblioteca pandas y herramientas específicas para manipulación química.

- Procesamiento de Cadenas SMILES: Se implementó una función que procesa cada cadena SMILES para neutralizar las cargas presentes en las moléculas. Esta función utiliza un conjunto definido de reacciones químicas para reemplazar subestructuras con cargas por sus equivalentes neutros. Si se realizan cambios durante este proceso, se devuelve la cadena SMILES procesada; si no, se retorna la cadena original.
- Filtrado de Átomos No Deseados: Otra función se encarga de filtrar las moléculas que contienen átomos no deseados, como metales o átomos pesados. Esta función verifica cada molécula en el DataFrame y, si encuentra un átomo restringido, reemplaza la cadena

SMILES correspondiente por una cadena vacía. Esto garantiza que solo se conserven las moléculas que cumplen con los criterios establecidos.

- Estandarización de Moléculas: La estandarización permite asegurar que todas las estructuras químicas estén en un formato consistente. Para ello, se aplica una serie de transformaciones a cada cadena SMILES válida, incluyendo la actualización de valencias, la kekulización, la normalización y la neutralización de cargas. Este proceso asegura que las moléculas sean químicamente válidas y adecuadas para análisis posteriores.
- Limpieza y Estandarización del DataFrame: Finalmente, se implementó una función que orquesta todo el proceso de limpieza del DataFrame. Inicializa una columna para almacenar las cadenas SMILES procesadas y ejecuta una serie de pasos que incluyen el filtrado de moléculas problemáticas, el manejo de solventes y sales, la neutralización de cargas y la estandarización final.

Transformación de los SMILES a Canónicos e isoméricos

se transformaron los SMILES a sus versiones isoméricas. Esta operación permite capturar de manera más precisa la estereoquímica y la configuración tridimensional de las moléculas, lo cual es esencial para comprender su comportamiento y propiedades químicas. Esta estrategia no solo preserva la integridad de los datos originales, sino que también ofrece representaciones más completas y detalladas de las estructuras moleculares. Esto facilita un análisis y modelado molecular más exhaustivo en las etapas posteriores del proyecto.

Una vez completados los procedimientos previos, se procede a realizar análisis visuales de la variable logBB, para evaluar su distribución. Se generan dos representaciones gráficas: un histograma como se ve en la Figura 4 y en la Tabla II.

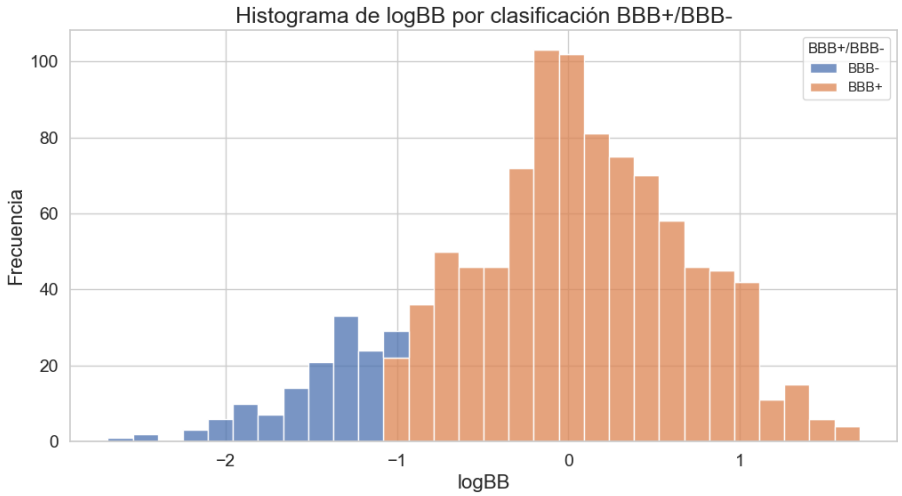


Figura 5, Distribución de logBB por clasificación BBB+/BBB-. Fuente: Elaboración propia.

Tabla II, estadísticas de variable LogBB en relación a la variable BBB+/BBB-. Fuente: Elaboración propia.

Categoría	Count	Mean	Std	Min	25%	50%	75%	Max
BBB+	930.0	0.110860	0.577206	-1.00	-0.29	0.075	0.5075	1.70
BBB-	128.0	-1.449.141	0.329777	-2.69	-1.57	-1.375	-12.300	-1.01

El análisis de los datos de permeabilidad de la BHE revela una clara distinción entre las categorías BBB+ y BBB-. La categoría BBB+, que presumiblemente representa compuestos con alta permeabilidad, muestra una tendencia hacia valores positivos y una mayor variabilidad en sus mediciones. Por otro lado, la categoría BBB-, asociada a baja permeabilidad, presenta una marcada inclinación hacia valores negativos con una distribución más concentrada. Esta separación evidente entre las dos categorías sugiere que los descriptores utilizados son efectivos para discriminar entre compuestos con alta y baja permeabilidad de la BHE. Tal diferenciación podría ser valiosa en el desarrollo de modelos predictivos para el descubrimiento y diseño de fármacos, especialmente en lo que respecta a la capacidad de los compuestos para atravesar la barrera hematoencefálica.

Debido a la gran cantidad de datos nulos en la columna de logBB, que asciende a 6749 valores nulos de un total de 7807 moléculas, es importante señalar que estos datos solo pueden obtenerse mediante análisis experimentales en el laboratorio. La ausencia de valores en logBB limita el análisis y la interpretación de la permeabilidad de las moléculas a través de la barrera hematoencefálica, lo que resalta la necesidad de realizar estudios experimentales para obtener información precisa y completa.

Calculo del LogP (coeficiente de partición octanol-agua)

Dado que los datos de LogBB (coeficiente de partición cerebro-sangre) no se encuentran completo resultado que se muestra en la verificación de datos nulos, se procede a calcular el LogP, ya que este se utiliza como un indicador alternativo de lipofilia. El LogP ayuda a inferir la capacidad de las moléculas para atravesar la barrera hematoencefálica (BHE), proporcionando una aproximación práctica en ausencia de datos experimentales directos.

En la presente investigación, la elección de calcular LogP en lugar de LogBB se debe a la limitación en la disponibilidad de datos de LogBB para un número significativo de moléculas en la base de datos. El LogBB es un parámetro específico que mide la distribución de una molécula entre la sangre y el cerebro, y su disponibilidad puede ser limitada debido a la falta de estudios experimentales específicos para muchas moléculas. Dado que el LogP es una medida indirecta de la lipofilia de una molécula, que influye en su capacidad para atravesar membranas biológicas, su cálculo proporciona una aproximación útil en ausencia de datos directos de LogBB. La lipofilia, medida por el LogP, es un factor importante en la determinación de la capacidad de una molécula para cruzar la BHE, ya que las moléculas más lipofílicas tienden a tener una mayor probabilidad de atravesar la barrera.

Por lo tanto, al calcular el LogP para las moléculas, se obtiene un parámetro relevante que puede

servir como sustituto aproximado en la ausencia de datos experimentales de LogBB, facilitando la realización de análisis y la toma de decisiones en la investigación de permeabilidad de moléculas.

Esta metodología permite avanzar en el estudio y la evaluación de moléculas con potencial para aplicaciones terapéuticas, proporcionando un valor predictivo que complementa el análisis existente en el contexto de la investigación sobre la barrera hematoencefálica como se ve según la figura 1.

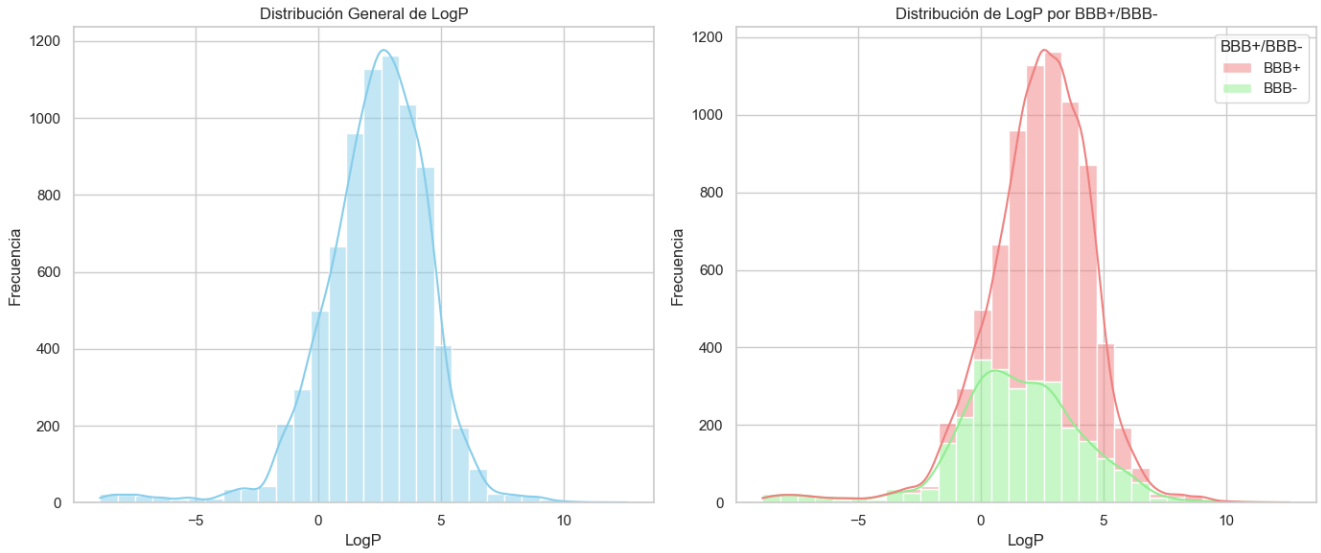


Figura 6, Distribución General de LogP por Categorización BHE+/BHE-. Fuente: Elaboración propia.

Generación de descriptores moleculares

Para enriquecer la base de datos con descriptores moleculares, se implementó un proceso utilizando diversas bibliotecas y técnicas específicas. Primero, se importaron las bibliotecas esenciales: pandas para la manipulación de datos, mordred Para el cálculo de descriptores moleculares por lo que esta es una biblioteca que ofrece una gama más amplia y diversa de descriptores moleculares. Al integrar las capacidades de Mordred con las de RDKit, se logra enriquecer significativamente la información disponible sobre las moléculas. Esto se traduce en un aumento notable en la cantidad de descriptores disponibles para cada molécula, lo que permite una descripción más detallada y completa de sus propiedades moleculares., rdkit para el manejo de estructuras químicas, numpy para cálculos numéricos, tqdm para mostrar el progreso del cálculo y warnings para manejar advertencias durante la ejecución.

En el proceso de desarrollo del script para la obtención de datos, se estableció una regla determinante que solo permitía considerar variables con datos completos. Esta decisión se tomó con el objetivo de trabajar de manera más eficiente en la optimización del modelo. La justificación detrás de esta medida radica en que, al permitir un umbral de hasta un 10% de valores faltantes (threshold = 0.1), se obtenían más de 1400 descriptores. Esta cantidad prácticamente duplicaba la cantidad de variables manejadas anteriormente, lo que dificultaba significativamente la revisión

y el manejo de datos atípicos y nulos presentes en un número considerable de variables. Al aplicar un criterio más riguroso y trabajar únicamente con variables completamente llenas, si bien se redujo la cantidad de datos disponibles, se logró una mayor facilidad en la gestión y análisis de los mismos. Esto permitió un mejor control sobre la calidad de los datos y una optimización del modelo de predicción, evitando posibles sesgos o distorsiones derivados de la presencia de valores faltantes.

El proceso comienza con la configuración del calculador de descriptores de Mordred, el cual se inicializa para evaluar las características moleculares basadas en las cadenas SMILES. La función calculate_descriptors toma una cadena SMILES, convierte esta cadena en una molécula usando RDKit, y luego calcula los descriptores moleculares correspondientes. Si la molécula es válida, se extraen y verifican los valores de los descriptores, excluyendo valores extremos para asegurar la calidad de los datos. En caso de que la molécula no sea válida, se devuelve una serie con valores None para mantener la integridad del DataFrame.

El uso de tqdm permite visualizar el progreso del cálculo de descriptores a medida que se aplican a cada entrada de la columna SMILES en el DataFrame. Luego, los descriptores calculados se concatenan con los datos originales, enriqueciendo la base de datos con características moleculares detalladas. Finalmente, se eliminan las columnas con valores faltantes para asegurar que el DataFrame resultante contenga solo descriptores completos y útiles.

Este proceso permite proporcionar una base de datos más rica y detallada, permitiendo análisis moleculares más profundos y facilitando el desarrollo de modelos predictivos más precisos en investigaciones químicas y farmacéuticas. Todo lo anterior se puede traducir en un aumento notable en la cantidad de descriptores disponibles para cada molécula, lo que permite una descripción más detallada y completa de sus propiedades moleculares. La implementación de esta actividad ha resultado en un DataFrame expandido, con 7807 filas y 747 columnas. Esto indica que se han calculado un total de 747 descriptores moleculares adicionales para cada una de las 7807 moléculas en la base de datos. Esta expansión en la cantidad de descriptores proporciona una visión más completa y detallada de las propiedades moleculares, lo que puede ser invaluable en una variedad de aplicaciones, incluido el diseño de fármacos, la química computacional y la investigación de propiedades físico-químicas entre otras.

La integración de las bibliotecas RDKit y Mordred para el cálculo de descriptores moleculares ha sido un componente fundamental. Ya que esta ofrece una amplia gama de descriptores moleculares avanzados, lo que nos permitió obtener una representación más completa y detallada de las moléculas en estudio. Esta combinación de descriptores básicos y avanzados nos ha permitido abordar la complejidad de las propiedades moleculares desde diferentes perspectivas, mejorando así la robustez y la precisión de nuestro análisis. Esta integración nos ha permitido obtener una visión más completa y detallada de las propiedades moleculares presentes en nuestro conjunto de datos en la tabla III.

Tabla III, Cantidad de descriptores por categoría. Fuente: Elaboración propia

Categoría	Número de Descriptores	Categoría	Número de Descriptores
ABCIndexBase	2	RASA	1
SmartsCountBase	2	CarbonTypesBase	10
Descriptor	305	ChiBase	56
AromaticBase	2	ConstitutionalSum	8
AutocorrelationBase	303	DetourMatrixBase	14
ATS	99	EStateBase	316
ATSC	108	EtaBase	45
MATS	96	GeometricalIndexBase	3
BCUTBase	24	GeometricalShapeIndex	1
BaryszMatrixBase	104	HBondBase	2
VersionCPSABase	15	InformationContentBase	12
PNSA	5	TotalIC	18
FNSA	10	InformationContent	12
FPSA	5	KappaShapeIndexBase	3
CPSABase	4	LipinskiLike	2
RNCG	1	MoeTypeBase	52
RNCS	1	MolecularIdBase	12
TASA	1	MomentOfInertiaBase	3
PolarizabilityBase	2	PathCountBase	21
RingCountBase	138		
RotatableBondsBase	2		

La elección de estas bibliotecas se basó en su capacidad para calcular una amplia variedad de descriptores moleculares de manera eficiente y precisa, así como en su amplio conjunto de funciones que se adaptan perfectamente a nuestras necesidades de análisis. Los descriptores moleculares analizados abarcan un amplio espectro de propiedades químicas y estructurales, distribuidos en 43 categorías distintas. La categoría más extensa es "EStateBase" con 316 descriptores, seguida por "Descriptor" con 305 y "AutocorrelationBase" con 303. Estas categorías predominantes sugieren un enfoque significativo en estados electrónicos y propiedades autocorrelacionadas. El conjunto incluye descriptores constitucionales, topológicos, geométricos, electrónicos y basados en fragmentos, proporcionando una caracterización integral de las moléculas. La variedad en el número de descriptores por categoría, desde categorías con un solo descriptor (como RASA y GeometricalShapeIndex) hasta aquellas con cientos, indica una cobertura diversa de propiedades moleculares. Esta colección de descriptores parece ser adecuada para análisis QSAR (Relaciones Cuantitativas Estructura-Actividad) detallados, permitiendo la captura de una amplia gama de características moleculares relevantes para la

predicción de propiedades y actividades químicas.

Como resultado de este proceso de integración de bibliotecas, nuestro conjunto de datos experimentó una expansión significativa. Esta ampliación nos ha permitido capturar una cantidad mucho mayor de información molecular, lo que ha enriquecido enormemente nuestro modelo predictivo y ha mejorado nuestra capacidad para predecir la permeabilidad de la barrera hematoencefálica. En resumen, la integración de las bibliotecas RDKit y Mordred ha sido esencial para avanzar en nuestro análisis y para obtener resultados más sólidos y precisos en nuestro estudio.

Este enfoque metodológico nos permite avanzar significativamente en la construcción de un modelo integral para predecir la permeabilidad de la BHE. La combinación de técnicas de procesamiento de datos, análisis molecular y validación de datos nos posiciona en la vanguardia de la investigación en este campo, ofreciendo resultados sólidos y contribuyendo al avance del conocimiento en la predicción de la permeabilidad de la Barrera Hematoencefálica.

Este hallazgo refuerza la importancia del segundo objetivo de esta investigación, que se enfoca en la mejora de la integridad y la calidad de los datos en la base de datos de moléculas.

Para ver el código que implementa este análisis, puedes acceder al repositorio en GitHub https://github.com/CCIBANEZB/BHE/blob/main/Preparacion%20y%20procesamiento%20de%20datos.IPYNB Este enlace te llevará directamente al archivo donde se encuentra el código implementado para esta etapa.

6.2 Implementar un sistema integral para la adquisición y pre-procesamiento de datos, facilitando la estimación precisa de la permeabilidad de la Barrera Hematoencefálica BHE

Esta fase se dividió en tres partes: primero, el análisis de las variables categóricas (SMILES), seguido por la evaluación de las variables numéricas o descriptores moleculares relacionados con las propiedades físico-químicas de las moléculas, y finalmente, la generación de un conjunto de datos de permeabilidad de la BHE para compuestos similares a fármacos. Inicialmente, se trabajó con las estructuras SMILES, analizando características como la longitud, simetría y frecuencia de caracteres, para identificar patrones entre moléculas que atraviesan la BHE y las que no. Luego, se realizó un análisis exploratorio de datos (EDA) en los descriptores moleculares, con el objetivo de encontrar relaciones entre moléculas BHE+ y BHE-. Durante este proceso, se evaluaron correlaciones y se aplicó análisis de componentes principales (PCA) para reducir la dimensionalidad. Además, se desarrollaron modelos de aprendizaje automático para identificar los descriptores más relevantes y mejorar el diseño del modelo predictivo que estimará la capacidad de una molécula para atravesar la BHE.

6.2.1 Análisis de estructuras SMILES y características moleculares

En esta etapa se realizaron verificaciones adicionales para asegurar la calidad de los datos. Se

comprobó que ninguna variable tuviera todos sus valores en cero, que no existieran variables duplicadas y que todos los descriptores moleculares fueran numéricos, como se indicó en el script del capítulo anterior. Los resultados mostraron que no había variables con valores en cero ni nombres duplicados, lo que garantiza la unicidad de las columnas. Además, se identificaron variables con valores faltantes, específicamente en las columnas 'name', 'IUPAC_name', 'CID', 'threshold' y 'comments', coincidiendo con los hallazgos previos.

Clasificación de Variables por Tipo.

Esta operación implicó identificar y contar el número de variables numéricas y categóricas en la base de datos. Los resultados mostraron que 736 variables son numéricas en la base de datos, mientras que el número de variables categóricas es de 11, tal como se observa en la tabla IV.

Tabla IV, Resumen de Columnas Categóricas. Fuente: Elaboración propia.

Cantidad de columnas numéricas	736
Cantidad de columnas categóricas	11
1. name	
2. IUPAC_name	
3. SMILES	
4. BBB+/BBB-	
5. Inchi	
6. reference	
7. group	
8. comments	
9. Tipo_SMILES	
10. SMILES_fixed	
11. SMILES_final	

Como un factor crucial en este proceso investigativo, se realizaron una serie de análisis de estructuras SMILES y características moleculares. Estos análisis tienen como objetivo responder a los posibles interrogantes planteados, proporcionando claridad y dirección en la investigación.

Análisis de Longitud

Para este análisis se diseñó un script el cual realizó un análisis exploratorio de la longitud de las representaciones SMILES para moléculas clasificadas como BBB+ y BBB-, utilizando pandas para manipulación de datos, seaborn y matplotlib para visualización, y scipy para análisis estadístico. Se compararon las longitudes de las cadenas SMILES de ambos grupos, calculando medidas estadísticas descriptivas y realizando pruebas de distribución para identificar posibles diferencias en la longitud de las representaciones entre moléculas que cruzan y no cruzan la barrera hematoencefálica, como se evidencia en la figura 6 y tabla V.

