ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA EN SISTEMAS

EVALUACIÓN DEL DESEMPEÑO COMPUTACIONAL DE SISTEMAS DE RECOMENDACIÓN APLICADO A BASES DE DATOS FARMACOLÓGICAS

SISTEMA DE RECOMENDACIÓN BASADO EN EL CONOCIMIENTO

TRABAJO DE INTEGRACIÓN CURRICULAR PRESENTADO COMO REQUISITO PARA LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN CIENCIAS DE LA COMPUTACION

RAMSÉS ALEXANDER PAREDES ARAUJO

ramses.paredes@epn.edu.ec

DIRECTOR: IVÁN MARCELO CARRERA IZURIETA

ivan.carrera@epn.edu.ec

Quito, Julio 2022

CERTIFICACIONES

Yo, Ramsés Paredes declaro que el trabajo de integración curricular aquí descrito es de mi autoría; no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

Ramses Alexander Paredes Araujo

Certifico que el presente trabajo de integración curricular fue desarrollado por Ramsés Paredes, bajo mi supervisión.

Ing. Iván Marcelo Carrera
DIRECTOR

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmamos que el trabajo de integración curricular aquí descrito, así como el (los) producto(s) resultante(s) del mismo, son públicos y estarán a disposición de la comunidad a través del repositorio institucional de la Escuela Politécnica Nacional; sin embargo, la titularidad de los derechos patrimoniales nos corresponde a los autores que hemos contribuido en el desarrollo del presente trabajo; observando para el efecto las disposiciones establecidas por el órgano competente en propiedad intelectual, la normativa interna y demás normas.

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

RESUMEN

El reposicionamiento de fármacos nace como una alternativa ante los grandes problemas existentes en el desarrollo de fármacos, permitiendo cubrir enfermedades para las que no existe un tratamiento específico o suplir medicamentos que se encuentran en escasez.

Se propone el desarrollo de un sistema de recomendación basado en el conocimiento para su uso en el reposicionamiento de fármacos. Se utilizó la base de datos farmacológica ChEMBL en su versión 29, que contiene información de actividades existentes entre compuestos y células. También se usó información de las interrelaciones entre células, por lo que se consideró un valor de intersección el cual indica que células se relacionan más entre sí. Con la implementación de esta información el sistema de recomendación genera predicciones de actividades para relaciones entre compuestos y células.

Se realizaron pruebas que consistían en encontrar una actividad para relaciones entre compuestos y células de la base de datos ChEMBL v30 que no se encontraban en la v29, las cuales muestran un 100% de resultados verdaderos positivos, junto con un 99.33% de resultados verdaderos negativos, estos dentro de una matriz de confusión con 305 recomendaciones.

Los resultados muestran que el sistema de recomendación basado en el conocimiento que se generó en el trabajo es capaz de generar recomendaciones de actividades para relaciones entre compuestos y células que aún no se encuentran en las bases de datos farmacológicas utilizadas.

PALABRAS CLAVE: reposicionamiento de fármacos, sistema de recomendación, sistema de recomendación basado en el conocimiento, bases de datos farmacológicas.

ABSTRACT

The drug repositioning was born as an alternative to the great problems that exist in the drug development process, making it possible to cover diseases that does not have a specific treatment or that are in short supply

We propose the development of a knowledge-based recommendation system to be used in drug repositioning. The ChEMBL pharmacological database version 29 was used, it contains information about activities between compounds and cells. Information about interrelationships between cells was also used, for which an intersection value was considered, it indicates which cells are more related to each other. Using this information, the recommendation system generates activity predictions for relationships between compounds and cells.

Tests were performed to find an activity for relationships between compounds and cells in the ChEMBL v30 database that were not found in v29 database, showing 100% true positive results, along with 99.33% true negative results. , these were obtained on a confusion matrix with 305 recommendations.

Based on the results, it was confirmed that the knowledge-based recommendation system generated can generate activity recommendations for relationships between compounds and cells that are not yet found in the pharmacological databases used.

KEYWORDS: drug repositioning, recommendation system, knowledge-based recommendation system, pharmacological databases.

