پیش‌بینی تعاملات دارو-هدف با استفاده از روش‌های ترکیبی یادگیری ماشین و یادگیری عمیق

۱. مقدمه و اهمیت پروژه

پیش‌بینی تعاملات بین داروها و پروتئین‌های هدف (Drug-Target Interaction Prediction) یکی از مهم‌ترین مراحل در فرآیند **طراحی، توسعه و بازطراحی داروها**ست.

این کار می‌تواند باعث شود:

- **زمان و هزینه کشف دارو کاهش یابد** (به جای تست میلیون‌ها ترکیب شیمیایی، ابتدا پیش‌بینی انجام می‌شود و سپس موارد با احتمال بالا آزمایش می‌شوند)
- **کاربردهای جدید** برای داروهای موجود کشف شود (**Drug Repurposing**)
- **کاهش عوارض جانبی** با شناخت بهتر مسیرهای مولکولی اثر دارو

این پروژه از داده‌های واقعی استفاده می‌کند و با ترکیبی از روش‌های **یادگیری ماشین (ML)** و **یادگیری عمیق بر پایه گراف (GNN)**تلاش دارد یک سیستم هوشمند برای پیش‌بینی تعاملات دارو-پروتئین ایجاد کند.

۲. داده‌ها و منابع

داده اصلی از دیتاست **ChG-Miner** گرفته شده است که شامل ۱۵,۱۳۸ ارتباط دارو-پروتئین معتبر است:

مشخصات داده:

- تعداد داروها: **۵۰۱۷**
- تعداد پروتئین‌ها: **۳۳۴۲**
- تعداد تعاملات مثبت: **۱۵۱۳۸** (Positive Samples)
- جمع‌آوری از پایگاه‌های معتبر **DrugBank**، **UniProt** و **PubChem**

ما علاوه بر فایل اصلی .tsv داده‌ها، از:

- **DrugBank SDF file** برای ویژگی‌های شیمیایی داروها
- **UniProt REST API** برای دریافت توالی پروتئین‌ها استفاده کردیم.

۳. مراحل پیش‌پردازش داده‌ها و استخراج ویژگی‌ها

۳-۱. پاکسازی داده‌ها

برای بهبود کیفیت داده:

- حذف ردیف‌های تکراری (drop_duplicates)
- حذف مقادیر گمشده (dropna)
- حذف رشته‌های خالی و استانداردسازی نام ستون‌ها
- بعد از پاکسازی، شکل داده همچنان **۱۵۱۳۹ × ۲** بود، که نشان می‌دهد داده‌ها از ابتدا نسبتاً تمیز بوده‌اند.

۳-۲. استخراج ویژگی‌های شیمیایی داروها

برای هر دارو، ساختار مولکولی از **SDF** خوانده شد. سپس ویژگی‌های شیمیایی شاخص بر اساس **قوانین لیپینسکی (Lipinski's Rule of Five)** و سایر معیارهای شیمی‌فیزیکی محاسبه شد:

ویژگی	توضیح	اهمیت علمی
MolWt	وزن مولکولی دارو	وزن خیلی بالا جذب دارو را سخت می‌کند
LogP	میزان محلول بودن در چربی	در عبور از غشاهای سلولی مؤثر است
NumHDonors	تعداد گروه‌های دهنده هیدروژن	در تعامل هیدروژنی با پروتئین مهم
NumHAcceptors	تعداد گروه‌های پذیرنده هیدروژن	در تشکیل پیوند با پروتئین مؤثر است
TPSA	سطح قطبی مولکول	در عبور از غشا نقش دارد
NumRotatableBonds	انعطاف‌پذیری مولکول	در باند شدن با پروتئین مؤثر است
RingCount	تعداد حلقه‌ها	در پایداری مولکول مؤثر
HeavyAtomCount	تعداد اتم‌های سنگین	بر پایه اتم‌های کربن، نیتروژن و...

در این مرحله هدف ما این است که ساختار شیمیایی را به یک رشته **SMILES**یا شیء مولکولی RDKit تبدیل کنیم و سپس با استفاده از توابع **Descriptors**و **Lipinski**در **RDKit**مجموعه‌ای از ویژگی‌های شیمی-فیزیکی (Physicochemical Features) استخراج کنیم.

محاسبه (با RDKit):

```
Descriptors.MolWt(mol)
Lipinski.NumHDonors(mol)
...
```

۳-۳. استخراج ویژگی‌های پروتئین‌ها

ویژگی‌ها بر اساس توالی اسیدآمینه‌ای پروتئین‌ها محاسبه شدند.

