دانشکده هوش مصنوعی، گروه کامپیوتر

عنوان

راهنمای پروژه تحلیل داده های ویروس کرونا و ارائه مناسب ترین راه مقابله دارویی با آن توسط مدل سازی داده ها و یادگیری ماشین

دانشجو

محمد حسین شوندی

استاد

دکتر امینه امینی

شماره دانشجویی

39911541054042

بهمن 1402

فهرست

مقدمه

با راهنمایی استاد محترم و علاقه‌ای که به حوزه بیوانفورماتیک دارم، خوشبختانه توانستم با این پروژه قدمی جدی برای ورود به این رشته جذاب و پر رمز و راز بردارم. این پروژه برایم فرصتی بود تا با موضوعاتی مثل یادگیری ماشین، مدل‌سازی، ساختارهای شیمیایی مولکول‌ها، و بیوتکنولوژی‌های مختلف بیشتر آشنا شوم و دانش خود را در این زمینه گسترش دهم.

این پروژه، در واقع بازسازی یکی از نمونه‌های اولیه و واقعی در حوزه بیوانفورماتیک است که توانستم با طبقه‌بندی و نکته‌برداری‌های خودم آن را به اجرا درآورم. برای درک بهتر و پیشبرد این پروژه، از منابع مختلفی همچون chatgpt, GitHub ، YouTube و Udemy (لینک :آموزش بیوانفورماتیک از پایه) بهره گرفتم.

این تازه آغاز مسیر من در بیوانفورماتیک است و با توجه به دانش به‌دست‌آمده و حمایتی که از جانب استاد محترم دارم، مطمئن هستم که می‌توانم در آینده جایگاه شایسته‌ای در این رشته کسب کنم.

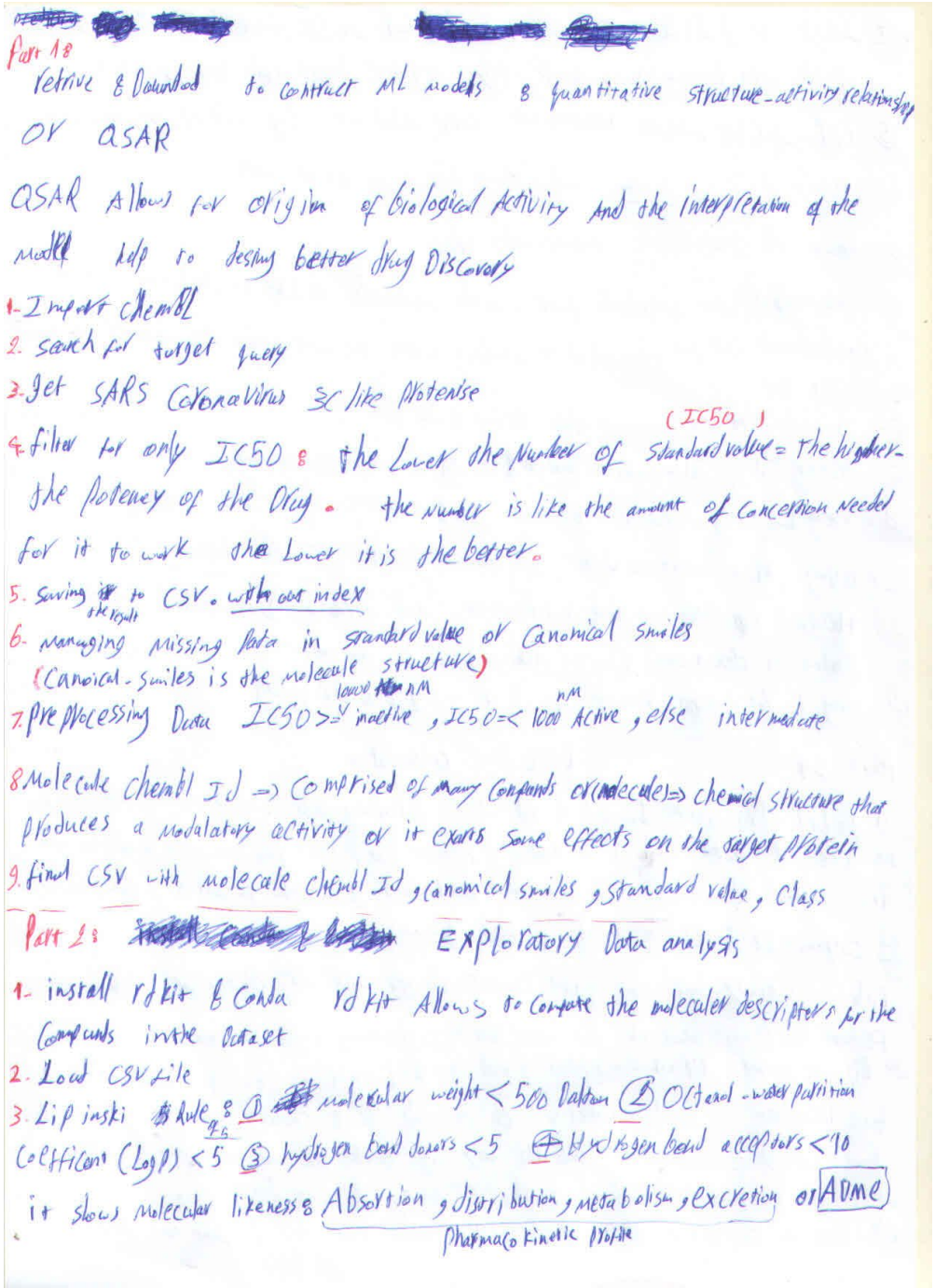
با تشکر از شما.