1. DESCRIPCIÓN DEL COMPONENTE DESARROLLADO

Con un presupuesto de 2.500 millones de dólares y alrededor de 10 a 13 años de estudios en miles de personas por alrededor de 6 años se da el largo proceso para desarrollar un nuevo fármaco [1]. Teniendo en cuenta lo costoso que es el desarrollo de un fármaco, existen casos en los que las compañías no toman en cuenta ciertas enfermedades, ya que su lucro económico será menor. Esto sucede con las enfermedades poco comunes que afectan al 8% de la población mundial, para las cuales realizar un estudio es complicado debido a la falta de pacientes [2]. Otro gran problema con lo costoso y tardío que es el desarrollo de un fármaco surge en emergencias sanitarias en donde es necesario el inmediato desarrollo de un fármaco, un ejemplo de esto es la pandemia del COVID-19, que en China desde el 31 de diciembre del 2019 hasta el 12 de marzo de 2020 presentó alrededor de 125 048 casos junto con 4 614 muertes y para la cual no existían vacunas o medicamentos aprobados [3].

Con la intención de solventar problemas como los mencionados surge el proceso conocido como el reposicionamiento de fármacos, el cual consiste en utilizar fármacos ya existentes que pueden estar aprobados o no a causa de su efectividad o efectos secundarios, pero con el objetivo de combatir una enfermedad diferente para la que fueron desarrollados.

1.1 Objetivo general

Desarrollo de un sistema de recomendación basado en el conocimiento de bases de datos farmacológicas para generar predicciones de actividades entre compuestos y células.

1.2 Objetivos específicos

1. Recolección y limpieza de información obtenida de bases de datos farmacológicas sobre relaciones entre compuestos y células
2. Análisis de la información obtenida para el diseño y generación del sistema de recomendación basado en el conocimiento
3. Generación de predicciones mediante el sistema de recomendación y evaluación de los resultados obtenidos

1.3 Alcance

Este trabajo plantea el desarrollo de un sistema de recomendación basado en el conocimiento de bases de datos farmacológicas con el fin de generar recomendaciones de actividades entre compuestos y células. Lo fundamental para este sistema de recomendación es el uso de información de actividades existentes entre compuestos y células, así como información de relaciones entre células. Esta información permitirá la generación de recomendaciones de actividades para células y compuestos que sé no se encuentran en la base de datos farmacológica.

Una vez se cuente con el sistema de recomendación, se llevará a cabo una fase de evaluación y pruebas que permitirá evaluar las recomendaciones obtenidas. La evaluación se la llevará a cabo con una versión actualizada de la base de datos farmacológica.

Después de realizar la evaluación se realizará un análisis de resultados, el cual permitirá conocer si el sistema de recomendación cumplió su objetivo.

1.4 Marco teórico

1.4.1 Conceptos teóricos

1.4.1.1 Reposicionamiento de fármacos

El reposicionamiento de fármacos nace como una solución ante los problemas de inversión en tiempo y capital existentes en el desarrollo de fármacos. La principal idea es hacer uso de fármacos que son actualmente comercializados, que pueden tener dificultades en ensayos clínicos debido a problemas de seguridad o fármacos que fueron cancelados, pero con nuevas indicaciones para las cuales no fueron prescritos [4]. Este proceso es posible ejecutarlo gracias a la polifarmacología de las moléculas, la cual consiste en el uso de fármacos que actúan en más de una enfermedad [5]. El ejemplo mas conocido de reposicionamiento de fármacos es el Sidenafil, que comenzó como un medicamento para combatir a la enfermedad de angina de pecho, pero actualmente es utilizado en el tratamiento de la disfunción eréctil [12].

1.4.1.2 Desarrollo de fármacos

En el desarrollo de un nuevo fármaco en primer lugar se lleva a cabo una investigación profunda sobre la enfermedad que se busca tratar, de manera que se logre identificar una diana molecular relacionada con esa enfermedad, para después determinar aquellos compuestos que sean activos con esta diana, esta determinación se realiza en muchos casos de manera computacional.

Aquellos compuestos que sean activos se los toma en cuenta para realizar ensayos in vitro en líneas celulares, en animales y para finalizar en humanos. Al pasar por todas estas etapas, estos compuestos pasan a ser aprobados por un agente regulatorio [7].
Existen casos en los que un compuesto presenta actividad en más de una diana molecular, por lo que podría interferir en más de un proceso fisiológico, esto es conocido como la polifarmacología.