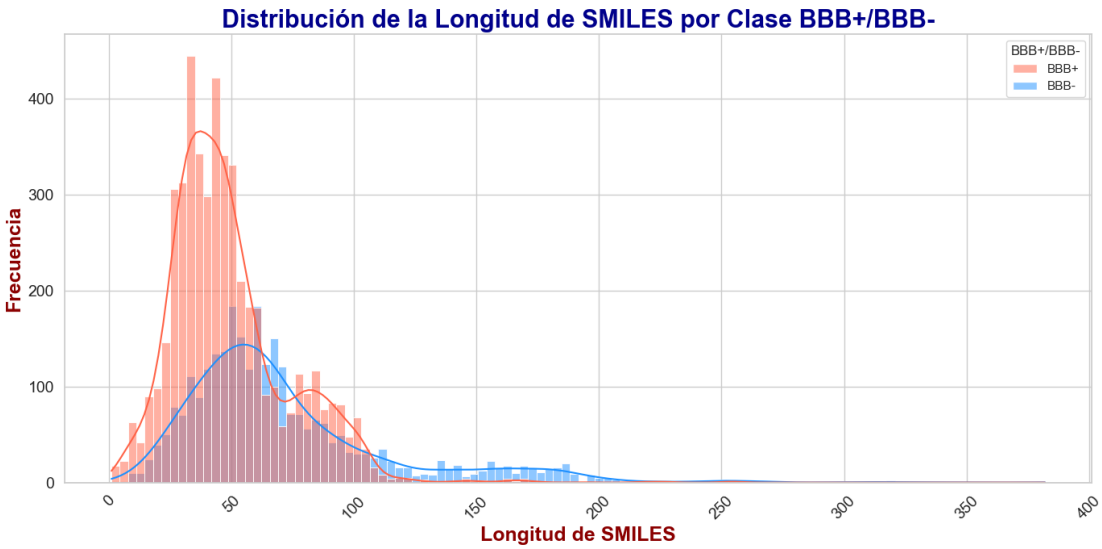


Figura 7, Distribución de la Longitud de SMILES por Clase BBB+/BBB-. Fuente: Elaboración propia.

Tabla V, Estadísticas descriptivas para SMILES. Fuente: Elaboración propia.

Categoría	Count	Mean	Std	Min	25%	50%	75%	Max
BBB+	4956.0	49.211259	24.472047	1.0	33.0	44.0	59.0	256.0
BBB-	2851.0	71.949491	43.568292	4.0	45.0	61.0	85.0	382.0

Desde la perspectiva de la química computacional, el análisis de 7807 SMILES revela que los compuestos que atraviesan la barrera hematoencefálica (BBB+) tienen una longitud promedio de 49.21 caracteres, mientras que los que no lo hacen (BBB-) promedian 71.95 caracteres. Esta diferencia sugiere que las moléculas más cortas son más propensas a ser permeables.

Este enfoque permite identificar patrones estructurales que influyen en la permeabilidad de las moléculas. Desde una perspectiva científica y química, este análisis es innovador, ya que aborda una característica esencial en el diseño de fármacos que podrían cruzar la BHE, una barrera crítica para el tratamiento de enfermedades neurológicas.

Análisis de simetría

Para este punto se realiza un análisis de la simetría molecular, utilizando SMILES como entrada, para evaluar su relación con la capacidad de una molécula para atravesar la BHE. En donde el script define una función para calcular la simetría de una molécula, utilizando el número de anillos como un proxy de simetría. Luego, aplica esta función a cada molécula en un DataFrame, descartando aquellas con SMILES inválidos. Después, se generan estadísticas descriptivas de la simetría para las clases BHE+ y BHE-, y se visualiza la distribución de la simetría mediante un diagrama de cajas evidenciados en la tabla VI y figura 7 descritas a continuacion.

Tabla VI, Estadísticas descriptivas de simetría. Fuente: Elaboración propia.

Categoría	Count	Mean	Std	Min	25%	50%	75%	Max
BBB+	4956.0	3.202583	1.542837	0.0	2.0	3.0	4.0	9.0
BBB-	2851.0	3.379867	1.718180	0.0	2.0	3.0	4.0	16.0

Las estadísticas de simetría para BBB+ (media: 3.20) y BBB- (media: 3.38) muestran que los compuestos BBB- son ligeramente más simétricos, con una mayor desviación estándar en BBB- (1.72) que indica mayor variabilidad estructural. Esto sugiere que la simetría puede influir en la capacidad de los compuestos para atravesar la barrera hematoencefálica.

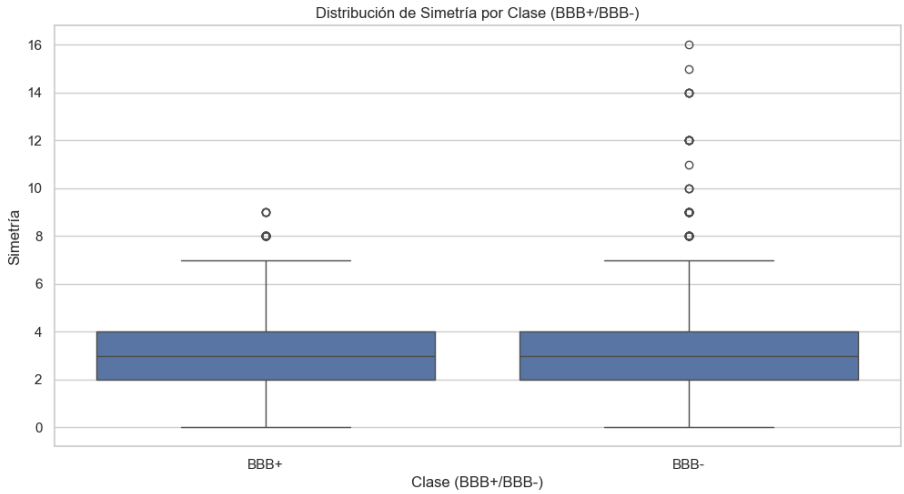


Figura 8, BoxPlot, Distribución de Simetría por Clase (BHE+/BHE-). Fuente: Elaboración propia.

Estos resultados resaltan diferencias en la distribución y tendencia de las calificaciones entre las dos categorías analizadas. Esto indica que las calificaciones BHE- muestran una mayor dispersión, con una desviación estándar de 1.72 en comparación con 1.54 para BHE+. Esto sugiere que los datos de BHE- son más variados y menos concentrados en torno a la media.

Análisis de quirales

Este análisis investiga la quiralidad de las moléculas. Utilizando la función is_chiral, se determina la quiralidad de las moléculas representadas por SMILES y se clasifica en función de si pueden atravesar la BHE (BHE+) o no (BHE-) como lo muestra la tabla VII y figura 8.

Tabla VII, Número de moléculas quirales por clase. Fuente: Elaboración propia.

Categoría	Número de Moléculas Quirales
BBB+	3716
BBB-	2363

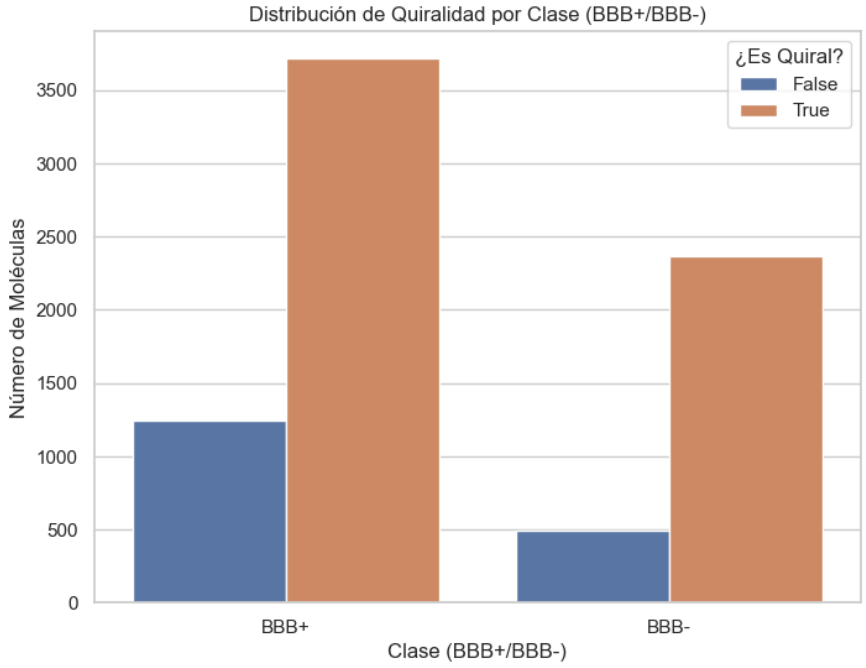


Figura 9, Distribución de Quiralidad por Clase (BHE+/BHE-). Fuente: Elaboración propia.

El estudio examinó la relación entre la quiralidad molecular y la capacidad de atravesar la BHE en un conjunto de 7,807 compuestos. Las moléculas quirales representaron la mayoría (77.9%) de la muestra, con 6,079 compuestos, frente a 1,728 no quirales (22.1%). Se observó una tendencia general de mayor permeabilidad BBB en ambas categorías, con el 61.1% de las moléculas quirales y el 71.8% de las no quirales capaces de atravesar la barrera. Esta distribución sugiere que la quiralidad por sí sola no es un factor determinante en la permeabilidad de la BBB, aunque podría influir en cierta medida. Ambos grupos mostraron similitudes en características estructurales y potencial farmacológico, incluyendo la presencia de heteroátomos y complejidad molecular. Estos hallazgos proporcionan información valiosa para el diseño de fármacos dirigidos al sistema nervioso central, destacando la importancia de considerar múltiples factores moleculares en la predicción de la permeabilidad de la BHE.

Análisis de frecuencia de Caracteres en SMILES BHE+/BHE-

Para esta parte se hicieron diferentes procedimientos primero se realizó un análisis de n-gramas (bigrama y trigrama) sobre representaciones SMILES para identificar patrones frecuentes y combinaciones específicas como en los bigramas más comunes fueron cc, CC, O), C(, y [C@, mientras que los trigramas incluyeron [C@, @H], ccc, =O), y (=O. Combinaciones específicas como Cl y N+ fueron representativas en el conjunto de datos tales resultados se evidencian en las siguientes figuras 9 y 10.

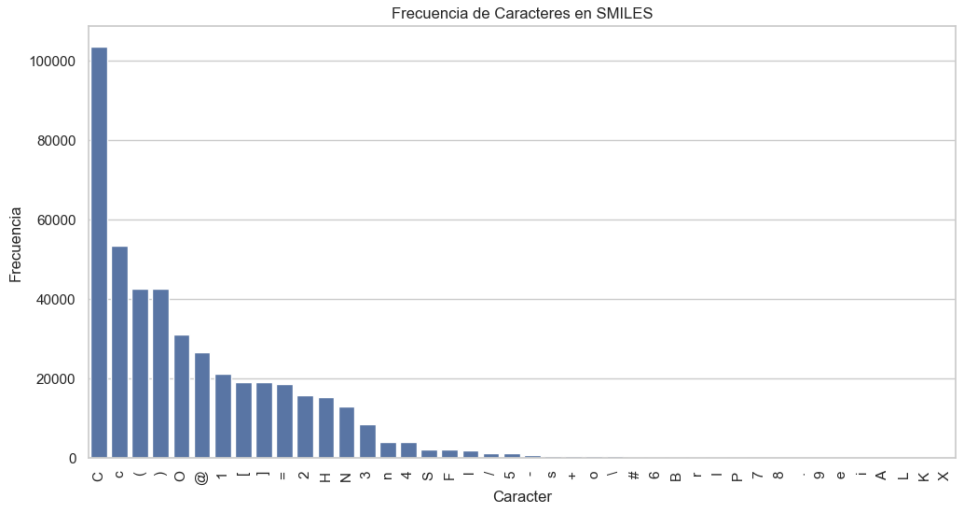


Figura 10, Frecuencia de caracteres SMILES. Fuente: Elaboración propia.

Posteriormente se hizo el análisis para comparar la frecuencia de caracteres en SMILES entre moléculas que atraviesan (BHE+) y no atraviesan (BHE-) la barrera hematoencefálica. Se calcula y visualiza la frecuencia de cada carácter para cada categoría, destacando las diferencias en su distribución. Los informes textuales y gráficos de barras muestran los caracteres más y menos comunes, proporcionando insights sobre cómo estas frecuencias varían según la permeabilidad de las moléculas.

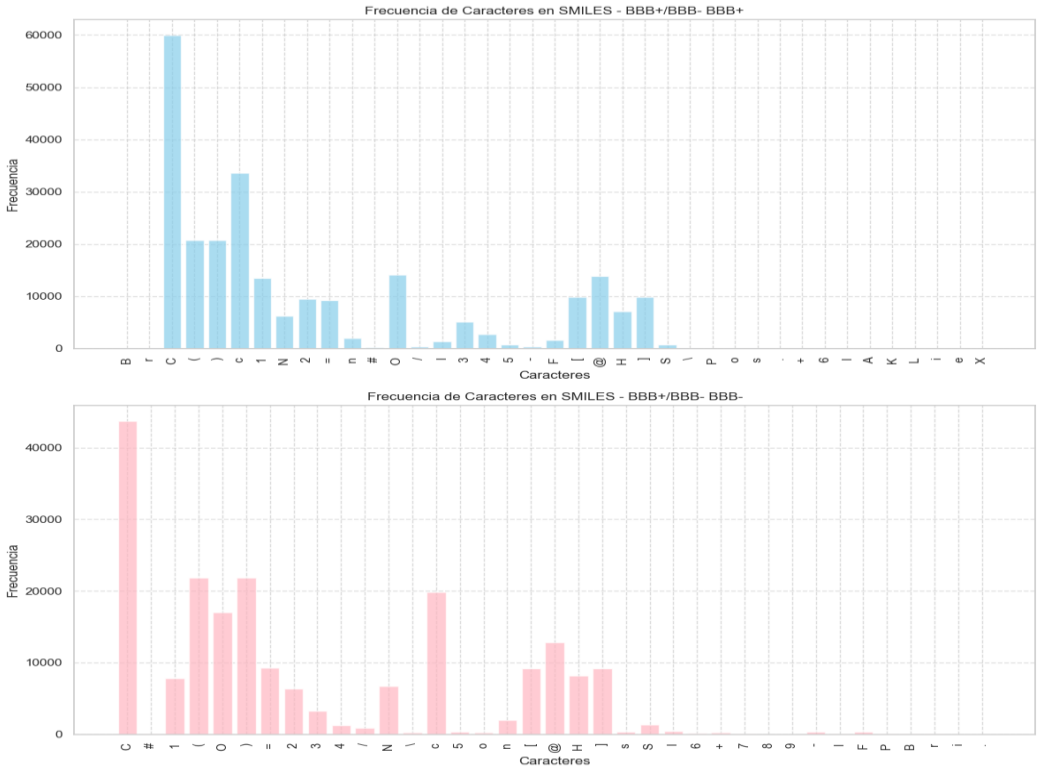
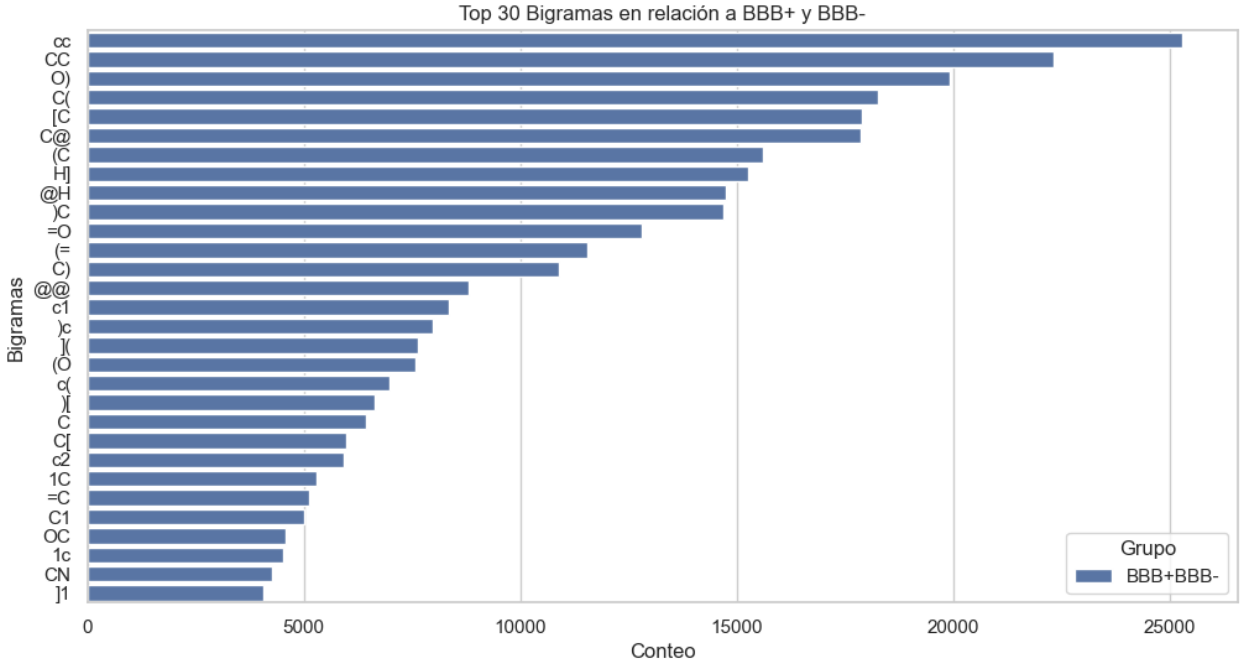


Figura 11, Frecuencia de Caracteres en SMILES en relación a BHE+/BHE-. Fuente: Elaboración propia

Los gráficos muestran que, para las moléculas que atraviesan la barrera hematoencefálica (BHE+), los caracteres más frecuentes son C, c, ( ), y O, mientras que los menos frecuentes incluyen X, i, L, K, y A. En contraste, para las moléculas que no atraviesan la barrera (BHE-), predominan C, (,), y c, mientras que los menos comunes son i, 9, ., y 8. Estos patrones pueden ofrecer información valiosa sobre las diferencias estructurales en función de la permeabilidad. EN OTRAS PALABRAS n la categoría BHE+, los caracteres más frecuentes son C (carbono), c (carbono aromático), (y) (delimitadores de anillos), y O (oxígeno), indicando la alta presencia de estos elementos en las moléculas. Los menos comunes, como X y i, no tienen significados estándar universales en SMILES pero pueden representar átomos o estructuras específicas. En BHE-, i y 9 son raros; i puede marcar posiciones en anillos, y 9 podría estar relacionado con numeraciones específicas en estructuras personalizadas.

Análisis de n-gramas en moléculas con diferente permeabilidad de la barrera hematoencefálica BHE

El análisis de bigramas y trigramas en los SMILES revela diferencias significativas entre las moléculas que atraviesan (BBB+) y no atraviesan (BBB-) la barrera hematoencefálica. Las diferencias en combinaciones como 'CN' y 'NC' sugieren variaciones estructurales que pueden influir en la permeabilidad. Sin embargo, hay similitudes en patrones como '(O' y 'C)', lo que indica que ciertos fragmentos químicos son comunes en ambos grupos. Este estudio proporciona información valiosa para entender la relación entre la estructura molecular y la capacidad de atravesar la BBB como lo demuestra la siguiente figura 11.



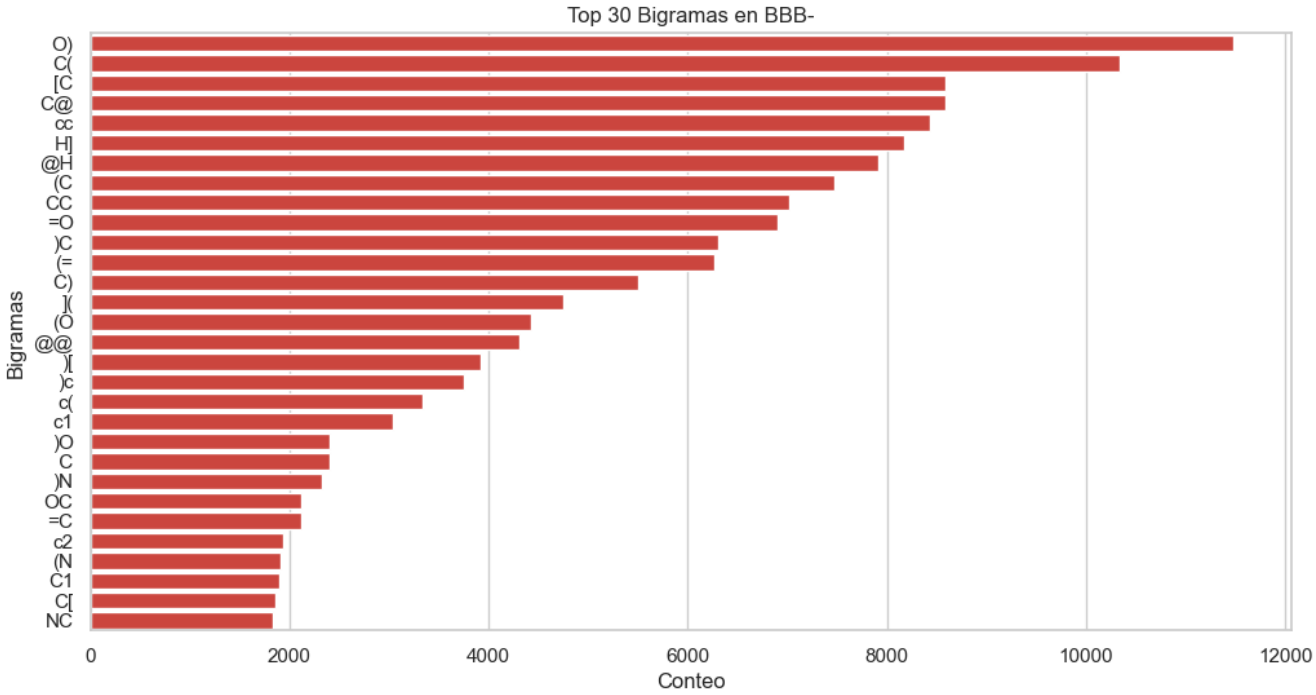


Figura 12, Top 30 Bigramas en relación a BBB+ Y BBB-. Fuente: Elaboración propia.