کتابخانه های استفاده شده:

Streamlit	pandas	base64	subprocess	os
pickle	padelpy	Numpy	seaborn	sklearn
lazypredict	scipy	chembl_webresource_client		

نکته برداری های دستی (به زبان انگلیسی)

در ادامه طبقه بندی ها و نکته برداری های خودم را در سه صفحه آینده به صورت عکس قرار می دهم.


Part 1 B

retrive & Download to construct ML models & quantitative structure-activity relationship OR QSAR

QSAR Allows for origin of biological Activity And the interpretation of the model help to design better drug Discovery

1- Import chembl

2. search for target query

3. Get SARS Coronavirus 3C like Protease

4. filter for only IC50 & the Lower the Number of (IC50) Standard value = the higher the potency of the Drug. the number is like the amount of concertion needed for it to work the Lower it is the better.

5. Saving the result to CSV. without our index

6- managing missing data in standard value or Canonical Smiles (Canoical-Smiles is the molecule structure)

7. preprocessing Data IC50 >= lower than nM inactive, IC50 =< 1000 nM Active, else intermadiate

8 Molecule chembl ID => Comprised of many Compounds or (molecule) => chemical structure that produces a modulatory activity or it exerts some effects on the target protein

9. final CSV with molecule chembl Id, canonical smiles, Standard value, Class

Part 2: Exploratory Data analysis

1- install rdkit & Conda rdkit Allows to compute the molecular descriptors for the Compounds in the Dataset

2. Load csv file

3. Lipinski Rule: ① molecular weight < 500 Dalton ② Octanol-water partition Coefficient (Log P) < 5 ③ hydrogen bond donors < 5 ④ Hydrogen bond acceptors < 10 it shows molecular likeness & Absorption, distribution, metabolism, excretion or ADME Pharmacokinetic Profile

④ IC50 to pIC50 is because the original IC50 value has uneven distribution of the data points. we should Apply Negetive Logarithmic transformation

⑤ Cap IC50 value to 100,000,000 cause other wise -Log will be Negetive Value

⑥ Apply pIC50 function and change pIC50 for IC50 value

⑦ Remove All Intermediate bioActivity class

⑧ Data Analysis for chemical space, Active molecule is Like constellation, we have constellation plot for chemical space Analysis where the more Active the molecule the Larger size it has

8-1 frequecy plot for bioActivity class using count plot

8-2 Scatter plot of MW (molecular weight Vs LogP) using scatter plot

8-3 using boxplot for pIC50 6 > Active 5 < = inactive

8-4 using Mann-whitney U test for statical significance of the difference

8-5 H0 (null Hypothesis) says there is no difference And difference is Accidental) but our p value is close to zero (less than Alpha = 0.05) so it's Not valid.