1.4.1.3 ChEMBL

La base de datos farmacológica de acceso abierto “ChEMBL” busca almacenar datos sobre química médica relacionada sobre moléculas y su actividad biológica, que se han obtenido por medio de investigaciones de varias revistas de química médica, además de información sobre fármacos como sus indicaciones terapéuticas y mecanismos de acción.[6]

Actualmente, la base de datos se encuentra en su versión 30, con alrededor de 2.2 millones de registros sobre compuestos, actualizada en febrero del 2022. En este trabajo se utilizará su versión 29 que fue actualizada en junio del 2021.

1.4.1.4 Sistemas de recomendación

El objetivo de estos sistemas es el de generar recomendaciones de ítems para un usuario, esto lo logran con la ayuda de grandes cantidades de información relevante para el usuario como pueden ser sus intereses, historial de búsqueda o historial de compras [9]. Existen diferentes sistemas de recomendación, que se diferencian por la información que emplean para generar recomendaciones, ya sea información brindada por el usuario o información descriptiva de lo que se quiere recomendar; algunos de los sistemas de recomendación más conocidos son: filtrado colaborativo, basado en el contenido, basado en el conocimiento y sistemas híbridos.

1.4.1.5 Modelos de filtrado colaborativo

Los modelos de filtrado colaborativo crean recomendaciones con ayuda de clasificaciones realizadas por usuarios, en donde los usuarios especifican lo que les gusta y lo que no les gusta [9]. Con la ayuda de estos datos se generan recomendaciones que podrían gustarle a un usuario en función de las reacciones de usuarios similares.

1.4.1.6 Sistemas de recomendación basados en contenido

Estos sistemas de recomendación usan las descripciones de los elementos para generar recomendaciones [9]. En este sistema la información más relevante son las calificaciones de los elementos, su comportamiento de compra, además de la información propia de los elementos.

1.4.1.7 Sistemas de recomendación basados en el conocimiento

Estos sistemas de recomendación buscan generar recomendaciones basadas en el tipo de contenido del artículo a recomendar; estos sistemas de recomendación se emplean cuando un artículo no cuenta con la suficiente información para ser recomendado [9].

1.4.1.8 Sistemas de recomendación híbridos

Los sistemas de recomendación híbridos son usados cuando se cuenta con varias fuentes de información, por lo que se podría utilizar diferentes sistemas de recomendación para ejecutar la misma tarea. Estos sistemas de recomendación, además de ser capaces de combinar el poder de diferentes fuentes de datos, permiten mejorar la efectividad de un sistema de recomendación, combinando varios modelos del mismo tipo [9].

1.4.2 Herramientas de desarrollo

1.4.2.1 Python

Python es conocido por ser un lenguaje de programación considerado de alto nivel, que permite el uso de estructuras de datos de alto nivel, además de tener una escritura y enlace dinámico, que lo hacen un lenguaje llamativo para aplicaciones que necesitan ser desarrolladas rápidamente, así como para la creación de scripts [10]. Además, admite el uso de módulos y paquetes externos, lo que permite reutilizar códigos.

1.4.2.2 Pandas

Pandas es un conjunto de módulos y librerías de código abierto desarrolladas en Python, que permiten la manipulación y análisis de datos [11]. Algunas de las facilidades que ofrece esta herramienta son el manejo inteligente de datos faltantes, la manipulación eficiente de datos con estructuras DataFrame, la lectura y escritura de datos en archivos de texto, archivos CSV y bases de datos, etc.

2. METODOLOGÍA

En este trabajo se realizará un sistema de recomendación basado en el conocimiento aplicado a bases de datos farmacológicas. Este sistema busca utilizarse para poder ayudar con el proceso de reposicionamiento de fármacos. El sistema de recomendación utilizará actividades entre compuestos y células, junto con información sobre relaciones existentes entre células, para poder generar predicciones de actividades que podrían existir entre células con compuestos existentes.

Para lograr la creación de este sistema de recomendación em primer lugar es necesario acudir a bases de datos farmacológicas que cuenten con información referente a actividades entre compuestos y células.

Con la intención de obtener esta información se acudió a la base de datos ChEMBL, la cual agrupo esta información a partir de resultados de ensayos ejecutados en documentos científicos.

En cuanto a la información de las relaciones entre las células se acudió a trabajos relacionados, los cuales exponen a la información de estas células en un plano espacial, permitiendo conocer las relaciones existentes entre las distintas células.