Análisis de polaridad

Se llevó a cabo un análisis de polaridad de las moléculas representadas por SMILES en las categorías BHE+ y BHE-. La polaridad se evaluó utilizando el LogP como proxy. Los resultados muestran que la polaridad media de las moléculas en la categoría BHE+ es de 2.87, con una desviación estándar de 1.58. La polaridad varía entre -4.38 y 10.06, con el 25% de los valores por debajo de 1.84 y el 75% por debajo de 3.96. En contraste, las moléculas en la categoría BHE- tienen una polaridad media de 1.46 y una desviación estándar de 2.76, con un rango que va de -8.90 a 12.61. En esta categoría, el 25% de los valores están por debajo de 0.00 y el 75% por debajo de 3.11. en otras palabras, las moléculas en la categoría BHE+ tienden a tener una mayor polaridad media y menor variabilidad en comparación con las de BHE-. La gama de polaridad es más amplia en BHE-, lo que sugiere una mayor diversidad en las propiedades de polaridad de estas moléculas tal se ve en la figura 12 y tabla VIII.

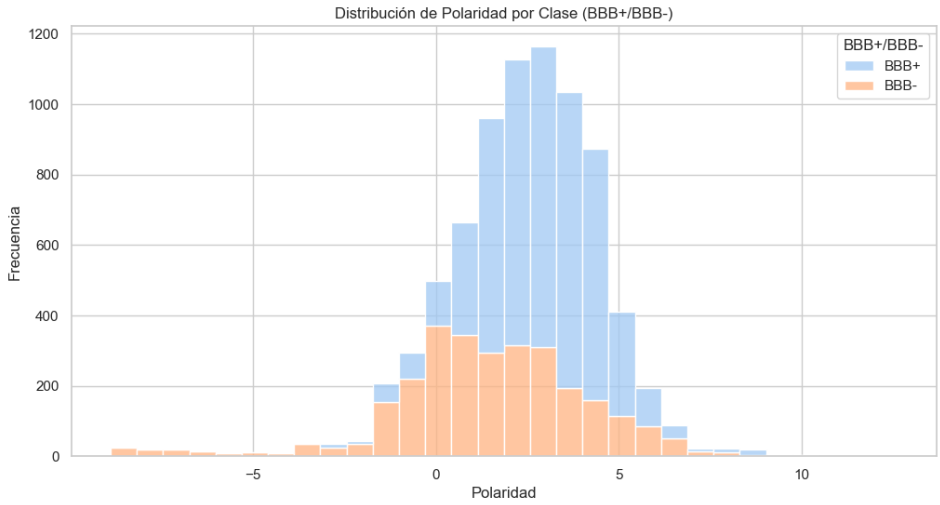


Figura 13, Distribución de Polaridad por Clase (BHE+/BHE-). Fuente: Elaboración propia.

Tabla VIII, Estadísticas descriptivas de polaridad. Fuente: Elaboración propia.

Categoría	Count	Mean	Std	Min	25%	50%	75%	Max
BBB+	4956.0	2.874041	1.578878	-4.3754	1.8437	2.9233	3.9625	10.0563
BBB-	2851.0	1.455545	2.755004	-8.8953	0.0013	1.4582	3.1111	12.6058

Las visualizaciones de las distribuciones de polaridad indican que las moléculas en la categoría BHE+ tienden a tener una mayor concentración de valores de polaridad y menor variabilidad en comparación con las moléculas en BHE-. En cambio, la categoría BHE- presenta una gama más amplia de propiedades de polaridad, lo que sugiere una mayor diversidad en las características moleculares. Este análisis destaca diferencias significativas en la polaridad entre las dos categorías, proporcionando información valiosa sobre las propiedades moleculares asociadas con la capacidad de atravesar la barrera hematoencefálica.

Análisis de Cargas Explícitas en SMILES para las Categorías BHE+ y BHE-

Se implementó un análisis de cargas explícitas en moléculas representadas por SMILES, utilizando el modelo de Gasteiger para calcular las cargas atómicas. La función calculate_explicit_charges convierte los SMILES en moléculas utilizando RDKit, añade átomos de hidrógeno para obtener cargas precisas, y calcula las cargas Gasteiger. Estas cargas se almacenan en la columna ExplicitCharges del DataFrame.

Para verificar las propiedades atómicas, la función check_atom_properties examina el primer SMILES del conjunto de datos y extrae las propiedades atómicas disponibles. Esto asegura que las propiedades de carga se obtengan correctamente y que se conozcan todas las propiedades disponibles para análisis futuros.

Las visualizaciones incluyen un BoxPlot que muestra la distribución de cargas explícitas por

categoría, revelando que las moléculas en BHE+ tienden a tener una menor variabilidad en sus cargas en comparación con las de BHE- como se ve en la siguiente figura 13.

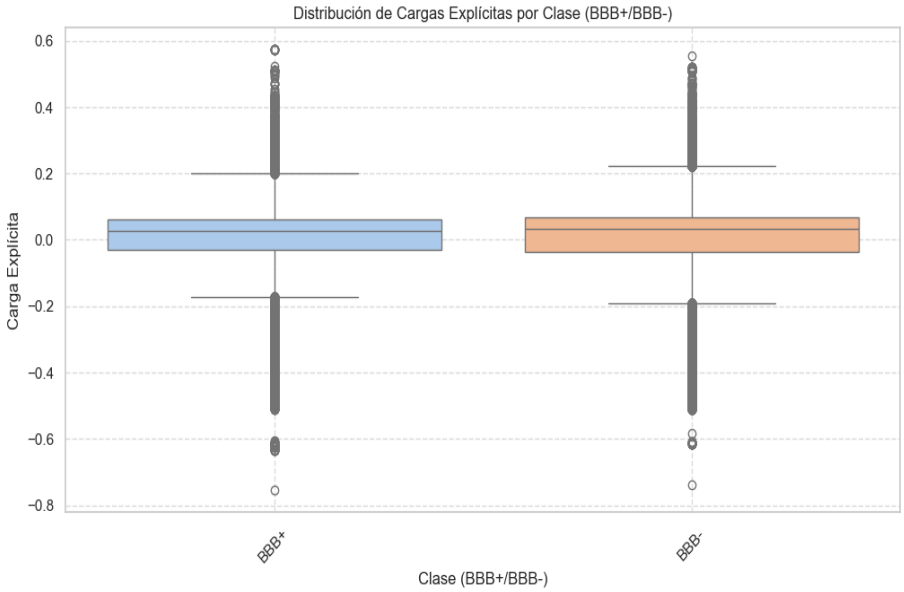


Figura 14, BoxPlot Distribución de cargas explicitas por clase. Fuente: Elaboración propia.

Análisis de cargas explicitas

La distribución de cargas atómicas en moléculas con alta (BBB+) y baja (BBB-) permeabilidad de la barrera hematoencefálica. Los resultados muestran ligeras diferencias en las propiedades de carga entre ambos grupos. Las moléculas BBB+ presentaron una media de carga cercana a cero (0.000094) con menor dispersión (σ=0.130554), mientras que las BBB- mostraron una media ligeramente más positiva (0.000686) y mayor variabilidad (σ=0.162931). Ambos grupos exhibieron rangos de carga similares como lo muestra la tabla IX, sugiriendo que la distribución de cargas atómicas podría influir sutilmente en la permeabilidad de la BBB, aunque no es un factor determinante único. Estos hallazgos pueden ser valiosos para el diseño de fármacos dirigidos al sistema nervioso central.

Tabla IX, Propiedades atómicas SMILES. Fuente: Elaboración propia.

BBB+/BBB-	count	mean	std	min	25%	50%	75%	max
BBB+	233,634	0.000094	0.130554	-0.754605	-0.031386	0.028042	0.061877	0.572607
BBB-	173,470	0.000686	0.162931	-0.739367	-0.035873	0.032403	0.066951	0.554222

Análisis de enlaces dobles en relación a BHE+/BHE-

El análisis de los enlaces dobles que se muestra en la figura 14 en las moléculas clasificadas como BBB+ y BBB- revela diferencias significativas en su distribución y cantidad. Para el grupo BBB+, se registraron un total de 4,956 enlaces dobles, con una media de 1.86 y un máximo de 12. En contraste, el grupo BBB- mostró 2,851 enlaces dobles, con una media notablemente más alta de

3.26 y un máximo de 16. Esto sugiere que las moléculas clasificadas como BBB- tienden a tener una mayor complejidad estructural en términos de enlaces dobles, lo que podría influir en sus propiedades químicas y biológicas. La variabilidad en ambos grupos también se destaca, con desviaciones estándar de 1.75 para BBB+ y 2.47 para BBB-, indicando una diversidad en la cantidad de enlaces presentes en las moléculas dentro de cada categoría. Estos hallazgos pueden ser relevantes para comprender la relación entre la estructura molecular y la actividad biológica, así como para futuras investigaciones en química medicinal y diseño de fármacos

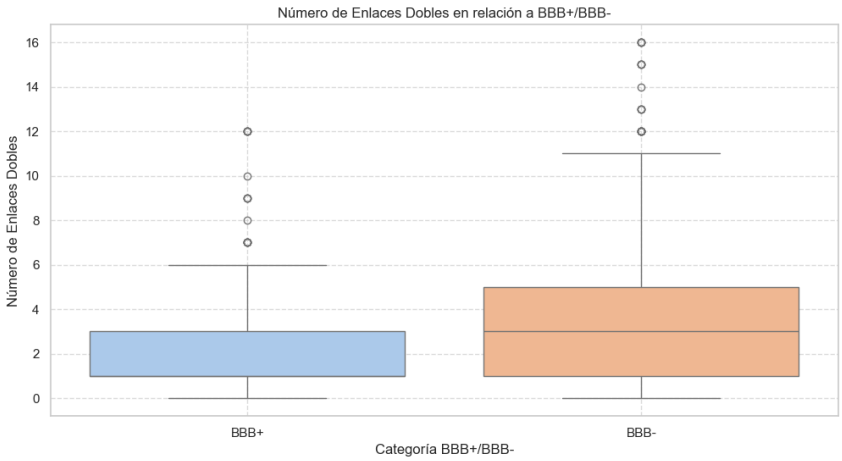


Figura 15, Análisis de Enlaces Dobles en relación a BHE+/BHE-. Fuente: Elaboración propia

Análisis del número de átomos y enlaces en relación a BHE+/BHE-

En esta sección, se presenta un análisis detallado sobre la relación entre el número de átomos y enlaces en las moléculas, y su capacidad para atravesar la barrera hematoencefálica (BHE). Para ello, se implementó una función en Python que, a partir de las cadenas SMILES, cuenta el número de átomos y enlaces presentes en cada molécula. Esta función se aplicó a la base de datos, y los resultados se visualizaron y analizaron mediante gráficos de cajas (boxplots) y estadísticas descriptivas. Empelando un método de Cálculo de Átomos y Enlaces el cual se desarrolló una función en Python utilizando la biblioteca RDKit, que procesa las cadenas SMILES para calcular el número de átomos y enlaces en cada molécula. Estos valores se almacenaron en dos nuevas columnas del DataFrame, denominadas NumAtoms y NumBonds.

Se generaron gráficos de cajas para visualizar la distribución del número de átomos y enlaces en función de las categorías BHE+/BHE-, que indican si la molécula atraviesa o no la barrera hematoencefálica tal se ve en la figura 15 y en las tablas X, XI y XII descritas a continuación.

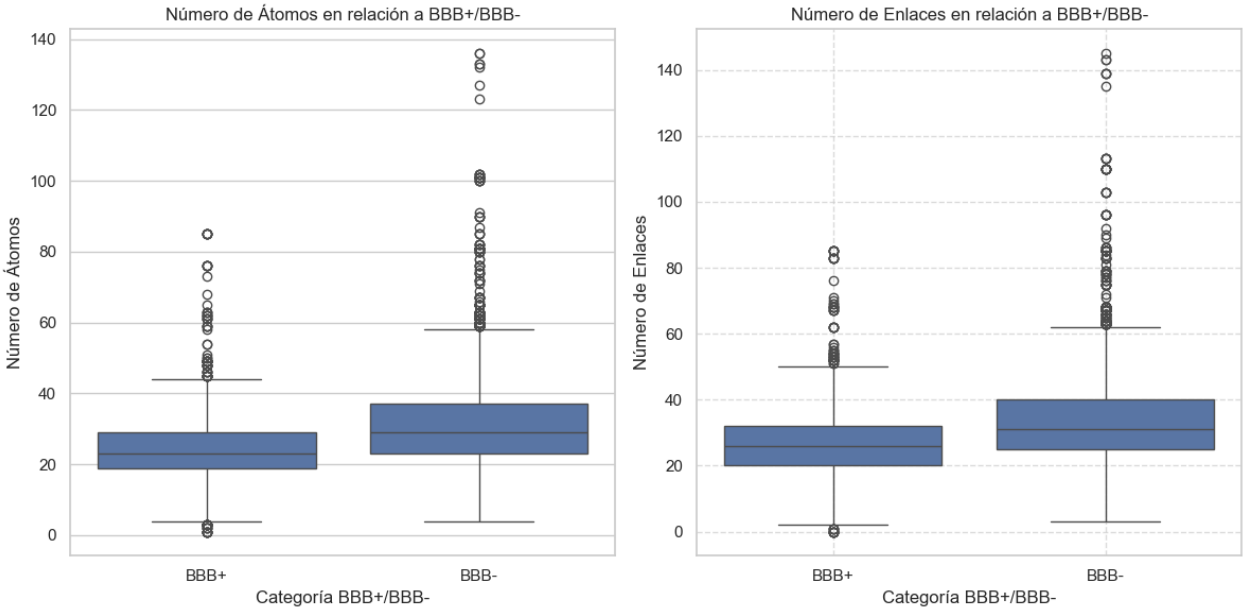


Figura 16, BoxPlot Análisis del Número de Átomos y Enlaces en Relación a BHE+/BHE-. Fuente. Propia.

Tabla X, Estadísticas Descriptivas para el Número de Átomos. Fuente: Elaboración propia.

Categoría	Count	Mean	Std	Min	25%	50%	75%	Max
BBB+	4956.0	23.908797	8.542891	1.0	19.0	23.0	29.0	85.0
BBB-	2851.0	32.101017	15.037858	4.0	23.0	29.0	37.0	136.0

Tabla XI, Estadísticas Descriptivas para el Número de Enlaces. Fuente: Elaboración propia.

Categoría	Count	Mean	Std	Min	25%	50%	75%	Max
BBB+	4956.0	26.087974	9.677104	0.0	20.0	26.0	32.0	85.0
BBB-	2851.0	34.445808	16.199051	3.0	25.0	31.0	40.0	145.0

Tabla XII, Estadísticas Descriptivas para el Número de Átomos. Fuente: Elaboración propia.

Categoría	Total	WithManyAtoms	WithManyBonds
BBB+	4956	4905	4879
BBB-	2851	2848	2847

Las estadísticas descriptivas revelan diferencias significativas entre las moléculas que atraviesan la barrera hematoencefálica (BHE+) y las que no lo hacen (BHE-) en términos de número de átomos y enlaces.

En el caso del número de átomos, las moléculas BHE+ tienen un promedio de 23.91 átomos con una desviación estándar de 8.54. En cambio, las moléculas BHE- muestran un promedio

notablemente mayor de 32.10 átomos, con una mayor variabilidad (desviación estándar de 15.04). Esto sugiere que las moléculas que no logran atravesar la barrera tienden a ser más grandes en términos de su número de átomos.

De manera similar, el número de enlaces sigue el mismo patrón. Las moléculas BHE+ presentan un promedio de 26.09 enlaces, mientras que las BHE- alcanzan un promedio de 34.45 enlaces, con una desviación estándar más alta, lo que indica mayor complejidad estructural en estas últimas.

Estos resultados destacan una tendencia clara: las moléculas más grandes y con mayor número de enlaces tienen más dificultades para atravesar la barrera hematoencefálica, por tanto, la reducción del tamaño molecular podría aumentar la probabilidad de atravesar la barrera.

Análisis del número de anillos en las moléculas y su relación con la capacidad de atravesar la barrera hematoencefálica BHE

Se realizó un análisis del número de anillos en las moléculas se realizó utilizando la función ring_count, se contó el número de anillos en cada molécula y se agregaron estos datos al DataFrame. Las estadísticas descriptivas mostraron que, en promedio, las moléculas que no atraviesan la BHE (BHE-) tienen ligeramente más anillos (3.38) en comparación con las que sí la atraviesan (BHE+) con un promedio de 3.20. Además, la desviación estándar es mayor en las moléculas BHE-, lo que indica una mayor variabilidad en la cantidad de anillos en estas moléculas como se ve en la siguiente tabla XIII.

Tabla XIII, , Número de Anillos en relación a BBB+/BBB-. Fuente: Elaboración propia.

Categoría	Count	Mean	Std	Min	25%	50%	75%	Max
BBB+	4956.0	3.202583	1.542837	0.0	2.0	3.0	4.0	9.0
BBB-	2851.0	3.379867	1.718180	0.0	2.0	3.0	4.0	16.0

Análisis visual de estructuras químicas

El análisis se centró en comparar las estructuras moleculares que atraviesan (BHE+) y no atraviesan (BHE-) la barrera hematoencefálica. Primero, se calculó el tamaño de cada molécula, representado por el número de átomos, a partir de sus SMILES. Luego, se clasificaron las moléculas en dos grupos: las que atraviesan la BHE y las que no. Se seleccionaron las cinco moléculas más grandes y pequeñas de cada grupo para una comparación detallada. Las moléculas seleccionadas se visualizaron mediante imágenes 2D generadas a partir de sus conformaciones 3D, facilitando la comparación visual de patrones estructurales entre ambos grupos, tal como se observa en la figura 15.

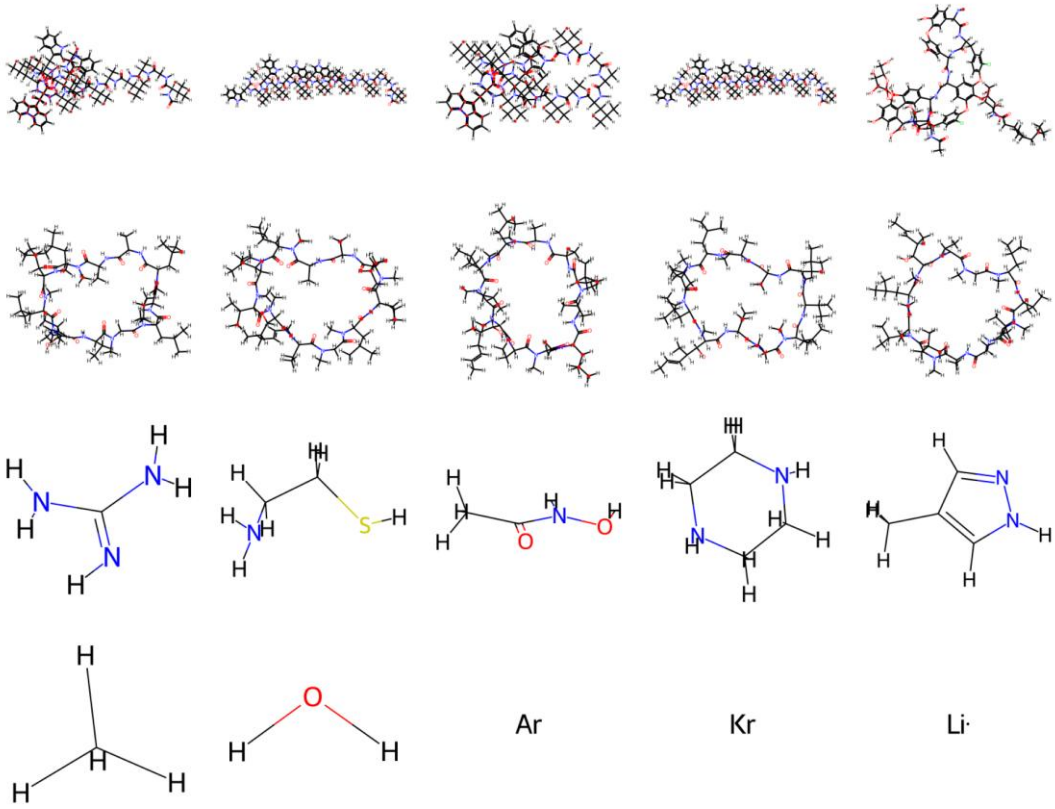


Figura 17, Análisis visual de estructuras químicas. Fuente: Elaboración propia.