⑨ using boxplot & Mann whitney for all 4 of Lipinski values

Part 3 Descriptor calculation

① Pa DeL - Descriptor Download a tool for calculating molecular descriptors And fingerprints from chemichal structure it is to Analyz & represent the properties for molecules in quantive matter its java-based API that shows 1 molecular Descriptors 2 molecular fingerprints 3 Compatibility 4 customizable Descriptor types 5 Batch Proccecing

pubchem a Binary-based value that encodes the presence or Absence of specific structural features. used for molecular similarity searching, clustering & classification

* Difference Between Lipinsk-Discripters & pubchem & Lipinski is gloal features (the rule of five) but Pubchem is local features & each molecule Like containg seveal Lego Blocks can have varaity of structures And the way lego blocks are connected it will create a unige set of properties And thats the essance of drug Disscovery.

we have to find A way that the molecule Provides the most potency and safe.

Part 4: Regression with Random forrest

(1) Pubchem fingerprints is unige to each molecule Like humans And we can create a module to distinguish Compounds that are Active & inactive And we Are using pIC50 => -Log IC50, IC50 is the inhibition Concentration at 50

(2) we Drop PIC50 to use it as y and rest for X

(3) Delete Low variance Data with low variance threshold => (0.8 * (1-0.8))

(4) Split Data to 80/20 with sklearn.module_selection

(5) Building Regression module with Random forest And make predictions

(6) with seaborn build plot to show difference of predicted pIC50 And our data

Part 5: module compression

(1) with Using Lazy predict Library we can do a quick & Rapid module building & classification and also Regression module in few lines of coding : Lazy predict and Lazy regressor

(2) we are comparing several ML module with three factor & [1] RMSE (Root mean Squared Error) & it predicts how far off on average the modles predictions are from the actual values, Lower the better [2] - R-squared (Coefficient of Determination) & Ranges from 0 to 1 the higher value indicates a better fit. R-squared 1 mens the model explains all the variance in the data. [3] time of the module Running.

Part 6: module deployment

we shall deploy A web App to use to predict the ~~target~~ protein target

More will be said in the pdf

در ادامه توضیح مختصری راجب پروژه به شما ارائه می دهم:

این پروژه نحوه استفاده از پایتون و یادگیری ماشین در کشف دارو را ارائه می‌دهد و شامل جمع‌آوری داده، پیش‌پردازش، تحلیل اکتشافی، ساخت مدل و استقرار آن به‌عنوان یک وب‌اپلیکیشن است:.

همچنین به مقایسه مدل‌ها و انتخاب مناسب‌ترین مدل پرداخته و یک اینفوگرافیک خلاصه برای مرور آسان فراهم می‌شود.

(برای اجرای بخش 1 تا 5 با استفاده از Jupyter Notebook فقط کافی است فایل ها اجرا شود. کتابخانه های لازمه به صورت خودکار نصب می شوند.)

معرفی پایگاه داده ChEMBL ، که شامل بیش از 2 میلیون ترکیب شیمیایی از 76,000 سند است، و نصب ChEMBL web resource clientبرای دسترسی به داده‌های فعالیت زیستی.

کاوش در پایگاه داده ChEMBL ، شامل ورژن و تعداد ترکیبات.

نصب ChEMBL client با استفاده از pip برای دانلود داده‌های فعالیت زیستی.

نحوه دریافت داده‌های بیواکتیویته برای پروتئین‌هایی مانند پروتئیناز ویروس کروناSARS ، با تمرکز بر قدرت تأثیر و یکپارچگی داده‌ها.

نمایش مراحل پیش‌پردازش داده‌های فعالیت زیستی، شامل دسته‌بندی ترکیبات بر اساس مقادیر IC50 به فعال، غیرفعال یا متوسط.

ایجاد پوشه‌ها، بررسی محتوای داده‌ها، و دسته‌بندی ترکیبات بر اساس IC50 برای استفاده در مدل‌های یادگیری ماشین در تحقیق دارویی.

محاسبه توصیف‌گرهای مولکولی با استفاده از RDKit و معرفی توصیف‌گرهای لیپینسکی، با توضیح اصول ارزیابی داروپذیری براساس ویژگی‌های فارماکوکینتیکی.

توضیح مقیاس‌بندی مقادیر IC50 به pIC50 برای تحلیل در پروژه بیوانفورماتیک.