A fin de alcanzar el objetivo de este trabajo se empleó la herramienta Anaconda, la cual hace uso del lenguaje Python y es utilizada en el análisis de datos. Esta herramienta junto con el uso de cuadernos Jupyter permite realizar las operaciones necesarias en el lenguaje Python para el tratamiento de datos, así como la generación del sistema de recomendación.

```
connection = sqlite3.connect('../../chembl_29.db')
cursor = connection.cursor()
query = "SELECT ACT.ACTIVITY_ID, ASY.ASSAY_ID, ACT.MOLREGNO ,
ASY.CELL_ID, ACT.STANDARD_VALUE,ACT.STANDARD_UNITS, ACT.STANDARD_TYPE " \
    " FROM ACTIVITIES ACT INNER JOIN ASSAYS ASY ON ACT.ASSAY_ID =
ASY.ASSAY_ID" \
    " WHERE ASY.CELL_ID IS NOT NULL"
cursor.execute(query)
activity_dictionary = cursor.fetchall()
activity_df =
pd.DataFrame(activity_dictionary,columns=['activity_id','assay_id','compo
und_id','cell_id','std_value','std_units','std_type'])
```

Fig. 1 Proceso de recolección de datos

Se acudió a la versión sqlite3 de la base de datos, además con la ayuda de una conexión a la base se obtuvo toda la información necesaria y se la mapeo en un DataFrame para facilitar su análisis y manipulación, este proceso se encuentra en la Fig. 1.

De la base de datos ChEMBL se recogió información referente a las actividades existentes entre compuestos y células, su valor, unidades y tipo de actividad que registran.

2.1 Tratamiento de información de células y compuestos

Al efectuar la revisión de la información requerida se obtuvo alrededor de 5.590.425 registros, en los cuales se realizó un proceso de limpieza y agrupación de información relevante.

Se consideraron los siguientes tipos de actividades entre células y compuestos:

- IC50: (Concentración inhibidora media máxima) Indica la dosis necesaria de un compuesto para reprimir una función biológica
- CC50: (Concentración citotóxica) Se refiere a la concentración de un compuesto que provoca el 50% de muerte celular.
- EC50: (Mitad de la concentración efectiva máxima) Plantea la dosis necesaria de un compuesto para ganar el 50% de un efecto máximo.
- GI50: (Determinación de la actividad antiprolifera) Describe a la concentración de un compuesto que inhibe el 50% del crecimiento celular

También se estandarizó las unidades de estas actividades para ser tratadas en una sola, las unidades consideradas son las micro Moles (uM). El proceso de agrupación de actividades junto con la selección de las unidades seleccionadas para la transformación se encuentra en la Fig. 2.

```
raw_act =
activity_df[activity_df['std_type'].isin(['IC50','CC50','EC50','GI50'])]
raw_act = raw_act[raw_act['std_units'].isin(['nM', '10^4M', '/uM', "10'-
11uM", "10'10uM", "10'8pM", "10'7pM", "10'6pM", "10'5pM", "10'-4nM",
"10'6uM", "10'5uM", 'µM'])]
```

Fig. 2 Agrupación de actividades por tipo y unidades

Al contar con un conjunto de datos como se muestra en Fig. 3, se procedió a clasificar las actividades, de modo que aquellas que presentaran un valor menor a 10 uM se las consideraría como una actividad y aquellas con un valor mayor a 10 uM se las consideraría

como una no actividad, esto debido a que se considera al valor de 10uM como la cantidad necesaria para que exista interacción entre un compuesto y enfermedad, siendo tratadas con valores de 1 y -1 respectivamente, este proceso se encuentra en Fig. 4.

activity_id	assay_id	compound_id	cell_id	std_value	std_units	std_type
161152	5	556	163	600000.0	uM	CC50
162404	5	82983	163	90000.0	uM	CC50
164804	5	1633207	163	340000.0	uM	CC50
166046	5	1633675	163	2400.0	uM	CC50
166057	5	47675	163	7000000.0	uM	CC50

Fig. 3 Conjunto de datos agrupados

```
raw_act['active'] = np.where(raw_act['std_value']<=10.0,1,-1)
```