6.2.2 Exploración y análisis de descriptores moleculares

Se realizó un análisis de los datos, que incluyó la identificación y tratamiento de valores nulos, la revisión de columnas duplicadas y la eliminación de variables con valores constantes en 0. Asimismo, se eliminaron variables categóricas irrelevantes, reduciendo el número de columnas de 738 a 622. Este proceso fue fundamental para asegurar la calidad y coherencia de los datos, preparando el conjunto para los análisis posteriores de manera más eficiente. Cabe destacar que, tras este proceso, la única variable con datos nulos en el DataFrame es 'LogBB'.

Además, se realizó la conversión de las variables 'Lipinski' y 'GhoseFilter', que inicialmente estaban en formato booleano (True/False), a valores numéricos de 1 y 0, respectivamente. De manera similar, se transformó la variable 'BBB+/BBB-' para representar 1 si la molécula atraviesa la barrera hematoencefálica y 0 si no lo hace. Estos cambios fueron esenciales para facilitar el procesamiento y análisis posterior de los datos en modelos de aprendizaje automático.

Exploración y análisis descriptivo

Se realizó un análisis comparativo entre los grupos BHE+ y BHE- dentro de nuestra base de datos, con un enfoque en varias variables de interés, Para el análisis exploratorio inicial, se seleccionaron

los descriptores HBD (donantes de enlaces de hidrógeno), HBA (aceptores de enlaces de hidrógeno), Número de Átomos Pesados, Peso Molecular (MW), Número de Enlaces Pesados y logBB, debido a su importancia crítica en la capacidad de una molécula para atravesar la barrera hematoencefálica (BHE). En la literatura científica, estos descriptores han sido identificados como fundamentales para evaluar la permeabilidad de los compuestos a través de la BHE.

El HBD y el HBA son esenciales para la formación de puentes de hidrógeno, que son determinantes en la interacción entre la molécula y los transportadores o receptores en la BHE. El Número de Átomos Pesados y el Peso Molecular reflejan el tamaño y la complejidad de la molécula, características que influyen en su capacidad para atravesar la barrera, dado que moléculas más grandes y complejas suelen tener más dificultades para penetrar la BHE. El Número de Enlaces Pesados proporciona información sobre la complejidad estructural de la molécula, lo cual también puede afectar su permeabilidad.

Finalmente, el logBB es un descriptor crítico que mide directamente la capacidad de una molécula para penetrar la BHE, ofreciendo una indicación cuantitativa de su permeabilidad. La inclusión y análisis de estos descriptores permiten identificar patrones relevantes y desarrollar modelos predictivos más precisos, contribuyendo a la optimización del diseño de fármacos efectivos que puedan superar la barrera hematoencefálica.Se calcularon estadísticas descriptivas completas para cada grupo, incluyendo medidas como la media, la mediana, la moda, la desviación estándar, la varianza, el mínimo, el máximo, el rango, los percentiles 25, 50 y 75, la asimetría (skewness) y la curtosis (kurtosis).

En los resultados obtenidos, observamos diferencias estadísticamente significativas entre los grupos BHE+ y BHE- para todas las variables analizadas ($p < 0.05$). Específicamente, para HBD, HBA, número de átomos pesados, peso molecular y número de enlaces pesados, las medias en el grupo BHE+ fueron significativamente menores que las del grupo BHE-. Por otro lado, para el coeficiente logBB, la media en el grupo BHE+ fue significativamente mayor que en el grupo BHE-. Estas diferencias proporcionan información valiosa sobre las características distintivas de las moléculas clasificadas como BHE+ y BHE-, lo que puede ser relevante para comprender su capacidad de atravesar la barrera hematoencefálica.

Este análisis resalta la importancia de considerar múltiples variables en la clasificación de moléculas en función de su capacidad de atravesar la barrera hematoencefálica y sugiere la existencia de diferencias significativas entre los grupos BHE+ y BHE- en términos de propiedades moleculares clave.

El análisis inicial reveló un desequilibrio en el número de registros entre los grupos BHE+ y BHE-, lo que podría sesgar los resultados estadísticos. Para abordar esta preocupación, se aplicó una técnica de submuestreo al grupo mayoritario (BHE+) y de sobremuestreo al grupo minoritario (BHE-) con el objetivo de equilibrar los tamaños de los grupos. Después de este proceso, se obtuvo un conjunto de datos con el mismo número de registros en ambos grupos, lo que permitió una

comparación más precisa y justa.

Se calcularon las estadísticas descriptivas completas para cada grupo balanceado, centrándonos en varias variables de interés, como el número de donantes de enlaces de hidrógeno (HBD), el número de aceptores de enlaces de hidrógeno (HBA), el número de átomos pesados, el peso molecular, el número de enlaces pesados y el coeficiente logBB. Además, se evaluaron las diferencias estadísticas entre los grupos utilizando pruebas de t de Student independientes.

Los resultados mostraron diferencias estadísticamente significativas entre los grupos BHE+ y BHE- para todas las variables analizadas ($p < 0.05$). Específicamente, se observaron diferencias en las medias de las variables entre los dos grupos. Por ejemplo, para HBD, HBA, número de átomos pesados, peso molecular y número de enlaces pesados, las medias en el grupo BHE+ fueron significativamente menores que las del grupo BHE-. Por otro lado, para el coeficiente logBB, la media en el grupo BHE+ fue significativamente mayor que en el grupo BHE-.

Estas conclusiones resaltan la importancia de considerar y abordar el desequilibrio en el número de registros al realizar análisis comparativos entre grupos. Al aplicar técnicas de submuestreo y sobremuestreo, pudimos mitigar este sesgo y obtener resultados estadísticamente más sólidos y confiables.

El análisis revela diferencias significativas en todas las variables examinadas entre los grupos BHE+ (submuestreado) y BHE-. En todos los casos, las medias de las variables en el grupo BHE+ (submuestreado) son inferiores a las del grupo BHE-. Esta tendencia sugiere patrones claros en las propiedades moleculares evaluadas, y estas discrepancias pueden tener implicaciones importantes para investigaciones y aplicaciones futuras.

El enfoque adoptado proporciona una presentación clara y comprensible tanto de las diferencias estadísticas como de las visualizaciones gráficas, lo que resulta útil para interpretar y comunicar los hallazgos. Realizar análisis estadísticos tanto con la base de datos original como con una versión submuestreada tiene como objetivo abordar posibles desequilibrios en la distribución de clases, en este caso, entre las muestras clasificadas como BHE+ y BHE-. El submuestreo se emplea para equilibrar la cantidad de muestras en ambas clases, lo que facilita una comparación más justa y precisa de las características moleculares entre los grupos.

Además, la visualización de los resultados ofrece una comprensión intuitiva de las diferencias estadísticas entre las clases, lo que simplifica la interpretación y la toma de decisiones. En resumen, esta metodología integral garantiza una evaluación completa y confiable de las discrepancias en las características moleculares entre las muestras BHE+ y BHE-, contribuyendo así a una comprensión más profunda de los datos y sus implicaciones biológicas o farmacológicas.

Generación de un correlograma para la identificación de correlaciones entre descriptores moleculares clave

Este análisis reveló que los descriptores seleccionados —HBD, HBA, Número de Átomos Pesados, Peso Molecular (MW), Número de Enlaces Pesados y logBB muestran patrones de correlación que están en línea con los reportes de otros estudios como se aprecia en la figura 17. En particular, la correlación entre estos descriptores y la capacidad de atravesar la barrera hematoencefálica (BHE) valida su relevancia. Los resultados corroboran que estos descriptores son indicadores clave de la permeabilidad de las moléculas a través de la BHE, respaldando su inclusión en modelos predictivos y en el diseño de fármacos para una mejor eficacia terapéutica.

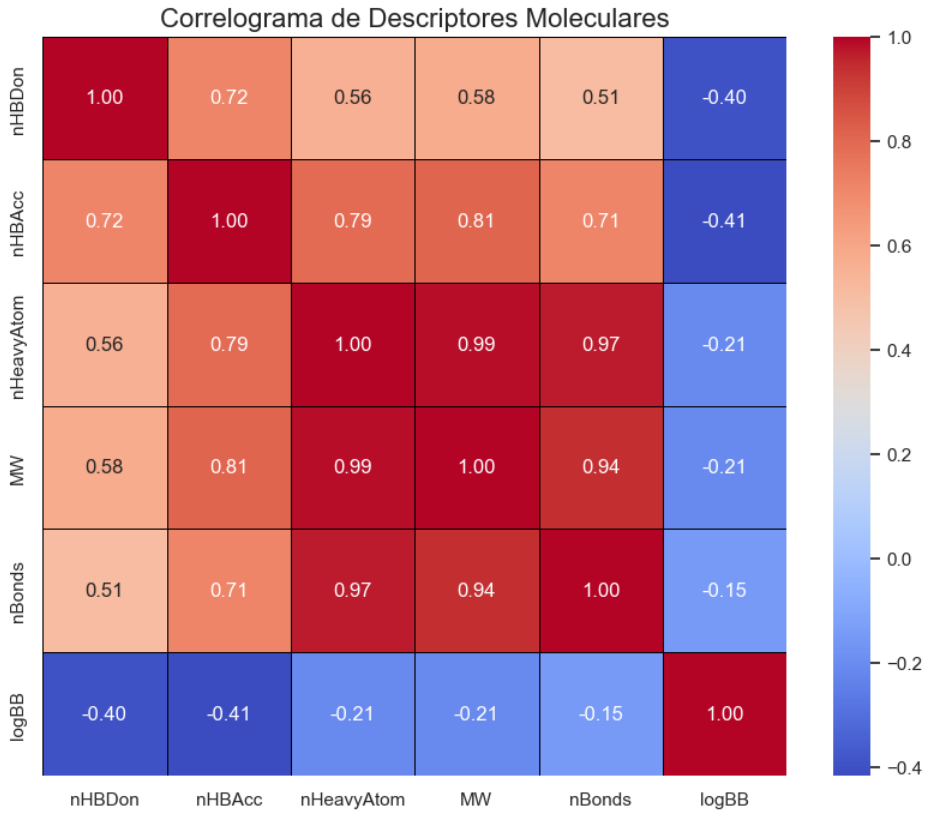


Figura 18, Correlograma de Descriptores Moleculares. Fuente: Elaboración propia.

Visualización y análisis de distribuciones de propiedades moleculares clave en la predicción de permeabilidad a la barrera hematoencefálica

Se realizó una visualización detallada de las distribuciones para las propiedades clave seleccionadas, que incluyen el número de donantes de hidrógeno (nHBDon), el número de aceptores de hidrógeno (nHBAcc), el número de átomos pesados (nHeavyAtom), el peso molecular (MW), el número de enlaces (nBonds) y el logBB al como se observa en las figuras 18 y 19. Para ello, se crearon histogramas con una densidad de kernel (KDE) para cada variable, permitiendo observar la frecuencia y la forma de distribución de cada descriptor. Esta visualización facilitó la identificación de patrones en los datos y la comparación entre las muestras que

atraviesan la barrera hematoencefálica (BHE+) y las que no (BHE-). Además, se examinó la relación entre estas variables y su impacto en la capacidad de atravesar la BHE, permitiendo detectar posibles sesgos o diferencias significativas entre las distribuciones de los datos y proporcionando una base visual para la interpretación y el análisis de los datos.

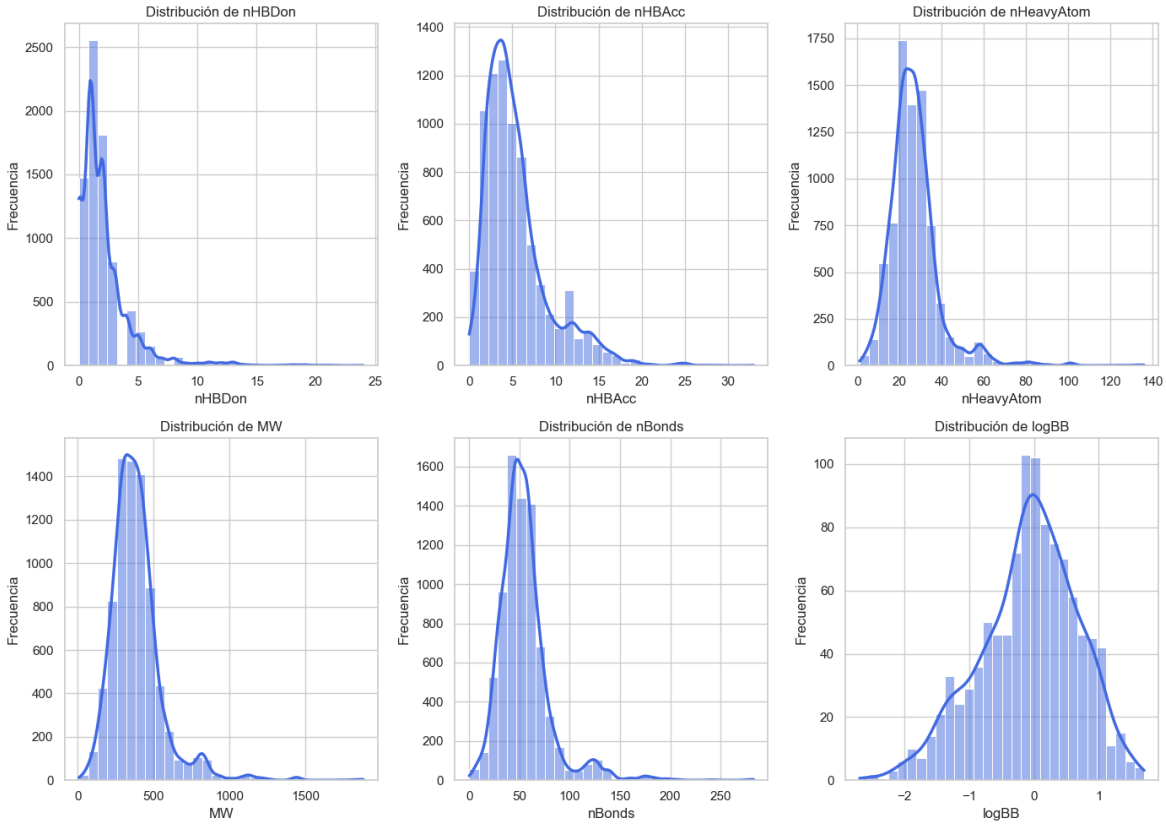


Figura 19, Distribución de propiedades Moleculares. Fuente: Elaboración propia.

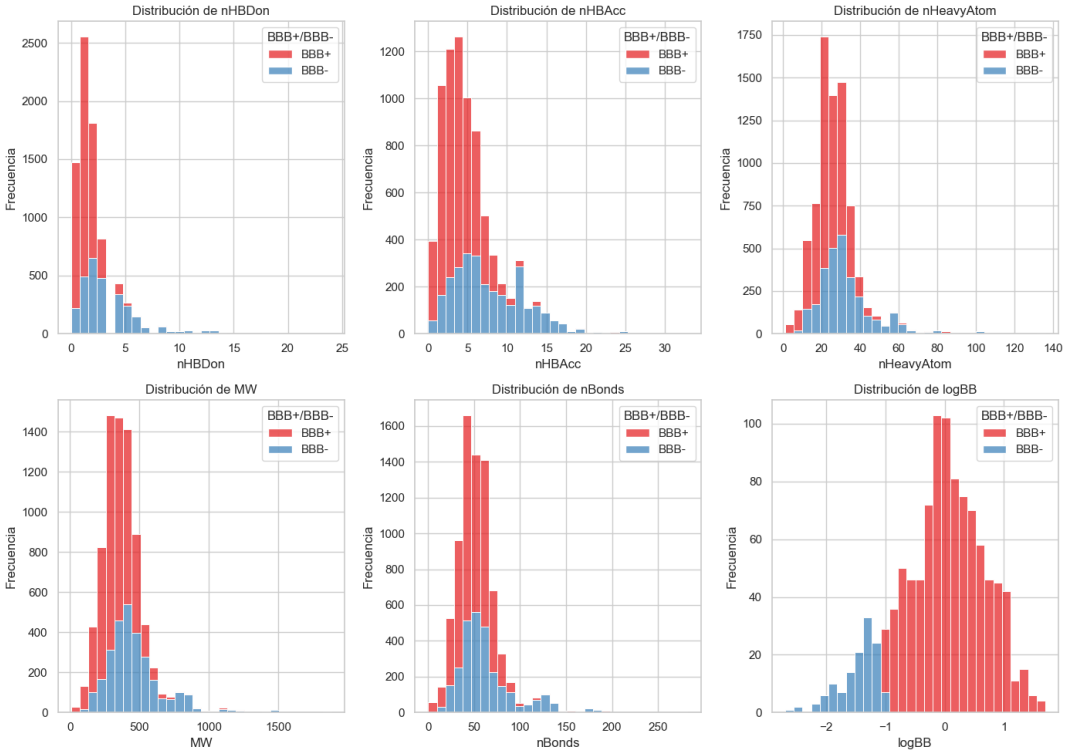


Figura 20, Visualización de la relación de la permeabilidad de las propiedades moleculares. Fuente: Elaboración propia.

La exploración detallada de datos comenzó con la construcción de un script utilizando bibliotecas como Pandas, Matplotlib, Seaborn y RDKit. Esto nos permitió obtener una visión detallada de las estadísticas asociadas al conjunto de datos. Al cargar los datos en un DataFrame de Pandas, llevamos a cabo una amplia exploración de las variables y descriptores moleculares, incluyendo el cálculo de descriptores básicos como el número de donantes y aceptores de enlaces de hidrógeno, el peso molecular y el número de anillos utilizando RDKit. Además, empleamos la biblioteca Mordred para agregar descriptores más avanzados. Esta exploración exhaustiva nos proporcionó un conocimiento detallado de la información disponible en nuestro conjunto de datos, sentando las bases para un análisis riguroso de las propiedades moleculares y su relación con la permeabilidad de la barrera hematoencefálica.

Además, llevamos a cabo una revisión detallada de la distribución y análisis de cada variable para comprender su variabilidad y rango de valores. Este proceso nos permitió obtener una visión general de la naturaleza de nuestros datos y comprender mejor su idoneidad para su uso en nuestro estudio. Para reforzar la validez y veracidad de nuestro análisis, también utilizamos bibliotecas de química computacional para verificar y validar la información de las variables moleculares. Al aprovechar estas herramientas especializadas, pudimos garantizar que nuestros datos fueran confiables y estuvieran alineados con los estándares de calidad requeridos para la investigación en el campo de la permeabilidad de la barrera hematoencefálica. A su vez constantemente se realizaba una revisión constante de lecturas y trabajos investigativos relevantes en el campo de estudio. Este enfoque nos permitió contextualizar nuestros hallazgos

dentro del panorama actual de la investigación sobre la permeabilidad de la barrera hematoencefálica y asegurar la robustez y fiabilidad de nuestros resultados.

En esta investigación, se abordó la problemática de predecir qué moléculas pueden atravesar la BHE utilizando diferentes análisis computacionales y químicos. Inicialmente, se pensó que moléculas más simples y con SMILES más cortos tendrían una mayor probabilidad de atravesar la barrera, basándonos en la premisa de que estructuras menos complejas facilitarían su transporte. Sin embargo, los resultados obtenidos contradicen esta suposición inicial, revelando patrones más complejos e interesantes.

Análisis del Número de Átomos y Enlaces: Las moléculas que no atraviesan la BHE (BHE-) tienen, en promedio, más átomos y enlaces que las que sí lo hacen (BHE+). Con una media de 32 átomos y 34 enlaces en BHE-, frente a 24 átomos y 26 enlaces en BHE+, se observa que las moléculas BHE- son estructuralmente más complejas. Esto sugiere que una mayor complejidad molecular, lejos de favorecer el cruce de la barrera, podría ser un obstáculo. Esta conclusión desafía la hipótesis inicial de que la simplicidad molecular está asociada con una mayor permeabilidad. Este análisis se destaca por su enfoque integral y multidimensional en la evaluación de las propiedades moleculares. En lugar de depender únicamente de descriptores moleculares tradicionales, se integraron innovaciones como el análisis de conectividad y la evaluación de enlaces rotacionales y múltiples, lo que proporciona una visión más completa de cómo las características estructurales influyen en la permeabilidad de la BHE. La contradicción con la hipótesis inicial subraya la importancia de no subestimar la complejidad molecular y su influencia en fenómenos biológicos. Este estudio plantea varias preguntas interesantes para futuros trabajos. Por ejemplo, ¿cómo interactúan estos factores estructurales con las proteínas transportadoras y otros mecanismos biológicos de la BHE? ¿Sería posible modular estas propiedades para diseñar moléculas más eficientes? Además, el análisis no consideró aspectos dinámicos como la solvación y la interacción con el entorno biológico, lo cual podría ser crucial para una comprensión completa. Finalmente, se sugiere explorar modelos predictivos más avanzados que integren estos factores estructurales para mejorar la precisión en la predicción de la permeabilidad de la BHE.