نمایش تکنیک‌های دست‌کاری و تحلیل داده‌ها در بیوانفورماتیک، شامل افزودن ستون‌های جدید، انجام تحلیل اکتشافی، و انجام آزمون‌های آماری روی کلاس‌های فعالیت زیستی.

بررسی داده‌های اکتشافی و تحلیل فضای شیمیایی با نمودارها و ذخیره نمودارها برای گزارش.

استفاده از ابزارهایی مانند Paddle برای محاسبه سریع‌تر توصیف‌گرهای مولکولی و بهبود دقت پروژه‌های بیوانفورماتیک.

دانلود، پردازش و آپلود فایل‌های محاسبه‌شده و آماده‌سازی ماتریس‌های داده برای مدل رگرسیون با الگوریتم جنگل تصادفی، با استفاده از اثر انگشت‌های PubChem برای بازدارنده‌های آنزیم استیل‌کولین استراز.

ساخت و مقایسه مدل‌های یادگیری ماشین با کتابخانهlazy predict ، با تمرکز بر ارزیابی عملکرد مدل و ایجاد وب‌اپلیکیشن برای پیش‌بینی‌های کاربران.

نحوه بارگذاری فایل ورودی حاوی داده ساختار شیمیایی، تولید اثر انگشت‌های مولکولی، پیش‌بینی مقادیر فعالیت زیستی، و بهینه‌سازی مدل با کاهش ویژگی‌های تکراری برای تسریع در ساخت مدل.

ساخت مدل یادگیری ماشین با استفاده از رگرسیون جنگل تصادفی و ذخیره مدل با استفاده از Pickle برای استفاده‌های آینده.

ساخت یک ابزار بیوانفورماتیک با پایتون، شامل بارگذاری داده‌ها، محاسبه توصیف‌گرهای مولکولی، ساخت مدل و ارائه پیش‌بینی‌ها، با تأکید بر یادگیری علم داده از طریق کاربرد عملی.

هدف پروژه:

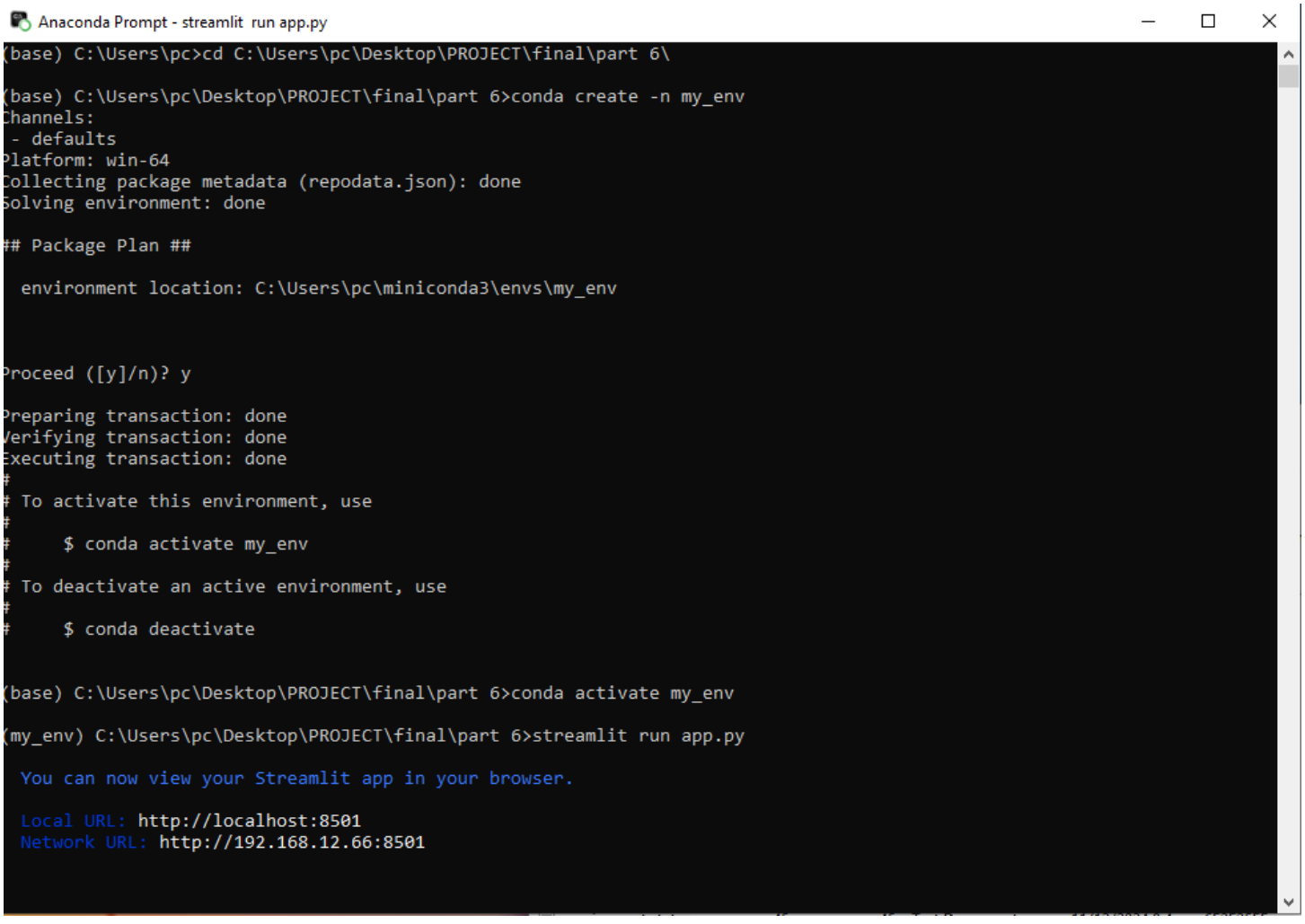
هدف نهایی این پروژه توسعه یک مدل یادگیری ماشین است که بتواند فعالیت زیستی ترکیبات را در برابر یک پروتئین هدف خاص پیش‌بینی کند. این کار با استفاده از مقایسه IC50 و PIC50 که نشان دهنده قدرت اثر این ترکیب در مقابل پروتئین هدف است.