Fig. 4 Proceso para asignar valores de actividades

En una revisión posterior se encontró varios valores repetidos e inconsistentes en las actividades, es decir que se encontró registros de relaciones entre compuestos y células para los que existía valores de actividades y no actividades, presentando una contradicción en los resultados. Estos problemas se deben a la procedencia de los datos ya que estos se recogieron en la base de datos ChEMBL, pero son parte de varios ensayos científicos, causando que existan casos donde las actividades entre compuestos y líneas celulares se repiten o se contradicen presentando valores de actividad y no actividad. Para tratar estos problemas se eliminaron en primer lugar los duplicados y seguido se tomó únicamente a aquellas relaciones entre compuesto y célula que presenten únicamente una actividad, debido a que no es posible determinar si se trata de una actividad o una no actividad en el caso que se presente una inconsistencia, este proceso se encuentra en la Fig. 5.

```
act=pd.read_csv('summary29.csv',index_col=0).drop_duplicates(keep='first'
)
df_count = act.groupby(by=['compound_id','cell_id'])
index1 = pd.MultiIndex.from_arrays([act[col] for col in ['compound_id',
'cell_id']])
index2 = df_count.size()[df_count.size()==1].index
summ_act=act.loc[index1.isin(index2)]
```

Fig. 5 Proceso para el tratamiento de datos

Con la eliminación de datos inconsistentes y duplicados de actividades entre compuestos y células, se consideró un Dataframe que contiene la información del compuesto, célula y la actividad existente entre estos, como se encuentra en la Fig. 6.

compound_id	cell_id	active
556	163	-1
82983	163	-1
1633207	163	-1
1633675	163	-1
47675	163	-1

Fig. 6 Dataframe consistente

2.2 Tratamiento de información de células

El conjunto de información de las relaciones entre las células presenta a las células en un modelo de hiperespacio, representando a cada célula como una hiperesfera. Se cuenta con el nombre de la célula, su ubicación en el espacio y el radio de la hiperesfera, de manera que los datos se encuentran como se muestra en la Fig. 7.

cell_name	center	r2
CVCL_2260	[0.13507872568263754, 0.0077602583237698875, -...	0.375213
CVCL_4806	[0.1438778206697353, 0.04852056233703366, 0.00...	0.596738
CVCL_M605	[0.16614108207377254, -0.03146658964662872, -0...	0.461877
CVCL_0464	[0.14870812739929176, -0.03909651168168045, -0...	0.389280
CVCL_8987	[0.14744802491027653, -0.006443613530958713, 0...	0.390232

Fig. 7 Información de celulas

Para encontrar la relación que existe entre las células en el hiperespacio, se consideró un concepto geométrico conocido como el valor de intersección, el cual consiste en la suma de sus radios menos la distancia existente entre sus puntos de origen y esto dividido para la suma de sus radios. Este cálculo se encuentra en Ec. 1, donde r_1 y r_2 son los radios de las células y d es la distancia entre estas.

$$intersección = \frac{r_1 + r_2 - d}{r_1 + r_2}$$

Ec. 1 Ecuación de cálculo de la intersección

Este valor de intersección permite establecer una métrica de similitud entre células, de manera que aquellas células que cuenten con un mayor valor de intersección, se debe a que entre estas células existe una mayor relación, el proceso para el cálculo de este valor se encuentra en Fig. 8.

```
for cell_a in cell_rad_df.iterrows():
    cell_a_centroid =  eval(cell_a[1][1])
    cell_r1 = cell_a[1][2]
    print(cell_a[1][0])
    for cell_b in cell_rad_df.iterrows():
        if cell_a[0]!=cell_b[0]:
            cell_b_centroid =  eval(cell_b[1][1])
            cell_r2 = cell_b[1][2]
            dist_cell =distancia(cell_a_centroid,cell_b_centroid)
            radios =cell_r1 + cell_r2
            intersec = (radios - dist_cell) / radios
            intersections_matrix[cell_a[0]][cell_b[0]] =intersec
        else:
            intersections_matrix[cell_a[0]][cell_b[0]] = 0
```