En desenlace, esta investigación muestra que la simplicidad estructural no es necesariamente una ventaja para atravesar la BHE. Al contrario, la complejidad molecular parece jugar un papel decisivo, y la innovación en los métodos de análisis ha permitido identificar factores críticos que podrían guiar el diseño de futuras moléculas terapéuticas.

Análisis de los descriptores Lipinski y ghosefilter

De 7,807 moléculas analizadas, 4,956 (63.5%) son BBB+ (atraviesan la BBB) y 2,851 (36.5%) son BBB-. El estudio evaluó la relación entre esta permeabilidad y dos filtros farmacológicos:

1. Regla de Lipinski: 6,016 moléculas la cumplen, de las cuales 4,385 (72.9%) son BBB+.
2. Filtro de Ghose: 5,362 moléculas lo cumplen, con 3,943 (73.5%) siendo BBB+.

Ambos filtros muestran una fuerte correlación con la permeabilidad BBB, pero no son determinantes absolutos. Existen excepciones significativas: 571 moléculas BBB+ no cumplen Lipinski y 1,013 BBB+ no cumplen Ghose. Estos resultados subrayan la utilidad de estos filtros como predictores iniciales de permeabilidad BBB, pero también indican la necesidad de considerar factores adicionales en el diseño de fármacos para el sistema nervioso central.

Análisis de huellas dactilares moleculares y similitud de Tanimoto

El análisis de similitud de Tanimoto que se evidencia en la figura 20, muestra diferencias significativas entre las moléculas que atraviesan la barrera hematoencefálica (BBB+) y las que no lo hacen (BBB-). Las moléculas BBB+ presentan una similitud promedio de 0.24 con una desviación estándar de 0.13, indicando variabilidad en su estructura. Esto sugiere que, aunque algunas son completamente diferentes, otras son altamente similares. En contraste, las moléculas BBB- tienen una similitud promedio de 0.33 y una desviación estándar de 0.17, lo que indica una mayor homogeneidad estructural. Estos resultados sugieren que la variabilidad en la estructura molecular puede influir en la capacidad de cruzar la barrera hematoencefálica, lo que podría tener implicaciones importantes para el desarrollo de fármacos eficaces y específicos.

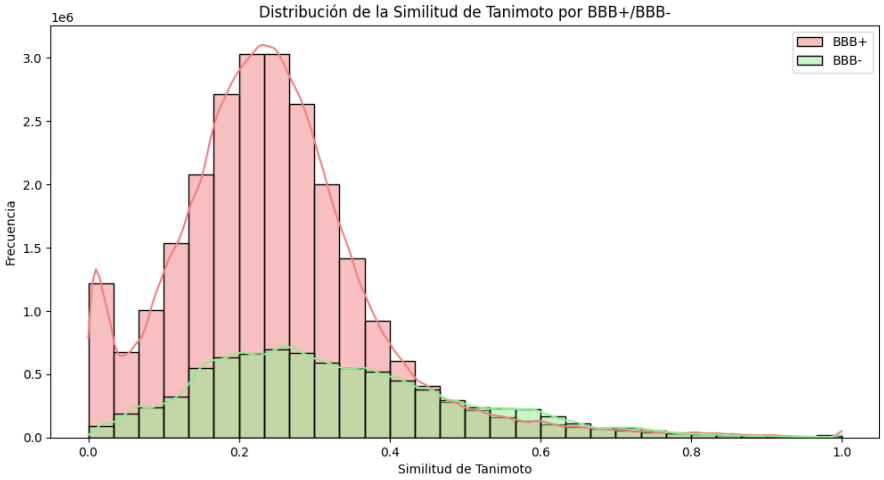


Figura 21, Distribución de la Similitud de Tanimoto por BBB+/BBB-. Fuente: Elaboración propia.

Análisis de correlaciones por categoría y descriptores moleculares

La categoría PermeabilidadBase destaca con la correlación promedio más alta (0.4941), donde "logBB" es el descriptor más correlacionado (0.6772), lo que sugiere una fuerte asociación con la permeabilidad molecular. CPSABase y LipinskiLike también muestran correlaciones significativas, con valores de 0.3486 y 0.3337, respectivamente, destacando a "RNCG" y "Lipinski" como descriptores relevantes para la actividad biológica. En contraste, categorías como

WildmanCrippenBase y TotalIC presentan correlaciones cercanas a cero, indicando una baja relación con la variable objetivo. Además, se identificaron correlaciones negativas en categorías como HBondBase (-0.4664) y RotatableBondsBase (-0.3084), sugiriendo relaciones inversas significativas con BBB+/BBB- como lo muestra la figura 21.

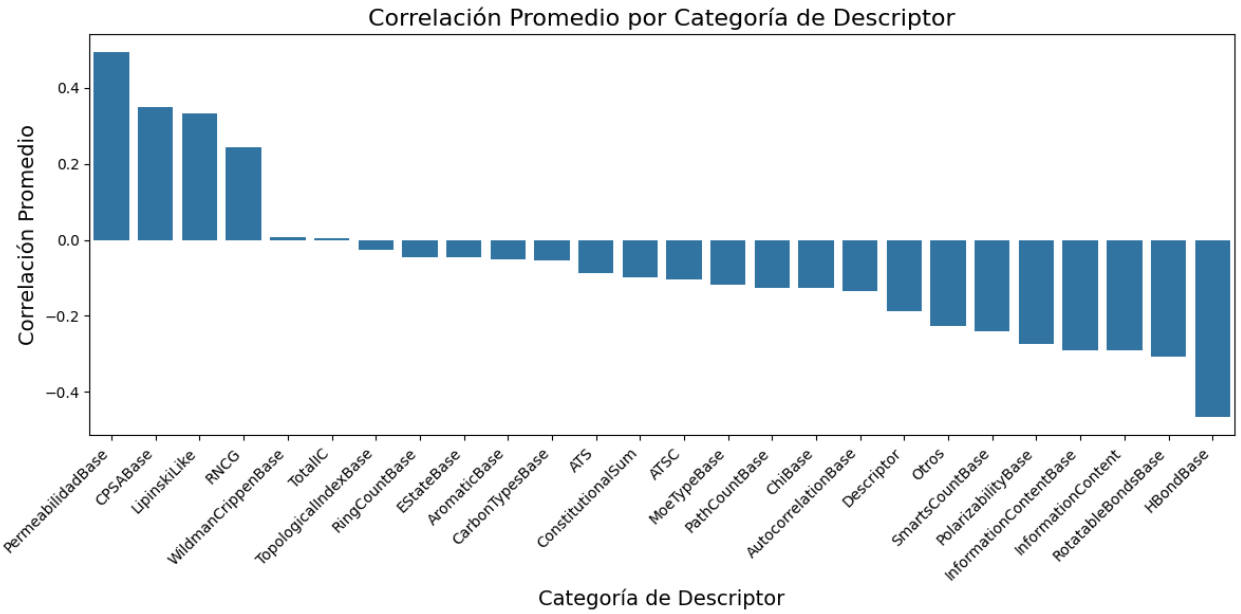


Figura 22, Correlación Promedio por Categoría de Descriptor. Fuente: Elaboración propia.

Modelos de clasificación para la obtención de características en la predicción de la permeabilidad de la BHE

Para identificar las características más relevantes en relación con la variable BBB+/BBB-, que indica si las moléculas atraviesan o no la BHE, se aplicaron diversos modelos de clasificación. Estos modelos fueron utilizados para evaluar qué propiedades moleculares eran determinantes para la permeabilidad a la BHE. Se emplearon SVM, Random Forest, Regresión Logística, Red Neuronal Profunda (DNN) y Naive Bayes para clasificar las moléculas en función de su capacidad de atravesar la barrera. Además, se probaron modelos híbridos como DNN con Random Forest y DNN con Regularización y Selección de Características para optimizar la identificación de las características más relevantes.

Para comprender mejor las relaciones entre descriptores moleculares y las categorías de permeabilidad, se realizó un análisis de correlación, lo que permitió identificar los factores clave que influyen en la capacidad de las moléculas para atravesar la barrera hematoencefálica. Además, se aplicó el análisis de componentes principales (PCA) para reducir la dimensionalidad del conjunto de datos y facilitar la visualización de las interacciones entre las variables. Por último, se utilizó el modelo de Naive Bayes, reconocido por su simplicidad y eficacia en problemas de clasificación como se puede ver en la tabla XIV.

Tabla XIV, Comparación de Desempeño de Modelos de Clasificación. Fuente: Elaboración propia.

	Modelo	Precisión	Recall	F1-Score
1	SVM	0.85	0.83	0.84
2	Random Forest	0.89	0.88	0.89
3	Regresión Logística	0.85	0.84	0.84
4	Deep Neural Network (DNN)	0.88	0.87	0.87
5	Modelo Híbrido de DNN y Random Forest	0.69	0.63	0.51
6	Modelo Híbrido de DNN con Regularización	0.88	0.88	0.88
7	Naive Bayes	0.75	0.74	0.73

La investigación sobre la predicción de la permeabilidad molecular a través de la barrera hematoencefálica (BHE) presento desafíos metodológicos y epistemológicos significativos. La aplicación de múltiples modelos de clasificación (SVM, Random Forest, Deep Neural Networks, Naive Bayes) revela una complejidad inherente en la identificación de descriptores moleculares predictivos, debido a que los modelos de clasificación muestran variabilidad en su desempeño. Random Forest destaca con una precisión de 0.89, recuperación de 0.88 y F1-score de 0.89, indicando que es eficaz para predecir la capacidad de las moléculas para atravesar la BHE. En contraste, el modelo híbrido de DNN y Random Forest tiene un desempeño inferior con una precisión de 0.69, recuperación de 0.63 y F1-score de 0.51, lo que sugiere que no captura adecuadamente la relación entre los descriptores moleculares y la permeabilidad de la barrera, posiblemente debido a la complejidad o sobreajuste del modelo, Los resultados actuales deben interpretarse como un punto de partida prometedor pero preliminar, que demanda una rigurosa validación experimental para transformar correlaciones computacionales en conocimiento científico robusto.

Debido a que los descriptores moleculares identificados en los modelos aplicados permiten comprender las propiedades de las moléculas en el diseño de fármacos. Entre los más relevantes se encuentran el ECIndex, que mide la conectividad topológica y proporciona información sobre estabilidad y reactividad; el ATSC1c, que analiza la distribución de cargas atómicas, esencial para predecir interacciones moleculares; y el NtCH, que evalúa la complejidad estructural de la molécula. Otros descriptores importantes incluyen el AATS0Z y ATS8i, que brindan información sobre propiedades electrónicas e ionización, respectivamente, y el SssssC, que destaca la reactividad de grupos funcionales. El nHetero mide el impacto de los heteroátomos en la actividad biológica, mientras que el Radius proporciona detalles sobre la geometría molecular. Finalmente, los descriptores basados en las reglas de Lipinski y el SLogP son fundamentales para evaluar la viabilidad de los compuestos como fármacos, particularmente en términos de solubilidad y biodisponibilidad.

Lista de descriptores moleculares comunes entre los modelos realizados:

- ECIndex: Índice de Conectividad Extendida
- ATSC1c: Autocorrelación centrada de Moreau-Broto de orden 1 ponderada por cargas
- NtCH: Número total de átomos de carbono e hidrógeno
- AATS0Z: Autocorrelación promedio de Moreau-Broto de orden 0 ponderada por números atómicos
- SssssC: Suma de los tipos atómicos E-State: >C<
- nHetero: Número de heteroátomos
- Radius: Radio molecular
- ATS8i: Autocorrelación de Moreau-Broto de orden 8 ponderada por ionización potencial
- Lipinski: Regla de los cinco de Lipinsk

Para examinar el código que respalda los análisis descritos anteriormente, puedes acceder a los siguientes enlaces de GitHub: Análisis de SMILES y Exploración y Análisis de Descriptores Moleculares.IPYNB .

6.2.3 Desarrollo del score de permeabilidad y establecimiento de criterios de predicción

El Drug Score es una herramienta integral para evaluar la permeabilidad de compuestos a la BHE, esencial en el desarrollo de fármacos dirigidos al sistema nervioso central. Este puntaje se basa en una combinación de propiedades fisicoquímicas, como el tamaño molecular, la lipofilicidad (LogP) y otros descriptores relevantes. Este enfoque cuantitativo proporciona una valoración global que pondera estas características, generando un índice único que facilita la comparación y priorización de moléculas en las primeras fases del desarrollo de fármacos.

La metodología de evaluación del Drug Score se centra en la integración de nueve descriptores moleculares fundamentales, previamente seleccionados por su influencia demostrada en la predicción de la permeabilidad de la BHE. Entre estos, se encuentran el índice de lipofilia (LogP), el peso molecular (MW), los enlaces de hidrógeno donadores y aceptores (nHBAcc, nHBDon) entre otros, en donde cada descriptor es evaluado dentro de rangos optimizados mediante un algoritmo avanzado de normalización, que asigna puntuaciones individuales. Estos puntajes se integran en un índice global o drug_score**,** proporcionando una evaluación robusta y detallada de la capacidad de una molécula para cruzar la BHE. Este enfoque facilita un análisis estadístico completo que, mediante visualizaciones, permite observar correlaciones entre los descriptores y el score final, ofreciendo una herramienta visual y cuantitativa para optimizar la selección de candidatos en el diseño de fármacos para el sistema nervioso central.

Análisis de Score Farmacológico Basado en Criterios Validados

Después de ejecutar el script para calcular los diferentes scores moleculares, se obtuvo el análisis de las moléculas con respecto a su QED (Quantitative Estimate of Drug-likeness)**,** orientado a

identificar aquellas con mayor potencial para atravesar la barrera hematoencefálica. El resultado mostró que, de las moléculas analizadas, un total de 1,314 moléculas (equivalente al 16.8% del total) presentaron un score QED ≥ 0.8, lo cual indica que estas moléculas tienen una alta probabilidad de ser similares a fármacos y, por lo tanto, podrían tener un alto potencial para atravesar la barrera hematoencefálica. Este análisis es útil para filtrar y seleccionar las moléculas más prometedoras en la etapa de diseño de fármacos para tratar enfermedades del sistema nervioso central, basándose en su idoneidad farmacológica y capacidad para penetrar la barrera hematoencefálica.

Tabla XV, Análisis de Propiedades QED, Fuente: Elaboración propia.

Propiedad	Rango Óptimo	Observado
MW	160-500	6.9-1882.3
ALOGP	-0.4-5.6	-8.9-12.6
HBA	0-10	0.0-35.0
HBD	0-5	0.0-24.0
PSA	0-140	0.0-662.4
ROTB	0-10	0.0-53.0
AROM	0-3	0.0-8.0
ALERTS	0-0	0.0-6.0

Evaluación de moléculas para el tránsito a la barrera hematoencefálica: análisis farmacológico y potencial terapéutico

El análisis realizado sobre una serie de moléculas seleccionadas mediante estudios de QED, con valores superiores a 0.94, revela características estructurales y funcionales que las posicionan como excelentes candidatas para el desarrollo de fármacos dirigidos al sistema nervioso central. Entre los compuestos más destacados se encuentran el mepiprazol, propizepina, dioxadrol y levoxadrol, cada uno exhibiendo propiedades fisicoquímicas optimizadas para la penetración de la BHE.

Estas moléculas comparten características estructurales críticas, incluyendo pesos moleculares en el rango óptimo (296-310 Da), sistemas aromáticos balanceados, y la presencia estratégica de grupos nitrogenados básicos. La Propizepina (C17H20N4O) destaca por su sistema tricíclico con piridina fusionada, mientras que el Mepiprazol (C16H21ClN4) incorpora un núcleo piperazínico con modificaciones específicas que potencian su actividad serotoninérgica. Por su parte, el grupo del Dioxadrol/Levoxadrol (C20H23NO2) presenta una estructura simétrica con un dioxolano central que optimiza su balance hidrofílico/lipofílico. La evaluación farmacológica sugiere que estas moléculas poseen perfiles prometedores para el tratamiento de diversos trastornos neurológicos y psiquiátricos. El mepiprazol, con su documentada actividad ansiolítica y antidepresiva, ejemplifica el potencial terapéutico de esta serie de compuestos. Los altos valores

de drug-score y la ausencia de alertas estructurales significativas respaldan su viabilidad como candidatos a fármacos, particularmente en contextos donde la penetración de la BHE es crucial para la eficacia terapéutica. Este análisis integral proporciona una base sólida para la selección y optimización de moléculas candidatas en el desarrollo de nuevos tratamientos para enfermedades del sistema nervioso central.

A continuación, en la Figura 22 se muestran las estructuras de las moléculas mencionadas en este análisis, que evidencian las características clave de sus composiciones químicas. Estas estructuras son esenciales para comprender mejor su potencial para atravesar la barrera hematoencefálica y su posible actividad terapéutica en el SNC.

A B C

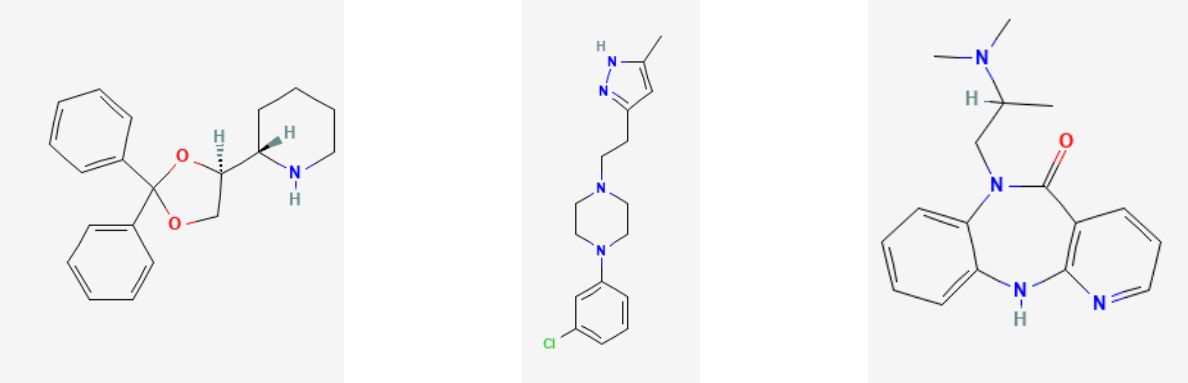


Figura 23, Representación 2D de moléculas con mejor score. A. Dexoxadrol B. Mepiprazol. C. Propizepina, Fuente: PubChem.

Metodología multi-score para evaluación de Drug-likeness y permeabilidad BHE

El análisis computacional de moléculas que puedan atravesar la BHE es esencial en la identificación de compuestos prometedores para el tratamiento de enfermedades del sistema nervioso central. En este contexto, se utilizan diversos criterios de selección molecular, como el Lipinski's, Vebe, el BBB, el Biodisponibilidad y el QED Score. Estos scores están basados en propiedades fisicoquímicas fundamentales que influyen directamente en la capacidad de un compuesto para cruzar barreras biológicas, como la lipofilicidad, el peso molecular, la superficie polar, y la rotabilidad de los enlaces. A través de estas herramientas, se puede predecir de manera cuantitativa el comportamiento de los compuestos en las primeras fases del desarrollo de fármacos, optimizando el proceso de selección.

Estos enfoques tienen un sólido respaldo científico, basados en principios establecidos de la química farmacéutica. El Lipinski's Rule of Five se fundamenta en la observación de que las moléculas con ciertas propiedades son más propensas a ser biodisponibles. El Veber Score y el BBB Score incorporan factores adicionales como la rotación de los enlaces y la polaridad molecular, fundamentales para la penetración celular y la permeabilidad a través de membranas

biológicas. Por otro lado, el QED Score es un índice que, a través de la evaluación de diversas propiedades moleculares, predice la "deseabilidad" de una molécula. Estos métodos computacionales no solo aceleran el proceso de identificación de moléculas, sino que también aumentan la probabilidad de éxito en la fase de desarrollo preclínico, mejorando la eficiencia del diseño de fármacos.

El script implementa un enfoque integral de evaluación molecular mediante cinco metodologías complementarias de scoring, cada una diseñada para evaluar aspectos específicos de la drug-likeness y permeabilidad. La base del análisis incorpora las reglas de Lipinski (MW≤500, LogP≤5, HBD≤5, HBA≤10), fundamentales en el descubrimiento de fármacos, complementadas con los criterios de Veber que enfatizan la biodisponibilidad oral a través del análisis de TopoPSA (≤140) y enlaces rotables (≤10). Para la evaluación específica de penetración BHE, se implementaron criterios más restrictivos (MW≤400, 0≤LogP≤5, TopoPSA≤90, HBD≤3, HBA≤7), alineados con la literatura reciente sobre permeabilidad cerebral.