IC50 یا "غلظت بازدارنده نیمه‌حداکثری" یک معیار رایج در بیوشیمی و داروسازی است که نشان می‌دهد چه غلظتی از یک ترکیب (مثل دارو) لازم است تا 50٪ از فعالیت یک آنزیم یا گیرنده خاص را مهار کند. این معیار به‌عنوان شاخصی از قدرت و کارآیی یک مهارکننده استفاده می‌شود؛ هرچه مقدار IC50 کمتر باشد، دارو یا ترکیب مؤثرتر است و می‌تواند با غلظت کمتری فعالیت زیستی موردنظر را مهار کند.

اجرای بخش 6

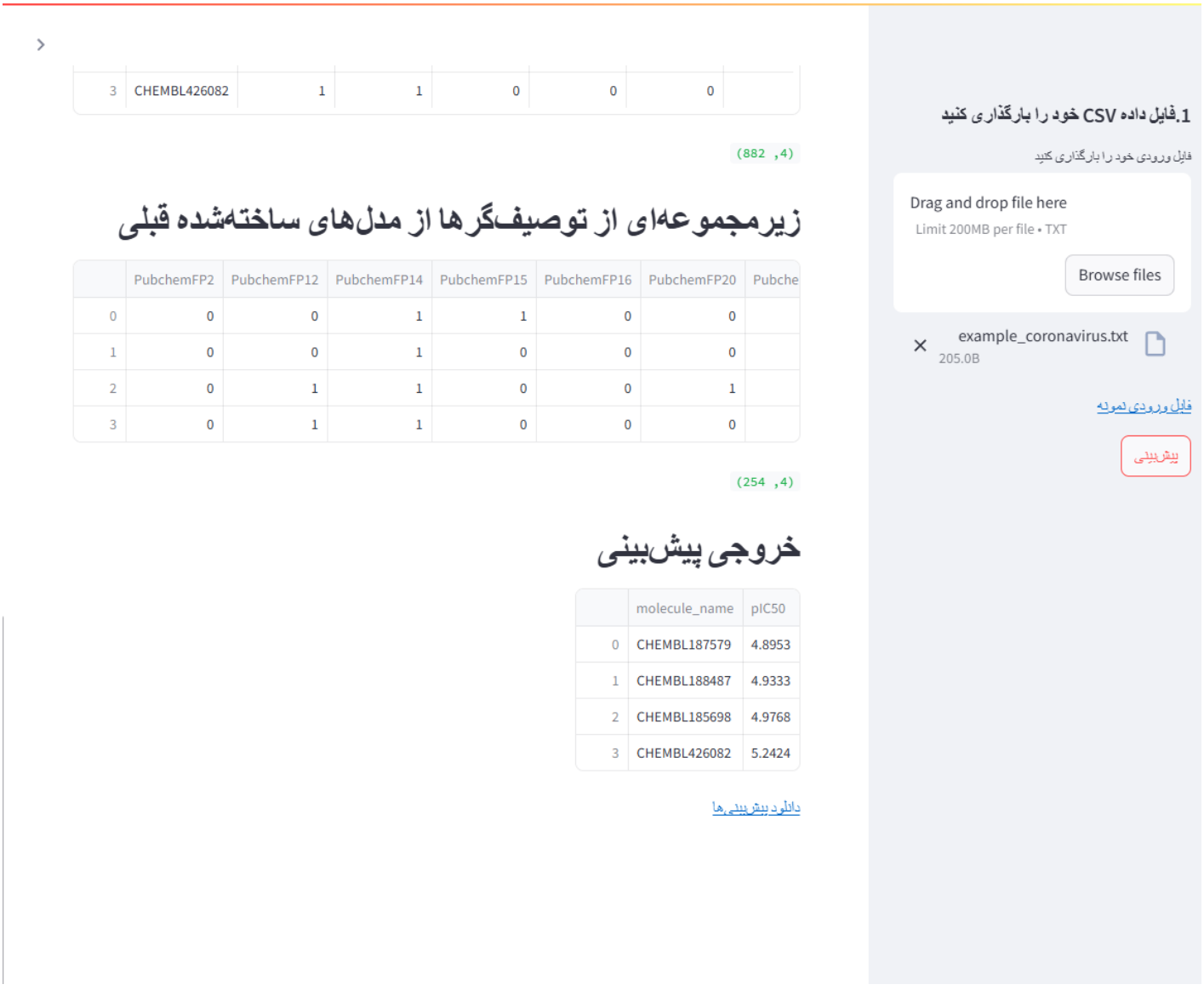
اول از همه در فایل part 6 این فایل را اجرا کرده : bioactivity_prediction_app.ipynb تا نتایجی مانند پنج بخش قبلی را در csv برای اجرای webapp آماده کنیم.

سپس از ترمینال anaconda prompt که در بخش دو نصب کردیم دستورات زیر را وارد کرده:



همانطور که در عکس میبیند با استفاده از streamlit برنامه app.py را برای اجرای بخش 6 اجرا می کنیم.

پس از اجرا با صفحه زیر مواجه می شویم:



1.فایل داده CSV خود را بارگذاری کنید

فایل ورودی خود را بارگذاری کنید

Drag and drop file here