Fig. 8 Proceso para encontrar valores de intersección

Con esta información se genera un conjunto de 10 células que más se relacionen con cada una de las células. Junto con este conjunto se añadió un valor de corte, el cual admite relaciones entre células que tengan un valor de intersección mayor a 0.8. También se realizó un mapeo de los nombres de células de este conjunto de datos a sus nombres pertenecientes a la base de datos ChEMBL para poder utilizar los datos en la generación del sistema de recomendación. Este último mapeo se realizó ya que la información de relaciones entre células se obtuvo de una fuente distinta mas no es parte de la base ChEMBL, el proceso donde se realiza este mapeo y la creación del conjunto de 10 células se encuentra en la Fig. 9. A fin de poder usar esta información se la ubico en un DataFrame como se presenta en la Fig.10, el cual contiene el nombre de la célula, los nombres del conjunto de 10 de células que tienen un mayor valor de intersección junto con el valor de intersección para cada célula de este conjunto.

```
lista= []
for label,content  in df_inter_cell.items():
    if label != 'cell_name':
        id = int(label)
        name = df_inter_cell['cell_name'][id]
        top10 =
df_inter_cell.nlargest(10,str(id))[['cell_name',str(id)]]
        lista_top10 = top10['cell_name'].tolist()
        lista_val_top10 = [val for val in top10[label].tolist() if val
>= 0.8]
        lista_top10= lista_top10[:len(lista_val_top10)]
        if len(lista_top10) !=0:
            df_name_map =
df_mapeo_cell[df_mapeo_cell['cellosaurus_id'].isin([name])]
            if  not df_name_map.empty:
                cells =[]
                values=[]
                for index,cell_name in enumerate(lista_top10):
                    df_name_map_top10 =
df_mapeo_cell[df_mapeo_cell['cellosaurus_id'].isin([cell_name])]
                    if not df_name_map_top10.empty:
                        cells.append(df_name_map_top10['cell_id'][0])
                        values.append(lista_val_top10[index])
                lista.append([df_name_map['cell_id'][0],cells,values])
listas_top10_mappings =pd.DataFrame(lista,
columns=['cell_id','cell_top10','value_top10'])
listas_top10_mappings.to_csv('listas_top10_mapping.csv',index=0)
```

Fig. 9 Proceso de mapeo y creación de top 10 de celulas

cell_id	cell_top10	value_top10
433	[1501, 1262, 1434, 622, 1478, 651, 1220, 646]	[0.8292894588490054, 0.8187765547886937, 0.817...
163	[698, 651, 646, 1501, 641, 786, 165, 633, 623,...	[0.8956966327324026, 0.8660311334932109, 0.860...
377	[408, 1506, 646, 633, 641, 786, 651, 405]	[0.9168315626718784, 0.8976510508479817, 0.849...
620	[1611]	[0.8098843080753313]
621	[478, 610, 1019]	[0.8294095294386521, 0.8142067128807, 0.812799...

Fig. 10 Top 10 de celulas con mayor valor de interseccion

2.3 Sistema de recomendación

El sistema de recomendación recibe como entrada un conjunto de relaciones entre compuestos y células, con el objetivo de recomendar una actividad para esta relación.

El proceso que se realiza por cada relación entre compuesto y célula es el siguiente:

1. Se obtiene el top 10 de células más parecidas con la célula que se relaciona el compuesto, se hace una revisión para conocer si existe actividad entre las células del top 10 y el compuesto de la relación original.

2. En el caso de que exista una actividad, esta es registrada para poder determinar el valor de actividad a recomendar.

3. Al obtener los valores de actividad para todas las células pertenecientes al top 10 que interactúan con el compuesto, se calcula el valor de actividad a recomendar, el cual será la actividad que se presente con mayor frecuencia, este proceso se encuentra en Fig. 11.

```
    for index, cell_interact in cell_interact_compound.iterrows():
        df_cell_interact_top10 =
lista_mapp[lista_mapp['cell_id'].isin([cell_interact['cell_id']])]
        lista_actividades=[]
        lista_cell =[]
        active_prediccion = 0
        cell_top10
=eval(df_cell_interact_top10['cell_top10'].tolist()[0])
        for cell in cell_top10:
            cell_exist = summary[summary['cell_id'].isin([cell])]
            cell_exist1=cell_exist[cell_exist['compound_id'].isin([comp])
]
            if len(cell_exist1)>0:
                lista_actividades.append(cell_exist1['active'].values[0])
            else:
                lista_cell.append(cell)
        if len(lista_actividades)>1:
            pred_a= lista_actividades.count(1)
            pred_i= lista_actividades.count(-1)
            if pred_a != pred_i :
                active_prediccion = mode(lista_actividades)
                lista_respuesta.append([comp,cell_interact['cell_id'],act
ive_prediccion])
                for cell in lista_cell:
                    lista_extendida.append([comp,cell_interact['cell_id']
,cell,active_prediccion])
```

Fig. 11 Sistema de recomendación

En el caso de que ninguna de las células del top 10 interactúe con el compuesto no es posible generar una recomendación. Tampoco es posible generar una recomendación en el caso de que no exista o no se tenga información del top 10 de una célula.