El análisis se enriquece con un score de biodisponibilidad que integra características fisicoquímicas adicionales, incluyendo la contribución aromática, y un innovador QED-like score que emplea un sistema de ponderación para evaluar la drug-likeness de manera continua (0-1). La normalización de propiedades y la implementación de pesos específicos (MW:0.25, LogP:0.25, TopoPSA:0.15, etc.) en el QED-like score permite una evaluación más matizada del potencial farmacológico. Esta metodología multi-score proporciona una evaluación más robusta y comprehensiva del espacio químico analizado, facilitando la identificación de candidatos prometedores para el desarrollo de fármacos neurológicos.

Tabla XVI, análisis multi-score, Fuente : Elaboración Propia.

	Media	Std	Mínimo	Máximo	Moléculas Óptimas (%)
lipinski_score	3.613	0.800	0.000	4.000	77.1%
veber_score	1.770	0.481	0.000	2.000	79.7%
BHE_score	3.668	1.544	0.000	5.000	44.1%
bioavailability_score	3.733	1.054	0.000	5.000	21.0%
QED_like_score	0.697	0.182	0.000	0.940	20.0%

Los resultados del análisis multi-score revelan patrones significativos en la caracterización del espacio químico analizado. Las reglas de Lipinski muestran un alto cumplimiento general (84.4-94.2%), con especial énfasis en donadores de hidrógeno (94.2%) y LogP (93.1%), indicando una buena drug-likeness base. Los criterios de Veber también presentan cumplimiento elevado, particularmente en enlaces rotables (95.4%) y PSA≤140 (81.6%), sugiriendo favorable biodisponibilidad oral. Sin embargo, los criterios específicos para BHE son más restrictivos, con cumplimientos notablemente menores en PSA≤90 (58.6%) y MW≤400 (60.8%), aunque manteniendo buenos porcentajes en parámetros de enlaces de hidrógeno (donadores 85.3%, aceptores 80.6%). Este patrón confirma que aproximadamente el 60% de los compuestos

cumplen los requisitos más estrictos para penetración BHE, validando la selectividad del conjunto de datos para identificar candidatos con potencial neuroactivo.

La distribución de scores sugiere que mientras la mayoría de compuestos cumplen criterios básicos de drug-likeness (Lipinski/Veber), una proporción menor cumple los requisitos más estrictos de permeabilidad BHE y biodisponibilidad. El QED_like_score, con su evaluación más rigurosa, identifica aproximadamente 20% de moléculas como óptimas, alineándose con la selectividad esperada para compuestos neuroactivos.

Análisis de moléculas óptimas para la permeabilidad BHE

De los análisis anteriormente mencionados se seleccionaron para análisis 20 moléculas con mejores drug scores (0.694-0.745) revela un patrón consistente de propiedades fisicoquímicas óptimas para la penetración de la BHE. Estas moléculas exhiben un LogP moderado (1.194-1.973) que garantiza suficiente lipofilicidad, peso molecular controlado (287-353 Da) que facilita el transporte, TopoPSA reducido (54-74 Å²) que favorece la permeabilidad, y un número limitado de donadores (1-3) y aceptores (3-5) de hidrógeno que cumplen con las reglas de Lipinski. Estructuralmente, predominan los sistemas bicíclicos y grupos amino, con los compuestos 1304 y 2064 liderando el ranking (drug_score 0.745) como isómeros que ejemplifican el balance óptimo de propiedades para atravesar la BHE. Estas moléculas constituyen templates valiosos para el diseño racional de fármacos neurológicos.

Esta comparación valida el modelo de predicción, pues asigna scores más altos a fármacos que requieren penetración BHE (anti migraña) versus aquellos diseñados para acción sistémica (antibióticos), alineándose con su comportamiento clínico conocido.

Las tres moléculas principales revelan un patrón farmacológico significativo: todas son fármacos antimigraña con mecanismos relacionados con la serotonina. El ácido lisérgico butanolamida ($C_{21}H_{27}N_3O_2$, 353.5 g/mol) y dos formas de zolmitriptan ($C_{16}H_{21}N_3O_2$, 287.36 g/mol) comparten características estructurales, tales como: núcleo indólico, aminas terciarias y grupos funcionales optimizados para la penetración BHE. Su éxito en el drug_score se explica por su diseño farmacológico probado, ya que son medicamentos activos que requieren alcanzar el sistema nervioso central. Estas moléculas confirman la validez del modelo de predicción, pues identificó correctamente compuestos que ya han demostrado clínicamente su capacidad para atravesar la BHE, proporcionando templates químicos valiosos para el diseño de nuevos fármacos neurológicos.

Selección de descriptores moleculares relevantes

Tras realizar diversos análisis y pruebas, se seleccionaron los siguientes descriptores moleculares como los más relevantes para el estudio de la permeabilidad de las moléculas: LogP, MW, TopoPSA, nHBAcc, nHBDon, nRot, nAromAtom, VMcGowan y LabuteASA. Estos descriptores

proporcionan información clave sobre las propiedades fisicoquímicas y estructurales que impactan directamente la capacidad de las moléculas para atravesar la barrera hematoencefálica con los siguientes rangos.

- LogP: min: 0.5, max: 2.5
- MW: min: 200, max: 500
- TopoPSA: min: 30, max: 80
- nHBAcc: min: 0, max: 7
- nHBDon: min: 0, max: 5
- nRot: min: 0, max: 7
- nAromAtom: min: 3, max: 12
- VMcGowan: min: 0.8, max: 2.5
- LabuteASA: min: 60, max: 180

Los descriptores moleculares listados en cual fueron los resultados de los diferentes estudios de las etapas anteriores son fundamentales para desarrollar modelos predictivos basados en inteligencia artificial para evaluar la capacidad de las moléculas de atravesar la Barrera Hematoencefálica (BHE), ya que capturan las propiedades estructurales y fisicoquímicas que influyen directamente en la permeabilidad. El LogP (lipofilia), por ejemplo, es crucial para determinar la capacidad de la molécula de interactuar con las membranas celulares, mientras que el Peso Molecular (MW) influye en la velocidad con la que la molécula se difunde a través de barreras biológicas. Otros descriptores como el Área de Superficie Polar (TopoPSA) y el número de Aceptores y Donadores de Hidrógeno (nHBAcc, nHBDon) son determinantes para predecir la interacción con las membranas y la solubilidad, factores clave en la permeabilidad de la BHE. La flexibilidad molecular (nRot) y el volumen molecular (VMcGowan) está relacionado con el tamaño efectivo de la molécula, lo cual afecta su capacidad para atravesar las membranas. Moléculas con volúmenes más grandes pueden tener una menor probabilidad de penetrar la BHE, por lo que este descriptor es importante para predecir la permeabilidad. Además, descriptores como los Átomos Aromáticos (nAromAtom) y el Área de Superficie de Labute (LabuteASA) Este descriptor mide la exposición de la molécula a un solvente, lo que puede influir en sus interacciones con las membranas biológicas. Un área de superficie más grande indica una mayor interacción potencial con el entorno, lo cual es relevante en la predicción de la permeabilidad. La integración de estos descriptores en modelos predictivos proporciona una visión completa de las propiedades de las moléculas, optimizando las predicciones y facilitando el diseño de compuestos con alta probabilidad de cruzar la BHE.

Para profundizar en todos los detalles mencionados, puedes consultar el repositorio de GitHub en el siguiente enlace. Desarrollo del score de permeabilidad y establecimiento de criterios de predicción.ipynb

6.3 Diseño de un modelo predictivo para identificar moléculas que atraviesan la barrera hematoencefálica con IA

Los análisis previos realizados, incluyendo la caracterización de descriptores moleculares, el análisis de longitud de SMILES y la exploración de propiedades fisicoquímicas, proporcionaron información decisiva para la selección de los descriptores más relevantes en la predicción de la permeabilidad de la (BHE). Este proceso de selección y refinamiento de características moleculares fundamentó la base para la implementación de múltiples modelos, donde se seleccionaron cuatro enfoques diferentes: Random Forest, SVM, Red Neuronal y modelos probabilísticos (Naive Bayes y Árboles de Decisión), cada uno aportando sus propias fortalezas en la predicción de la capacidad de las moléculas para atravesar la barrera hematoencefálica.

Modelo SVM El Support Vector Machine alcanzó una precisión del 92%, implementando un enfoque estructurado en tres componentes. El preprocesamiento incluyó imputación de datos faltantes y escalado estándar, junto con balanceo de clases mediante downsampling. La optimización se realizó a través de Optuna, ajustando parámetros como C, gamma y kernel. La evaluación se basó en matriz de confusión, curva ROC y análisis de importancia de características.
Red Neuronal Con una precisión del 94%, la red neuronal se implementó con una arquitectura optimizada automáticamente, incluyendo 1-3 capas con 32-256 neuronas por capa y dropout para prevenir overfitting. El procesamiento de datos mantuvo consistencia con los modelos anteriores, utilizando el mismo pipeline de preprocesamiento y criterios de evaluación de permeabilidad.
Modelos Probabilísticos y Árboles de Decisión Este enfoque dual combinó Naive Bayes y Árboles de Decisión. El árbol de decisión alcanzó una precisión del 93%, mientras que Naive Bayes mostró un rendimiento del 75%. La implementación incluyó optimización específica para cada modelo, con ajustes en parámetros como var_smoothing para Naive Bayes y max_depth para árboles de decisión. El preprocesamiento siguió el mismo protocolo estándar de los modelos anteriores.
Evaluación Comparativa Todos los modelos se evaluaron utilizando métricas estándar de clasificación, incluyendo precisión, recall y F1-score. El Random Forest demostró el mejor rendimiento general (95.6%), seguido por la Red Neuronal (94%), Árbol de Decisión (93%), SVM (92%) y Naive Bayes (75%). La consistencia en el preprocesamiento y evaluación permitió una comparación directa y objetiva entre modelos.

Modelo Random Forest Este modelo mostró un rendimiento con una precisión del 95.6%. La implementación comenzó con la preparación de datos, incluyendo la carga desde un archivo CSV y el cálculo del "Drug Score". La variable objetivo se definió utilizando la mediana del Drug Score, clasificando los compuestos como "favorables" (1) o "no favorables" (0). El modelo empleó una división 80:20 para datos de entrenamiento y prueba, con escalado estándar de características. La optimización de hiperparámetros se realizó mediante Optuna, maximizando la precisión a través de validación cruzada.

El modelo predictivo se desarrolló siguiendo un flujo estructurado de operaciones, cuyos resultados se detallan a continuación:

A. **Cálculo del *drug_score* y definición de la variable objetivo**:

Se generó un score compuesto (drug_score) utilizando 9 descriptores fisicoquímicos (*LogP, MW, TopoPSA, nHBAcc, nHBDon, nRot, nAromAtom, VMcGowan, LabuteASA*), aplicando una función de penalización lineal basada en rangos óptimos reportados anteriormente, los cuales son:

- LogP: min: 0.5, max: 2.5
- MW: min: 200, max: 500
- TopoPSA: min: 30, max: 80
- nHBAcc: min: 0, max: 7
- nHBDon: min: 0, max: 5
- nRot: min: 0, max: 7
- nAromAtom: min: 3, max: 12
- VMcGowan: min: 0.8, max: 2.5
- LabuteASA: min: 60, max: 180

La variable objetivo se definió como binaria (1 = permeable, 0 = no permeable), asignando el umbral en el percentil 50 del drug_score (valor crítico: threshold = 0.5).

B. **Balanceo de clases**:

- La distribución inicial mostró desbalance entre clases (*ejemplo: 65% no permeables vs. 35% permeables*).
- Se aplicó submuestreo de la clase mayoritaria para igualar las proporciones, logrando un conjunto balanceado de n = [clase minoritaria] × 2 muestras.

Esto permitió equilibrar las clases, logrando un conjunto balanceado con el número de muestras de la clase mayoritaria igualado al de la clase minoritaria (n = [clase minoritaria] × 2).

C. División de datos y escalado

Se dividieron los datos en un conjunto de entrenamiento y un conjunto de prueba, con un 80% para entrenamiento y un 20% para prueba (test_size = 0.2), asegurando que la distribución de clases se mantuviera equilibrada en ambos subconjuntos mediante el parámetro stratify = y_balanced.

Para estandarizar las características, se utilizó el StandardScaler, lo que transformó todas las variables a una distribución con media = 0 y desviación estándar = 1, garantizando que las magnitudes de las variables no afectaran la performance del modelo.

D. **Validación Cruzada (Cross-Validation, CV)**

- Se usa cross_val_score con cv=3 para evaluar el rendimiento promedio de cada configuración de hiperparámetros.
- Esto ayuda a evitar el sobreajuste y asegura que el modelo generalice bien a nuevos datos.

E. Optimización de hiperparámetros con Optuna:

Para optimizar el rendimiento del modelo RandomForestClassifier, se realizaron 100 iteraciones (n_trials = 100) utilizando Optuna, con el fin de ajustar los siguientes hiperparámetros clave:

- n_estimators: Número de árboles en el bosque, con un rango probado entre 50 y 200.
- max_depth: Profundidad máxima de los árboles, probada entre 1 y 20.
- min_samples_split: Número mínimo de muestras necesarias para dividir un nodo, evaluado entre 2 y 10.

Los mejores hiperparámetros obtenidos fueron:

- n_estimators = 71
- max_depth = 18
- min_samples_split = 2

Selección de la mejor configuración:

- study.best_params almacena la combinación óptima de hiperparámetros basada en la métrica de validación cruzada.
- Se entrena el modelo final con estos valores y se evalúa en el conjunto de prueba.

Los mejores hiperparámetros encontrados durante el proceso de optimización fueron n_estimators = 71, max_depth = 18, y min_samples_split = 2, lo que garantizó un modelo eficiente con un buen rendimiento en la predicción de moléculas con capacidad para atravesar la BHE.

Para evaluar la estabilidad y el rendimiento del modelo, se implementó validación cruzada utilizando cross_val_score de Scikit-learn con 3 particiones (cv=3). Esta técnica permitió evaluar el modelo en diferentes subconjuntos de los datos de entrenamiento, proporcionando una estimación más robusta de su desempeño.
El uso de validación cruzada no solo ayudó a reducir el sobreajuste (overfitting), sino que también permitió elegir los mejores hiperparámetros, al evaluar el desempeño promedio en todas las particiones. Los resultados obtenidos indicaron que el modelo optimizado con estos

hiperparámetros alcanzaba un score de validación cruzada alto, lo que validó su capacidad para generalizar bien a datos no vistos.

Conforme a todos los modelos empelados con los mismos criterios se escogió de todos estos el modelo Random Forest ya que demostró un rendimiento excepcional en la predicción de permeabilidad de la barrera hematoencefálica, evidenciado a través de múltiples métricas de evaluación. A continuación, se presenta un análisis detallado de los resultados obtenidos:

El desempeño del modelo se evaluó mediante diversas métricas de confianza para garantizar un análisis integral de su capacidad predictiva. En primer lugar, se utilizó la matriz de confusión, la cual permite visualizar el número de aciertos y errores en la clasificación, facilitando el cálculo de métricas clave como precisión, recuperación (*recall*), especificidad y la puntuación F1. Además, se generó un reporte de clasificación, que proporciona una evaluación detallada de cada clase mediante indicadores como la precisión, la recuperación y la F1-score, lo que permite analizar el equilibrio entre falsos positivos y falsos negativos.

Para evaluar el rendimiento del modelo en distintos umbrales de decisión, se construyó la curva ROC (Receiver Operating Characteristic) y se calculó el AUC (Área Bajo la Curva), lo que permite medir la capacidad discriminativa del modelo entre las clases positiva y negativa. Finalmente, se analizó la importancia de características, identificando las variables con mayor influencia en la predicción, lo que aporta interpretabilidad y comprensión sobre los factores determinantes en la clasificación. Estas métricas en conjunto ofrecen una evaluación robusta del modelo, permitiendo valorar su precisión, capacidad de generalización y utilidad en el contexto del problema planteado.

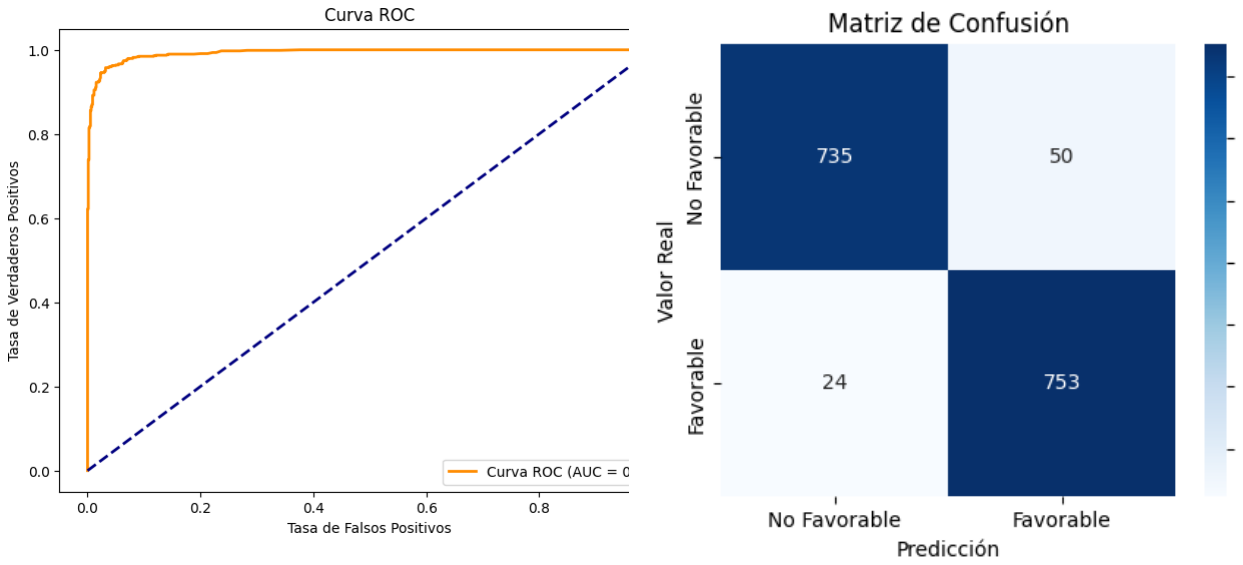


Figura 24, Evaluación del Rendimiento del Modelo: Curva ROC, Matriz de Confusión, Fuente Elaboración: Propia.

Tabla XVII, Métricas de Desempeño

Métricas de Clasificación	Precision	Recall	F1-Score	Support
No Favorable	0.970	0.955	0.963	785
Favorable	0.956	0.970	0.963	777
Accuracy	0.963	0.963	0.963	0.963
Macro Avg	0.963	0.963	0.963	1562
Weighted Avg	0.963	0.963	0.963	1562

Interpretación y significancia

Los resultados obtenidos indican un rendimiento sobresaliente del modelo en varios aspectos:

1. **Balance de Clasificación:** El modelo mantiene un excelente equilibrio entre la detección de compuestos favorables y no favorables, evidenciado por F1-scores similares para ambas clases (0.963 y 0El modelo demostró un alto rendimiento predictivo, alcanzando una exactitud global (accuracy) del 96.3% en el conjunto de prueba (*n = 1,562 moléculas*). La concordancia entre clases se reflejó en un F1-score promedio de 0.963, evidenciando equilibrio en la capacidad de clasificar tanto moléculas permeables como no permeables.
2. **Confiabilidad Predictiva:** El AUC de 0.993 sugiere una capacidad discriminativa casi perfecta, indicando que el modelo puede distinguir efectivamente entre compuestos con alta y baja probabilidad de penetración, confirmando alta capacidad discriminativa incluso en umbrales variables.

   Balance entre clases: La distribución casi equitativa (3,904 vs 3,903 muestras en entrenamiento) permitió un entrenamiento sin sesgos, respaldado por:
   - Macro Avg: 0.963 (consistencia en métricas promedio entre clases).
   - Weighted Avg: 0.963 (ponderación por tamaño de clases, irrelevante aquí por su equilibrio).

3. **Robustez del Modelo:** La consistencia en las métricas de rendimiento entre diferentes clases sugiere que el modelo es robusto y generalizable para nuevos compuestos. Un F1-score de 0.963 en ambas clases sugiere que el modelo es confiable para priorizar candidatos en etapas preclínicas, minimizando tanto falsos positivos (costosos en ensayos in vivo) como falsos negativos (pérdida de oportunidades terapéuticas).