Limit 200MB per file • TXT

Browse files

example_coronavirus.txt 205.0B

فایل ورودی نمونه

پیش‌بینی

3	CHEMBL426082	1	1	0	0	0

(882 ,4)

زیرمجموعه‌ای از توصیف‌گرها از مدل‌های ساخته‌شده قبلی

	PubchemFP2	PubchemFP12	PubchemFP14	PubchemFP15	PubchemFP16	PubchemFP20	Pubche
0	0	0	1	1	0	0	
1	0	0	1	0	0	0	
2	0	1	1	0	0	1	
3	0	1	1	0	0	0	

(254 ,4)

خروجی پیش‌بینی

	molecule_name	pIC50
0	CHEMBL187579	4.8953
1	CHEMBL188487	4.9333
2	CHEMBL185698	4.9768
3	CHEMBL426082	5.2424

دانلود پیش‌بینی‌ها

با استفاده از فایل ورودی نمونه در سمت راست می توانید به github من و فایل txt داخلش دسترسی پیدا کنید که داخل آن چهار نمونه پروتئین آزمایشی هست.

با اضافه کردن فایل txt و زدن دکمه پیش بینی می توان بهترین گزینه ها برای استفاده در دارو سازی که کمترین PIC50 را دارند را به دست آورد.

(نکته: فرمت فایل ورودی باید مانند فایل نمونه باشد که شامل نام مولکول و ترکیب chembl و ترکیب شیمیایی آن مولکول است.)

با تشکر از صبر و لطف شما.