Al terminar este proceso, se generará una actividad para cada relación entre compuestos y células que se recibieron en el conjunto de datos que se ingresó en un principio.

3. PRUEBAS, RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Pruebas

Como prueba para el sistema de recomendación, se realiza una evaluación por utilidad, donde se divide a los datos en datos de entrenamiento y testeo. Ya que el modelo actual se entrenó con la base de datos ChEMBL en su versión 29 se necesitaría de un conjunto de datos de prueba, para lo cual se utiliza la base de datos ChEMBL en su versión 30, generando recomendaciones para aquellos compuestos de la versión 30 que no se encuentran en la versión 29.

El conjunto de datos que se usara para evaluar la eficacia del sistema de recomendación se lo almacena en un DataFrame que se obtiene de la resta entre la información de la versión 30 menos la versión 29, este proceso se encuentra en Fig. 12.

```
v30 = pd.read_csv('act_summary30.csv',index_col=0)
v29 = pd.read_csv('act_summary29.csv',index_col=0)
#resta v30 - v29
index1 = pd.MultiIndex.from_arrays([v30[col] for col in ['compound_id',
'cell_id']])
index2 = pd.MultiIndex.from_arrays([v29[col] for col in ['compound_id',
'cell_id']])
v30_ni=v30.loc[~index1.isin(index2)]
v30_ni.columns =['compuesto','cell','active']
v30_ni
```

Fig. 12 Creación del conjunto de datos para usarlo en pruebas

Estas pruebas determinarán en qué medida el sistema de recomendación es capaz de recomendar actividades para relaciones entre compuestos y líneas celulares que aún no hayan sido expuestas en ensayos químicos y recogidas por la base de datos ChEMBL.

3.2 Resultados

Se llevo a cabo el uso de un conjunto de datos construidos a partir de la versión 30 de la base de datos ChEMBL en el sistema de recomendación, donde se obtuvo 305 recomendaciones de actividades y no actividades. Es necesario tener en cuenta que el modelo del sistema de recomendación es un trabajo inicial y se entrenó con registros de la base de datos ChEBML en su versión 29, por lo que se llevó a cabo su evaluación con datos de la versión 30 que no se incluían en la versión 29 de la base de datos, lo que resulto en un universo muy pequeño para el cual encontrar resultados. Los datos que se

obtuvieron de resultado se agruparon en un Dataframe de manera que se pueden visualizar aquellos valores diferentes entre el valor real y el que se obtuvo por el sistema de recomendación, como se muestra en la Fig. 13.

compuesto	cell	prediccion	realidad
1907	789	-1	-1
1836	789	-1	-1
111027	498	-1	-1
110451	498	-1	-1

Fig. 13 Conjunto de datos con resultados obtenidos por el sistema de recomendacion

Para interpretar de mejor manera los resultados obtenidos, se realizó una representación de estos a través de una matriz de confusión, para la cual se utilizó la librería de Python “sklearn metrics” que implementa varias funciones para medir el rendimiento de una clasificación, este proceso encuentra en la Fig. 14.

```
from sklearn.metrics import confusion_matrix
y_true = df_comparaciones['prediccion'].tolist()
y_pred = df_comparaciones['realidad'].tolist()
confusion_matrix(y_pred, y_true)
```

Fig. 14 Calculo de la matriz de confusión

Se obtuvo una matriz de confusión, que muestra 299 recomendaciones de no actividades junto a 6 recomendaciones de actividades, los resultados se encuentran en la Tabla 1.

Tabla 1 Tabla de confusión para el sistema de recomendación

297	2
0	6

La matriz de confusión nos presenta que en las recomendaciones de actividad positiva existen 6 valores verdaderos positivos junto con ningún valor para falsos positivos, que representaría a un 100% de recomendaciones verdaderas para una actividad positiva. Además, para aquellas recomendaciones de actividad negativa existen 297 valores verdaderos negativos, junto con 2 valores falsos negativos, que representaría un 99.33% de recomendaciones verdaderas para una actividad negativa.