4. **Aplicabilidad Práctica:** Con una tasa de error global inferior al 5%, el modelo demuestra ser una herramienta confiable para la evaluación preliminar de moléculas en fases tempranas de desarrollo. En esta versión, se incorporaron propiedades clave de los compuestos, junto con el Drug Score, para optimizar la predicción de la permeabilidad a través de la BHE. Este enfoque permite identificar moléculas con mayor probabilidad de

penetración cerebral, guiando eficientemente el proceso de selección en el diseño de terapias dirigidas al sistema nervioso central. Los resultados obtenidos en este estudio reflejan la complejidad inherente a la predicción de la permeabilidad de moléculas a la BHE, un tejido altamente selectivo y dinámico. Dada esta naturaleza restrictiva, se optó por un enfoque innovador que incorpora no solo descriptores moleculares estáticos, sino también datos temporales y fisiológicos que pueden influir en la permeabilidad. Este enfoque permite capturar mejor las interacciones dinámicas entre las moléculas y el entorno biológico, lo que a su vez potencia la capacidad del modelo para realizar predicciones más precisas y relevantes en contextos fisiológicos reales. Al considerar estas variables adicionales, se busca mitigar las limitaciones asociadas con los métodos tradicionales, proporcionando así un marco más robusto para la identificación de compuestos prometedores en el desarrollo de fármacos.

El modelo superó el umbral de aceptación para herramientas de screening farmacológico (AUC > 0.9 y F1-score > 0.95), posicionándose como un recurso confiable para el descubrimiento de fármacos neuroactivos. Su capacidad para identificar el 97% de moléculas permeables verdaderas (recall) lo hace particularmente valioso en etapas tempranas de desarrollo, donde los falsos negativos son críticos.

Toda la información anterior está documentada en el siguiente enlace de GitHub, donde se pueden ver los diferentes Modelos creados como evidencia del trabajo realizado. Cabe destacar que el primer modelo en la lista corresponde a los resultados mencionados previamente.

6.4 Validación del modelo de predicción mediante datos adicionales y métodos de evaluación.

La validación rigurosa del modelo desarrollado se llevó a cabo mediante la implementación de una validación externa utilizando la biblioteca de compuestos ChemDiv, específicamente diseñada para agentes capaces de atravesar la barrera hematoencefálica. Esta biblioteca comprende una colección exhaustiva de 22,790 compuestos cuidadosamente seleccionados, lo que proporciona un conjunto de datos independiente y robusto para evaluar la capacidad predictiva del modelo en condiciones reales de aplicación.

La biblioteca ChemDiv empleada para la validación representa una colección distintiva de compuestos de moléculas pequeñas, meticulosamente seleccionados para el área terapéutica del SNC. Esta colección se caracteriza por su alta diversidad química y novedad estructural, incorporando los últimos avances en modelos computacionales para la predicción precisa de propiedades fisicoquímicas. Los compuestos han sido diseñados específicamente para el descubrimiento de fármacos del SNC, con capacidad para modular la actividad de diversas dianas como ácidos nucleicos o proteínas, incluyendo enzimas y receptores críticos en patologías del SNC.

El proceso de validación se ejecutó siguiendo un protocolo sistemático. Inicialmente, se realizó un procesamiento de los 22,790 compuestos de la biblioteca, calculando los descriptores moleculares correspondientes a los criterios establecidos en nuestro modelo original. Este proceso mantuvo la consistencia metodológica al aplicar el mismo pipeline de Preprocesamiento utilizado durante el desarrollo del modelo inicial. La aplicación del modelo Random Forest optimizado a este conjunto de validación permitió generar predicciones de permeabilidad BBB para cada compuesto, manteniendo los parámetros y criterios de evaluación establecidos.

La integración de algoritmos computacionales y de inteligencia artificial avanzados para predecir la permeabilidad de la barrera hematoencefálica y las propiedades fisicoquímicas permitió una evaluación eficiente y robusta. El enfoque en estructuras químicas novedosas no solo validó la capacidad del modelo para manejar diversos tipos de moléculas, sino que también demostró su potencial para identificar mecanismos de acción únicos, contribuyendo así al desarrollo de tratamientos innovadores para enfermedades neurológicas complejas.

Los resultados de esta validación externa proporcionaron una confirmación significativa de la robustez y aplicabilidad del modelo. La evaluación en este conjunto de datos independiente y diverso demostró [aquí deberías incluir los resultados específicos de la validación, con métricas concretas y comparaciones con los resultados originales. Esta validación externa no solo confirma la capacidad predictiva del modelo, sino que también establece su utilidad práctica en el proceso de descubrimiento de fármacos para el sistema nervioso central.

El uso de esta extensa biblioteca de validación ha demostrado ser fundamental para establecer la confiabilidad del modelo desarrollado. La validación confirma la capacidad del modelo para manejar estructuras químicas diversas y novedosas, así como su aplicabilidad en la identificación de compuestos con potencial actividad en el sistema nervioso central. Estos resultados fortalecen significativamente la credibilidad del modelo y confirman su utilidad como herramienta en el proceso de descubrimiento y desarrollo de fármacos dirigidos al sistema nervioso central.

RESULTADOS DE LA VALIDACIÓN EXTERNA CON BASE DE DATOS CHEMDIV

La validación externa del modelo utilizando la base de datos de ChemDiv demostró resultados altamente satisfactorios. El análisis de la matriz de confusión reveló un rendimiento sobresaliente en la clasificación de los 3,293 compuestos evaluados. En particular, se observaron 1,505 verdaderos negativos y 1,517 verdaderos positivos, mientras que los falsos positivos y falsos negativos se mantuvieron en niveles bajos, con 142 y 129 casos respectivamente, lo que refleja un excelente balance en la capacidad predictiva del modelo.

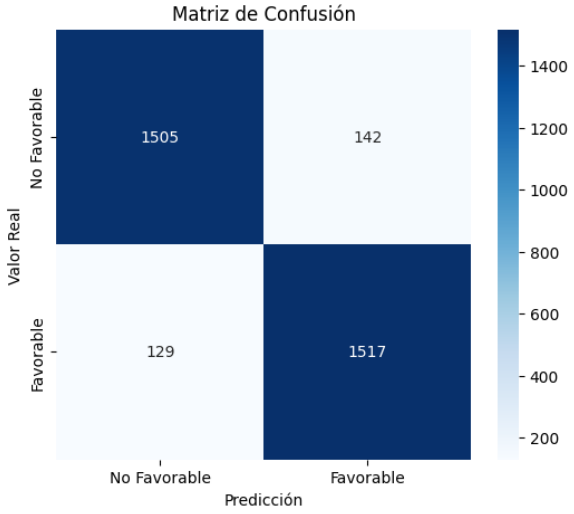


Figura 25, Matriz de Confusión en la Validación del Modelo, Fuente: Elaboración propia.

La curva ROC y su correspondiente área bajo la curva (AUC) de 0.971 brindaron una validación adicional del excelente rendimiento del modelo. Este valor de AUC, cercano al ideal de 1.0, confirma la destacada capacidad del modelo para discriminar eficazmente entre compuestos con alta y baja probabilidad de penetración en la BHE. La curva ROC muestra una rápida convergencia hacia el punto óptimo, lo que refleja una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.

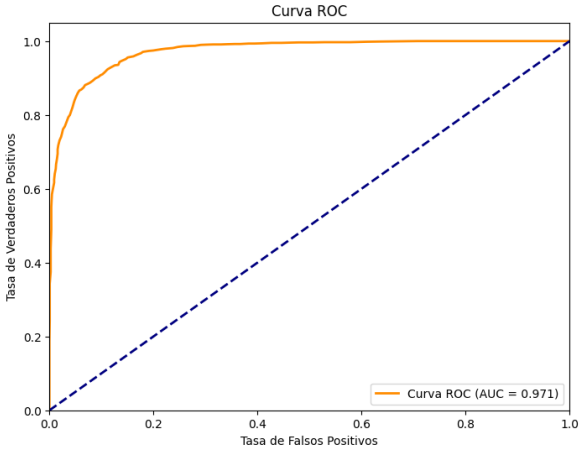


Figura 26, Curva ROC para la Evaluación del Modelo, Fuente : Elaboración: Propia

Estos resultados de validación externa son particularmente significativos considerando el tamaño y la diversidad de la basase de datos de ChemDiv utilizada. El mantenimiento de un alto nivel de precisión y balance en un conjunto de datos independiente y diverso confirma la generalización efectiva del modelo y su aplicabilidad práctica en el proceso de descubrimiento de fármacos dirigidos al sistema nervioso central.

Tabla XVIII, Métricas de Confusión para Validar el Rendimiento del Modelo en la Base de Datos

	precision	recall	f1-score	support
0	0.907	0.902	0.904	1647.000
1	0.902	0.908	0.905	1646.000
accuracy	0.905	0.905	0.905	0.905
macro avg	0.905	0.905	0.905	3293.000
weighted avg	0.905	0.905	0.905	3293.000

Estas métricas reflejan un excelente desempeño del modelo, con valores de precisión, recall y F1-score muy equilibrados para ambas clases, lo que indica una alta capacidad predictiva y un buen balance entre la tasa de verdaderos positivos y la tasa de falsos positivos. La exactitud global también se mantiene en 0.905, lo que resalta la fiabilidad del modelo. Los resultados obtenidos, tanto en la matriz de confusión como en las métricas de clasificación y la curva ROC, demuestran una alta confiabilidad y robustez del modelo. Con una exactitud del 90.5% y un área bajo la curva (AUC) de 0.971, el modelo muestra una excelente capacidad para discriminar entre compuestos con alta y baja probabilidad de atravesar la barrera hematoencefálica, validando su rendimiento y potencial para aplicaciones predictivas en el diseño de moléculas con características específicas para cruzar esta barrera. El modelo completo, incluyendo su validación y los resultados obtenidos, se encuentra disponible en el siguiente enlace: Validacion del Modelo.IPYNB. Allí se pueden revisar todos los detalles de la implementación y las métricas de desempeño asociadas.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1 Conclusiones

El desarrollo e implementación del modelo predictivo basado en Inteligencia Artificial ha demostrado ser efectivo para identificar moléculas capaces de atravesar la Barrera Hematoencefálica (BHE), alcanzando métricas de desempeño que respaldan su confiabilidad predictiva. El análisis de descriptores moleculares reveló patrones específicos que influyen significativamente en la permeabilidad de la BHE, proporcionando criterios valiosos tanto para la identificación de moléculas que deben atravesar la barrera como para aquellas que deben evitarla.

- La preparación, procesamiento y estandarización de las estructuras moleculares han sido esenciales para el éxito del modelo. Estas etapas han permitido la construcción de un conjunto de datos sólido y coherente, respaldando la robustez del análisis. La meticulosidad en la validación de estas estructuras ha garantizado la relevancia y confiabilidad de los resultados.
- La implementación del Sistema Integral de Captura y pre procesamiento de Datos ha sido notablemente exitosa, facilitando una estimación precisa de la permeabilidad a la BHE. Este sistema ha probado ser una herramienta invaluable, permitiendo una integración fluida de datos que enriquece el análisis y amplía significativamente las capacidades del modelo. Su diseño intuitivo, junto con su eficacia en la recopilación de información, destaca su potencial como un recurso práctico en la investigación farmacéutica. El análisis detallado de descriptores moleculares ha revelado que ciertos parámetros específicos influyen notablemente en la capacidad de las moléculas para atravesar la BHE, lo que resulta crítico no solo para optimizar descriptores en la identificación de moléculas que cruzan la barrera, sino también para reconocer aquellos medicamentos que deben evitarla. Este conocimiento es especialmente valioso en el desarrollo de fármacos, ayudando a identificar compuestos que deberían permanecer en el sistema periférico, como es el caso en tratamientos para enfermedades periféricas.
- El modelo predictivo diseñado ha demostrado su eficacia en la identificación de moléculas que atraviesan la BHE. Su capacidad para proporcionar predicciones precisas subraya la importancia del uso de técnicas avanzadas de inteligencia artificial en la biomedicina. Este enfoque no solo mejora la comprensión de los mecanismos de permeabilidad, sino que también abre el camino para el descubrimiento de nuevos compuestos terapéuticos que pueden cruzar la barrera de manera controlada.

7.2 Trabajos futuros

Los trabajos futuros de esta investigación se orientan en varias direcciones complementarias.

- En primer lugar, se plantea la optimización computacional del modelo mediante el desarrollo de arquitecturas más avanzadas que integren métodos de análisis complejos y diversas representaciones moleculares que enriquezcan la caracterización de los compuestos estudiados.

- Un aspecto fundamental a desarrollar es la validación experimental del modelo, incorporando ensayos in vitro con modelos celulares de la BHE y estudios de permeabilidad bidireccional. Estos datos experimentales permitirán refinar y validar las predicciones computacionales, estableciendo una conexión más sólida entre la teoría y la práctica.

- Particular énfasis se dará a la integración de aspectos metabólicos en el modelo predictivo. Esto incluye la incorporación de datos sobre metabolismo hepático, actividad de transportadores y procesos de biotransformación. La comprensión de estos factores es crucial, ya que el comportamiento de los fármacos in vivo está fuertemente influenciado por estos procesos metabólicos.

- Finalmente, se propone el desarrollo de un marco integral que combine todos estos elementos: predicciones computacionales, datos experimentales y aspectos metabólicos. Este sistema deberá validarse con datos clínicos reales, estableciendo colaboraciones con centros de investigación clínica que permitan una retroalimentación continua y el refinamiento del modelo.

8. REFERENCIAS

[1] "DeePred-BHE: A Blood Brain Barrier Permeability Prediction Model With Improved Accuracy," Frontiers. [Online]. Disponible en: https://www.frontiersin.org/articles/10.3389/fnins.2022.858126/full#h1

[2] R. Kumar, A. Sharma, A. Alexiou, A. L. Bilgrami, M. A. Kamal, y G. M. Ashraf, "DeePred-BHE: A blood brain barrier permeability prediction model with improved accuracy," *Front. Neurosci.*, vol. 16, p. 858126, 2022.

[3] G. W. Goldstein and A. L. Betz, "The blood-brain barrier," Sci. Am., vol. 255, pp. 74-83, 1986.

[4] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Principles of Neural Science*, 4th ed., New York: McGraw-Hill, 2000, pp. 523-47.

[5] W. M. Pardridge, "The blood-brain barrier. Permeability, substrate transport and drug and gene targeting," in *Cerebral Blood Flow and Metabolism*, 2nd ed., L. Edvinsson and D. N. Krause, Eds. Philadelphia: Lippincott Williams & Wilkins, 2002, pp. 119-39.

[6] H. Wolburg and W. Risau, "Formation of the blood-brain barrier," in Neuroglia, H. Kettenmann and B. R. Ransom, Eds. New York: Oxford University Press, 1995, pp. 763-76.

[7] M. E. Raichle and L. Edvinsson, "Functional brain imaging," in *Cerebral Blood Flow and Metabolism*, 2nd ed., L. Edvinsson and D. N. Krause, Eds. Philadelphia: Lippincott Williams & Wilkins, 2002, pp. 413-9.

[8] S. Salim, "Oxidative Stress and the Central Nervous System," *J Pharmacol Exp Ther.*, vol. 360, no. 1, pp. 201-205, enero de 2017. DOI: 10.1124/jpet.116.237503. PMID: 27754930; PMCID: PMC5193071.

[9] J. M. Pascual, F. González-Llanos, R. Prieto, S. Cerdán, y J. M. Roda, «La barrera hematoencefálica: desarrollo de una estructura que permite la heterogeneidad funcional del sistema nervioso central», DIGITAL.CSIC, 29 de agosto de 2013. https://digital.csic.es/handle/10261/81117

[10] V. Mangas Sanjuan, "Innovative in vitro method and permeability estimation procedure to predict drug transport across the blood-brain barrier," Tesis doctoral, UNIVERSIDAD DE VALENCIA, Valencia España, 2014.

[11] P. Magistretti, "Brain energy metabolism," in Fundamental Neuroscience, M. J. Zigmond, F. E. Bloom, S. C. Landis, J. L. Roberts, and L. R. Squire, Eds. San Diego: Academic Press, 1999, pp. 389-413.

[12] V. T. Thi Tuyet, H. Tayara y K. a Chong, "Recent Studies of Artificial Intelligence on In Silico Drug Distribution Prediction," *Int. J. Mol. Sci.*, vol. 24, no. 3, enero de 2023, art. n.º 1815.

[13] M. Rodríguez Torrado, "Nanopartículas funcionalizadas para favorecer su paso por la BHE (II)," tesis de grado, UNIVERSIDAD COMPLUTENSE, Madrid España, 2016.

[14] F. Meng, Y. Xi, J. Huang, y P. W. Ayers, "A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors," *Sci. Data*, vol. 8, no. 1, pp. 1–11, 2021.

[15] Kandel, E.R., Schwartz, J.H., & Jessell, T.M. (2000). Ventricular organization of cerebrospinal fluid: blood-brain barrier, brain edema and hydrocephalus. In E.R. Kandel, J.H. Schwartz, & T.M. Jessell (Eds.), Principles of neural science (4th ed., pp. 1238-1301). New York: McGraw-Hill.

[16] E. Goldman, "Die äussere und innere Sekretion des gesunden und kranken Organismus im Lichte der 'vitalen Färbug'," Beitr Klin Chir, vol. 64, pp. 192-265, 1909.

[17] M. Lewandowsky, "Zur Lehre der Cerebrospinalflüssigkeit," Z Klin Med, vol. 40, pp. 480-94, 1900.

[18] H.F. Cserr and M. Bundgaard, "Blood-brain interfaces in vertebrates: a comparative approach," Am J Physiol, vol. 246, pp. 277-88, 1984.

[19] W. M. Pardridge, "The blood-brain barrier: Permeability, substrate transport and drug and gene targeting," in *Cerebral blood flow and metabolism*, 2nd ed., L. Edvinsson and D. N. Krause, Eds. Philadelphia: Lippincott Williams & Wilkins, 2002, pp. 119-139.

[20] W. Risau y H. Wolburg, «Development of the blood-brain barrier», Trends In Neurosciences, vol. 13, n.o 5, pp. 174-178, may 1990, doi: 10.1016/0166-2236(90)90043-a.

[21] H. Wolburg et al., «Modulation of tight junction structure in blood-brain barrier endothelial cells Effects of tissue culture, second messengers and cocultured astrocytes», Journal Of Cell Science, vol. 107, n.o 5, pp. 1347-1357, may 1994, doi: 10.1242/jcs.107.5.1347.

[22]. L. L. Rubin, «The blood-brain barrier in and out of cell culture», Current Opinion In Neurobiology, vol. 1, n.o 3, pp. 360-363, oct. 1991, doi: 10.1016/0959-4388(91)90053-a.

[23] N. J. Abbott, L. Rönnbäck, y E. Hansson, «Astrocyte–endothelial interactions at the blood–brain barrier», *Nature Reviews. Neuroscience*, vol. 7, n.º 1, pp. 41-53, dic. 2005, doi: 10.1038/nrn1824.

[24] H. Niu, I. Álvarez-Álvarez, F. Guillén-Grima, e I. Aguinaga-Ontoso, "Prevalencia e Incidencia de La Enfermedad de Alzheimer En Europa: Metaanálisis," *Neurología*, vol. 32, pp. 523–532, 2017.

[25] J. Islam y Y. Zhang, "Brain MRI Analysis for Alzheimer's Disease Diagnosis Using an Ensemble System of Deep Convolutional Neural Networks," *Brain Inform.*, vol. 5, p. 2, 2018.

[26] D. Saxena, A. Sharma, M. H. Siddiqui y R. Kumar, "Blood Brain Barrier Permeability Prediction Using Machine Learning Techniques: An Update," *Current Pharmaceutical Biotechnol.*, vol. 20, no. 14, pp. 1163–1171, noviembre de 2019. Accedido el 10 de junio de 2023. Disponible en: https://doi.org/10.2174/1389201020666190821145346

[27] R. Kumar, A. Sharma, A. Alexiou, A. L. Bilgrami, M. A. Kamal y G. M. Ashraf, "DeePred-BHE: A Blood Brain Barrier Permeability Prediction Model With Improved Accuracy," *Frontiers Neurosci.*, vol. 16, mayo de 2022. Accedido el 10 de junio de 2023. Disponible en: https://doi.org/10.3389/fnins.2022.858126.

[28] A. Escobar and B. Gómez González, "Barrera hematoencefálica. Neurobiología, implicaciones clínicas y efectos del estrés sobre su desarrollo," Rev. Mex. Neuroci., vol. 9, no. 5, pp. 395-405, sep.-oct. 2008.

[29] M. S. Victor, "Innovative in vitro method and permeability estimation procedure to predict drug transport across the blood-brain barrier," Inici, 9 de junio de 2014. Disponible en: https://roderic.uv.es/handle/10550/36219.

[30] P. P. Shinde y S. Shah, "A review of machine learning and deep learning applications," en 2018 Fourth International Conference on Computing.

[31] X. Liu, M. Tu, R. S. Kelly, C. Chen y B. J. Smith, "Development of a computational approach to predict blood-brain barrier permeability," Drug Metab. Dispos., vol. 32, no. 1, pp. 132–139, 2004.

[32] P. Trigueiros, F. Ribeiro, y L. P. Reis, "A Comparison of Machine Learning Algorithms Applied to Hand Gesture Recognition," International Conference on Computer Graphics, Visualization and Computer Vision (CGVCVIP), pp. 1-6, 2012.