3.3 Conclusiones

Se logró obtener un sistema de recomendación basado en el conocimiento capaz de realizar recomendaciones de actividades entre compuestos y líneas celulares, con una fidelidad del 100% en actividades y con un 99.33% en el caso de no actividades para un conjunto de datos construido a partir de la versión 30 de la base de datos ChEMBL.

El correcto análisis de información de la base de datos farmacológica ChEMBL permitió contar con un conjunto de datos estandarizados con los cuales se logró establecer relaciones existentes entre compuestos y líneas celulares, que son de gran ayuda para la generación de recomendaciones.

El uso de un valor de interacción entre las células ayudo a cuantificar la relación existente entre células, relación de suma importancia que ayudaría a conocer cuando un compuesto podría tener una actividad o no actividad con una célula con la cual no existe relación, pero cuenta con un alto valor de interacción con una célula con la cual si existe relación para este compuesto.

El sistema de recomendación generado puede ser actualizado y mejorado con información de nuevas versiones de la base datos, que, junto con información adicional de las relaciones entre células existentes, sería capaz de generar predicciones de actividades con mayor precisión.

3.4 Recomendaciones

Se recomienda realizar una investigación más profunda con el objetivo de aumentar la información de actividades entre compuestos y células, así como información de las interrelaciones entre células, lo cual permitiría aumentar el número de recomendaciones generadas por el sistema.

De igual manera, se recomendaría tomar en cuenta diferentes tipos de actividades aparte de las ya consideradas que son IC50, CC50, EC50 y GI50, lo cual podría aumentar el número de resultados obtenidos con el sistema de recomendación actual.

.

4. REFERENCIAS BIBLIOGRÁFICAS

[1] Farmaindustria, «¿Cuánto tiempo se tarda en desarrollar un medicamento?,» farmaindustria.es, 01 Abril 2020. [En línea]. Available: https://www.farmaindustria.es/web/reportaje/cuanto-tiempo-se-tarda-y-por-que-en-desarrollar-un-medicamento/. [Último acceso: 10 Agosto 2022].

[2] J. Quintana y F. Palau, Drug repurposing for rare diseases. The cure for the 21st century., Cosmocaixa Barcelona: Biocat, 2016, p. 3.

[3] S. Rosa y W. Santos, «Clinical trials on drug repositioning for COVID-19 treatment,» *Revista Panamericana de Salud Pública,* vol. 44, p. 1, 2020.

[4] Y. Hua y X. Dai, «Drug repositioning: Progress and challenges in drug discovery for various diseases,» *European Journal of Medicinal Chemistry,* vol. 234, p. 4, Abril 2022.

[5] J. Naveja, A. Dueñas y J. Medina, «Drug Repurposing for Epigenetic Targets Guided by Computational Methods,» *Epi-Informatics,* p. 327, 2016.

[6] D. Mendez, «ChEMBL: towards direct deposition of bioassay data,» *Nucleic Acids Research,* p. 3, 2018.

[7] J. Medina y F. Saldívar, «Descubrimiento y desarrollo de fármacos: un enfoque computacional,» *Educación Química,* vol. 28, nº 1, 2017.

[8] C. Aggarwal, Data Mining, New York: Springer International Publishing, 2015, pp. 619-620.

[9] C. Aggarwal, Recommender Systems, New York: Springer International Publishing, 2016, pp. 15-16.

[10] Python, «What is Python?,» Python Software Foundation, 2022. [En línea]. Available: https://www.python.org/doc/essays/blurb/. [Último acceso: 2022].

[11] Pandas development team, «About pandas,» Pandas, 23 01 2022. [En línea]. Available: https://pandas.pydata.org/about/index.html. [Último acceso: 08 2022].

[12] A. Carranza y A. Segura, «Reposicionamiento de fármacos identificados por métodos computacionales (SVBS), para su uso como terapias contra el cáncer,» *Salud Jalisco,* vol. 7, nº 1, p. 1, 2020.

5. ANEXOS

I. ANEXO 1: Enlace al repositorio

Trabajo De integración Curricular Knowledge Based Recommender System

https://gitlab.com/Ramsesboom98/Trabajo-de-Integracion-Curricular-knowledge-based-recommender-system.git