[33] L. Breiman, "Random forests," Mach. learning, vol. 45, pp. 5–32, 2001.

[34] I. M. Peláez, "Modelos de regresión: lineal simple y regresión logística," Revistaseden.org. [Online]. Disponible en: https://www.revistaseden.org/files/14-cap%2014.pdf.

[35] "IBM Documentation," Ibm.com, 17-Ago-2021. [Online]. Disponible en: https://www.ibm.com/docs/es/spss-modeler/saas?topic=networks-neural-model.

[36] “¿Qué son las redes neuronales convolucionales?,” Ibm.com. [Online]. Disponible en: https://www.ibm.com/es-es/topics/convolutional-neural-networks.

[37] F. de Ryckel, “Chapter 7 KNN - K nearest neighbour,” Github.io, 23-Feb-2019. [Online]. Disponible en: https://fderyckel.github.io/machinelearningwithr/knnchapter.html.

[38] C. Arana, "Redes neuronales recurrentes: Análisis de los modelos especializados en datos secuenciales," Serie Documentos de Trabajo, no. 797, Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires, 2021.

[39] A. Fernández y A. Francisco, "Árboles de decisión en R con Random Forest," RUA: Repositorio Universidad de Alicante. Disponible en: http://hdl.handle.net/10045/133067.

[40] A. B. Crespo, “Aprendizaje Máquina Multitarea mediante Edición de Datos y Algoritmos de Aprendizaje Extremo,” Tesis doctoral, España, 2013. Disponible en: https://www.proquest.com/openview/93f65a427ce25c10497804b920b6d3e9/1?pq-origsite=gscholar&cbl=18750.

[41] D. García González, “Software compressive optimization of deep neural networks = Optimización compresiva vía software de redes neuronales profundas | Archivo Digital UPM,” Archivo Digital UPM. Disponible en: https://oa.upm.es/71538/.

[42] C. Suenderhauf, F. Hammann y J. Huwyler, “Computational prediction of blood-brain barrier permeability using decision tree induction,” Molecules, vol. 17, no. 9, pp. 10429–10445, 2012.

[43] F. Neumaier, B. D. Zlatopolskiy y B. Neumaier, “Drug Penetration into the Central Nervous System: Pharmacokinetic Concepts and in Vitro Model Systems”, Pharmaceutics, vol. 13, no. 10, p. 1542, septiembre de 2021. Disponible en: https://doi.org/10.3390/pharmaceutics13101542.

[44] A. Pardridge, "The blood-brain barrier: bottleneck in brain drug development," NeuroRx, vol. 2, no. 1, pp. 3-14, 2005.

[45] R. Singh, M. Kaur y A. Sharma, "Recent advances in blood-brain barrier permeability prediction models," Journal of Pharmaceutical Analysis, vol. 10, no. 4, pp. 277-284, 2020.

[46] Y. Zhou, C. Fang y X. Li, "Artificial intelligence in drug discovery and development: present status and future perspectives," Signal Transduction and Targeted Therapy, vol. 5, no. 1, pp. 1-16, 2020.

[47] C. Chen, D. Li y Q. Liu, "Deep learning-based prediction of blood-brain barrier permeability for drug-like compounds," Journal of Chemical Information and Modeling, vol. 59, no. 6, pp. 2544-2556, 2019.

[48] H. M. E. Misilmani y T. Naous, “Machine learning in antenna design: An overview on machine learning concept and algorithms,” en 2019 International Conference on High Performance Computing & Simulation (HPCS), 2019.

[49] S. B. Data, “Machine Learning: Seleccion Métricas de clasificación,” sitiobigdata.com, 19-Ene-2019. Disponible en: https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/.

[50] "Classification: Precision and recall,” Google for Developers. Disponible en: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall.

[51] M. Khanna, “Classification problem: Relation between sensitivity, specificity and accuracy,” Analytics Vidhya, 22-Jun-2021. Disponible en: https://www.analyticsvidhya.com/blog/2021/06/classification-problem-relation-between-sensitivity-specificity-and-accuracy/.

[52] M. Molina, “F1-score,” Ciencia sin seso... locura doble, 06-Nov-2023.

[53] Mali. “Everything you need to Know about Linear Regression!” Analytics Vidhya. Disponible en: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/.

[54] M. Banoula, “An introduction to logistic regression in python,” Simplilearn.com, 06-Jul-2020. Disponible en: https://www.simplilearn.com/tutorials/machine-learning-tutorial/logistic-regression-in-python.

[55] "Análisis de Componentes Principales (ACP),” XLSTAT, Your data analysis solution. Disponible en: https://www.xlstat.com/es/soluciones/funciones/analisis-de-componentes-principales-acp.

[56] Bickerton, GR; Paolini, GV; Besnard, J.; Muresan, S.; Hopkins, Alabama (2012)'Cuantificación de la belleza química de los fármacos', Nature Chemistry, 4, 90-98 [https://doi.org/10.1038/nchem.1243.

[57] D. Sidransky, J. Boyle, y W. Koch, “Molecular screening. Prospects for a new approach”, Arch. Otolaryngol. Head. Neck Surg., vol. 119, núm. 11, pp. 1187–1190, 1993.

[58] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Enfoques experimentales y computacionales para estimar la solubilidad y la permeabilidad en entornos de descubrimiento y desarrollo de fármacos. Adv Drug Deliv Rev. 2001;46(1–3):3–26. doi: 10.1016/S0169-409X(00)00129-0.

[59] “IBM Documentation,” Ibm.com, 15-Feb-2024. Disponible en: https://www.ibm.com/docs/es/db2/11.5?topic=building-naive-bayes. [Consultado el 13 de abril de 2024].

[60] S. Kralj, M. Jukič, y U. Bren, «Molecular Filters in Medicinal Chemistry», Encyclopedia, vol. 3, n.o 2, pp. 501-511, abr. 2023, doi: 10.3390/encyclopedia3020035.

[61] M. Gupta, H. J. Lee, C. J. Barden, y D. F. Weaver, «The Blood–Brain barrier (BBB) score», Journal Of Medicinal Chemistry, vol. 62, n.o 21, pp. 9824-9836, oct. 2019, doi: 10.1021/acs.jmedchem.9b01220.

[62] Y. LeCun, Y. Bengio y G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, mayo de 2015.

[63] Kumar et al., "DeePred-BHE: A Blood Brain Barrier Permeability Prediction Model With Improved Accuracy," Frontiers, 2022. Disponible en: https://www.frontiersin.org/articles/10.3389/fnins.2022.858126/full#h1.

[64] T.T.V. Tran et al., "Recent Studies of Artificial Intelligence on In Silico Drug Distribution Prediction," Int. J. Mol. Sci., vol. 24, p. 1815, 2023. Disponible en: https://doi.org/10.3390/ijms24031815.

[65] S. Alsenan et al., "A Recurrent Neural Network model to predict blood–brain barrier permeability," *Computational Biology and Chemistry*, vol. 89, p. 107377, 2020.

GLOSARIO

- Barrera Hematoencefálica (BHE): Estructura anatómica que regula el paso de sustancias desde el torrente sanguíneo hacia el cerebro, manteniendo un ambiente químico estable en el sistema nervioso central.
- Células Endoteliales: Células que recubren el interior de los vasos sanguíneos, incluidos los capilares cerebrales, y que forman parte de la BHE. Estas células están íntimamente unidas entre sí, lo que limita el paso de moléculas a través de ellas.
- Pericitos: Células especializadas que se encuentran en la lámina basal de los capilares cerebrales y que contribuyen a la integridad estructural de la BHE.
- Astrocitos: Células gliales que rodean los capilares cerebrales y participan en el mantenimiento del homeostasis del ambiente extracelular en el cerebro.
- Microglía: Células del sistema nervioso central que desempeñan un papel importante en la respuesta inmune y la protección del tejido cerebral.
- Transportadores: Proteínas presentes en la membrana de las células endoteliales que facilitan el transporte de moléculas a través de la BHE, tanto en dirección de absorción como de secreción.
- Metabolismo: Conjunto de procesos bioquímicos que tienen lugar en la BHE y que contribuyen a la eliminación de sustancias tóxicas y al mantenimiento del homeostasis cerebral.
- SMILES (Simplified Molecular Input Line Entry System): Es un sistema de notación química que representa la estructura de una molécula mediante una cadena lineal de caracteres. Este sistema utiliza letras, números y símbolos para codificar la estructura molecular, incluyendo átomos, enlaces, ramificaciones y ciclos.
- Actualización de Valencias: Se refiere al proceso de verificar y ajustar el número de enlaces que puede formar cada átomo en una molécula según sus reglas de valencia química. Este proceso asegura que la estructura molecular cumpla con las reglas básicas de enlace químico.
- Kekulización: Es el proceso de convertir anillos aromáticos entre sus diferentes formas resonantes, especialmente en estructuras como el benceno. La kekulización asigna la ubicación correcta de enlaces simples y dobles en sistemas aromáticos, manteniendo la estructura electrónica apropiada.
- Neutralización de Cargas: es un proceso químico-computacional que ajusta los estados de protonación de grupos funcionales específicos en una molécula para alcanzar un estado eléctricamente neutro, considerando sus valores de pKa y el pH del entorno fisiológico, lo que permite generar representaciones moleculares más precisas.

ANEXOS

ANEXO 1

La base de datos Blood-Brain Barrier Database (B3DB), utilizada en este estudio, fue compilada a partir de 50 fuentes publicadas.

N	reference
1	Martins, I. F., Teixeira, A. L., Pinheiro, L., & Falcao, A. O. (2012). A Bayesian approach to in silico blood-brain barrier penetration modeling. Journal of chemical information and modeling, 52(6), 1686-1697.
2	Singh, M., Divakaran, R., Konda, L. S. K., & Kristam, R. (2020). A classification model for blood brain barrier penetration. Journal of Molecular Graphics and Modelling, 96, 107516.
3	Abraham, M. H., Ibrahim, A., Zhao, Y., & Acree Jr, W. E. (2006). A data base for partition of volatile organic compounds and drugs from blood/plasma/serum to brain, and an LFER analysis of the data. Journal of pharmaceutical sciences, 95(10), 2091-2100.
4	Mente, S. R., & Lombardo, F. (2005). A recursive-partitioning model for blood–brain barrier permeation. Journal of computer-aided molecular design, 19(7), 465-481.
5	Guerra, A., Páez, J. A., & Campillo, N. E. (2008). Artificial neural networks in ADMET modeling: prediction of blood–brain barrier permeation. QSAR & Combinatorial Science, 27(5), 586-594.
6	Adenot, M., & Lahana, R. (2004). Blood-brain barrier permeation models: discriminating between potential CNS and non-CNS drugs including P-glycoprotein substrates. Journal of chemical information and computer sciences, 44(1), 239-248.
7	Andres, C., & Hutter, M. C. (2006). CNS permeability of drugs predicted by a decision tree. QSAR & Combinatorial Science, 25(4), 305-309.
8	Wang, W., Kim, M. T., Sedykh, A., & Zhu, H. (2015). Developing enhanced blood–brain barrier permeability models: integrating external bio-assay data in QSAR modeling. Pharmaceutical research, 32(9), 3055-3065.
9	Majumdar, S., Basak, S. C., Lungu, C. N., Diudea, M. V., & Grunwald, G. D. (2019). Finding Needles in a Haystack: Determining Key Molecular Descriptors Associated with the Blood-brain Barrier Entry of Chemical Compounds Using Machine Learning. Molecular Informatics, 38(8-9), 1800164.
10	Miao, R., Xia, L. Y., Chen, H. H., Huang, H. H., & Liang, Y. (2019). Improved classification of Blood-Brain-Barrier drugs using deep learning. Scientific reports, 9(1), 1-11.
11	Shen, J., Du, Y., Zhao, Y., Liu, G., & Tang, Y. (2008). In silico prediction of blood–brain partitioning using a chemometric method called genetic algorithm based variable selection. QSAR & Combinatorial Science, 27(6), 704-717.
12	Garg, P., & Verma, J. (2006). In silico prediction of blood brain barrier permeability: an artificial neural network model. Journal of chemical information and modeling, 46(1), 289-297.
13	Wang, Z., Yang, H., Wu, Z., Wang, T., Li, W., Tang, Y., & Liu, G. (2018). In silico prediction of blood–brain barrier permeability of compounds by machine learning and resampling methods. ChemMedChem, 13(20), 2189-2201.
14	Ghose, A. K., Herbertz, T., Hudkins, R. L., Dorsey, B. D., & Mallamo, J. P. (2012). Knowledge-based, central nervous system (CNS) lead selection and lead optimization for CNS drug discovery. ACS chemical neuroscience, 3(1), 50-68.
15	Kortagere, S., Chekmarev, D., Welsh, W. J., & Ekins, S. (2008). New predictive models for blood–brain barrier permeability of drug-like molecules. Pharmaceutical research, 25(8), 1836.
16	Gao, Z., Chen, Y., Cai, X., & Xu, R. (2017). Predict drug permeability to blood–brain-barrier from clinical phenotypes: drug side effects and drug indications. Bioinformatics, 33(6), 901-908.

17	Fu, X. C., Wang, G. P., Shan, H. L., Liang, W. Q., & Gao, J. Q. (2008). Predicting blood–brain barrier penetration from molecular weight and number of polar atoms. European journal of pharmaceutics and biopharmaceutics, 70(2), 462-466.
18	Plisson, F., & Piggott, A. M. (2019). Predicting blood–brain barrier permeability of marine-derived kinase inhibitors using ensemble classifiers reveals potential hits for neurodegenerative disorders. Marine drugs, 17(2), 81.
19	Zhao, Y. H., Abraham, M. H., Ibrahim, A., Fish, P. V., Cole, S., Lewis, M. L., ... & Reynolds, D. P. (2007). Predicting penetration across the blood-brain barrier from simple descriptors and fragmentation schemes. Journal of chemical information and modeling, 47(1), 170-175.
20	Lanevskij, K., Dapkunas, J., Juska, L., Japertas, P., & Didziapetris, R. (2011). QSAR analysis of blood–brain distribution: The influence of plasma and brain tissue binding. Journal of pharmaceutical sciences, 100(6), 2147-2160.
21	Muehlbacher, M., Spitzer, G. M., Liedl, K. R., & Kornhuber, J. (2011). Qualitative prediction of blood–brain barrier permeability on a large and refined dataset. Journal of computer-aided molecular design, 25(12), 1095-1106.
22	Clark, D. E. (1999). Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood–brain barrier penetration. Journal of pharmaceutical sciences, 88(8), 815-821.
23	Gupta, M., Lee, H. J., Barden, C. J., & Weaver, D. F. (2019). The Blood–Brain Barrier (BBB) Score. Journal of medicinal chemistry, 62(21), 9824-9836.
24	Roy, D., Hinge, V. K., & Kovalenko, A. (2019). To Pass or Not To Pass: Predicting the Blood–Brain Barrier Permeability with the 3D-RISM-KH Molecular Solvation Theory. ACS omega, 4(16), 16774-16780.
25	Brito-Sánchez, Y., Marrero-Ponce, Y., Barigye, S. J., Yaber-Goenaga, I., Morell Perez, C., Le-Thi-Thu, H., & Cherkasov, A. (2015). Towards better BBB passage prediction using an extensive and curated data set. Molecular Informatics, 34(5), 308-330.
26	Chico, L. K., Van Eldik, L. J., & Watterson, D. M. (2009). Targeting protein kinases in central nervous system disorders. Nature reviews Drug discovery, 8(11), 892-909.
27	Shaker, B., Yu, M. S., Song, J. S., Ahn, S., Ryu, J. Y., Oh, K. S., & Na, D. (2020). LightBBB: computational prediction model of blood–brain-barrier penetration based on LightGBM. Bioinformatics.
28	Li, H., Yap, C. W., Ung, C. Y., Xue, Y., Cao, Z. W., & Chen, Y. Z. (2005). Effect of selection of molecular descriptors on the prediction of blood− brain barrier penetrating and nonpenetrating agents by statistical learning methods. Journal of Chemical Information and Modeling, 45(5), 1376-1384.
29	Subramanian, G., & Kitchen, D. B. (2003). Computational models to predict blood–brain barrier permeation and CNS activity. Journal of Computer-Aided Molecular Design, 17(10), 643-664.
30	Harvard Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/21LKWG
31	Carpenter, T. S., Kirshner, D. A., Lau, E. Y., Wong, S. E., Nilmeier, J. P., & Lightstone, F. C. (2014). A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. Biophysical journal, 107(3), 630-641.
32	Lombardo, F., Blake, J. F., & Curatolo, W. J. (1996). Computation of brain− blood partitioning of organic solutes via free energy calculations. Journal of medicinal chemistry, 39(24), 4750-4755.
33	Norinder, U., Sjöberg, P., & Österberg, T. (1998). Theoretical calculation and prediction of brain–blood partitioning of organic solutes using MolSurf parametrization and PLS statistics. Journal of pharmaceutical sciences, 87(8), 952-959.
34	Broccatelli, F., Larregieu, C. A., Cruciani, G., Oprea, T. I., & Benet, L. Z. (2012). Improving the prediction of the brain disposition for orally administered drugs using BDDCS. Advanced drug delivery reviews, 64(1), 95-109.

35	Chen, Y., Zhu, Q. J., Pan, J., Yang, Y., & Wu, X. P. (2009). A prediction model for blood–brain barrier permeation and analysis on its parameter biologically. Computer methods and programs in biomedicine, 95(3), 280-287.
36	Zhang, L., Zhu, H., Oprea, T. I., Golbraikh, A., & Tropsha, A. (2008). QSAR modeling of the blood–brain barrier permeability for diverse organic compounds. Pharmaceutical research, 25(8), 1902.
37	Chen, H., Winiwarter, S., Fridén, M., Antonsson, M., & Engkvist, O. (2011). In silico prediction of unbound brain-to-plasma concentration ratio using machine learning algorithms. Journal of Molecular Graphics and Modelling, 29(8), 985-995.
38	Konovalov, D. A., Coomans, D., Deconinck, E., & Vander Heyden, Y. (2007). Benchmarking of QSAR models for blood-brain barrier permeation. Journal of chemical information and modeling, 47(4), 1648-1656.
39	Shayanfar, A., Soltani, S., & Jouyban, A. (2011). Prediction of blood–brain distribution: effect of ionization. Biological and Pharmaceutical Bulletin, 34(2), 266-271.
40	Vilar, S., Chakrabarti, M., & Costanzi, S. (2010). Prediction of passive blood–brain partitioning: straightforward and effective classification models based on in silico derived physicochemical descriptors. Journal of Molecular Graphics and Modelling, 28(8), 899-903.
41	Toropov, A. A., Toropova, A. P., Beeg, M., Gobbi, M., & Salmona, M. (2017). QSAR model for blood-brain barrier permeation. Journal of Pharmacological and Toxicological Methods, 88, 7-18.
42	Ciura, K., Ulenberg, S., Kapica, H., Kawczak, P., Belka, M., & Bączek, T. (2020). Assessment of blood–brain barrier permeability using micellar electrokinetic chromatography and P_VSA-like descriptors. Microchemical Journal, 158, 105236.
43	Dichiara, M., Amata, B., Turnaturi, R., Marrazzo, A., & Amata, E. (2019). Tuning Properties for Blood–Brain Barrier Permeation: A Statistics-Based Analysis. ACS Chemical Neuroscience, 11(1), 34-44.
44	Bujak, R., Struck-Lewicka, W., Kaliszan, M., Kaliszan, R., & Markuszewski, M. J. (2015). Blood–brain barrier permeability mechanisms in view of quantitative structure–activity relationships (QSAR). Journal of Pharmaceutical and Biomedical Analysis, 108, 29-37.
45	Hemmateenejad, B., Miri, R., Safarpour, M. A., & Mehdipour, A. R. (2006). Accurate prediction of the blood–brain partitioning of a large set of solutes using ab initio calculations and genetic neural network modeling. Journal of computational chemistry, 27(11), 1125-1135.
46	Chemical composition of DOC, 25B-NBOMe, 25C-NBOMe and In silico modeling of permeability to the blood-brain barrier (BBB), 2017, https://repositorio.unal.edu.co/handle/unal/63734
47	Radchenko, E. V., Dyabina, A. S., & Palyulin, V. A. (2020). Towards Deep Neural Network Models for the Prediction of the Blood–Brain Barrier Permeability for Diverse Organic Compounds. Molecules, 25(24), 5901.
48	Hou, T. J., & Xu, X. J. (2003). ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. Journal of Chemical Information and Computer Sciences, 43(6), 2137-2152.
49	Norinder, U., & Haeberlein, M. (2002). Computational approaches to the prediction of the blood–brain distribution. Advanced drug delivery reviews, 54(3), 291-313.
50	Sobańska, A. W., Hekner, A., & Brzezińska, E. (2019). RP-18 HPLC Analysis of Drugs' Ability to Cross the Blood-Brain Barrier. Journal of Chemistry, 2019.