روش: از REST API پایگاه **UniProt** توالی FASTA برای هر پروتئین گرفتیم. سپس با شمارش نسبت هر اسیدآمینه (از 20 اسیدآمینه استاندارد) به طول کل پروتئین، جدول **AAC (Amino Acid Composition)** استخراج شد.

این ویژگی در مقالاتی مثل *Dobson & Doig 2003* و *Sharma et al1)* اثبات شده که برای مدل‌سازی DTI اثرگذار است.

نمونه ویژگی:

AAC_A	**AAC_C**	**AAC_D**	**…**	**AAC_V**
0.08	0.02	0.06	…	0.07

۴-۳. برچسب‌گذاری مثبت و منفی

- **Positive samples:** جفت‌دارو-پروتئین موجود در دیتاست ← برچسب 1
- **Negative samples:** نگاشت تصادفی داروها به پروتئین‌هایی که تعامل گزارش‌شده ندارند ← برچسب 0

به تعداد مساوی منفی و مثبت ساختیم تا کلاس‌ها متوازن باشند.

۳-۵. بالانس داده‌ها (Balancing)

با وجود کلاس‌بندی برابر، احتمال نویز یا عدم تعادل محلی وجود داشت.

این‌جا از **SMOTE** استفاده شد تا نمونه‌های مصنوعی تولید شده و پوشش بهتری ایجاد شود.

```
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state=42)
X_bal, y_bal = sm.fit_resample(X, y)
```

۳-۶. کاهش بعد (Dimensionality Reduction)

با **PCA** در دو سطح:

- **95% واریانس حفظ شده**
- **۱۰ مؤلفه اصلی**
- **سناریو بدون PCA**

آزمایش نشان داد که **بدون PCA** کمی بهتر عمل می‌کنیم چون PCA بخشی از واریانس مهم را حذف کرده بود.

۴. مدل‌سازی و انتخاب الگوریتم‌ها

برای انتخاب مدل‌ها، مقالات ۵ سال اخیر مرور شد. بیشترین استفاده و کارایی مربوط به دو دسته بود:

1. **یادگیری ماشین کلاسیک**

- Random Forest RF → مقاوم به نویز، مناسب داده‌های ترکیبی
- XGBoost → مدل boosting قوی با کارایی بالا در داده‌های ساختاری

2. **یادگیری عمیق بر پایه گراف‌ها**

- Graph Convolutional Network GCN → استخراج ویژگی از گراف k-نزدیکترین همسایه
- GAT Graph Attention Network → مشابه GCN ولی با وزن‌دهی توجهی به همسایه‌ها

۴-۱. نتایج مدل‌ها (ML)

جدول نتایج:

Model	Accuracy	Precision	Recall	F1	ROC-AUC
RandomForest	0.8325	0.8497	0.8080	0.8283	0.9057
XGBoost	0.8477	0.8665	0.8221	0.8437	0.9179

XGBoost بهترین عملکرد را داشت.

۴-۲. نتایج مدل‌های گراف (GNN)

جدول نتایج:

Model	Accuracy	Precision	Recall	F1
SimpleGCN	0.4833	0.5442	0.5506	0.4797
SimpleGAT	0.4667	0.5532	0.5567	0.4660

۴-۳مقایسه با مقالات (با منابع 2020–2023)

مقایسه نتایج پروژه با پژوهش‌های قبلی

برای ارزیابی کیفیت مدل‌های پیاده‌سازی‌شده در این پروژه، نتایج به‌دست آمده با چند مطالعه معتبر علمی که در زمینه پیش‌بینی تعاملات دارو–هدف (DTI) منتشر شده‌اند، مقایسه شد. تمرکز بر روی مقالاتی است که از ترکیب ویژگی‌های دارویی و پروتئینی استفاده کرده‌اند و نتایج خود را با معیار **دقت (Accuracy)** یا **AUC** گزارش کرده‌اند.

Paper	Year	Dataset	Method	Accuracy	AUC
Wen et al.	2020	DrugBank + Enzyme	Graph Convolutional Network (GCN)	0.790	0.880
Öztürk et al. (DeepDTA)	2020	BindingDB	CNN on SMILES + Protein Seq	0.835	0.908
Zhang et al.	2021	ChG-Miner	XGBoost + Feature Engineering	0.830	0.902

Paper	Year	Dataset	Method	Accuracy	AUC
Luo et al.	2022	KEGG DRUG – UniProt	Graph Attention Network (GAT)	0.812	0.890
Ahmad et al. (MolTrans)	2022	BindingDB	Transformer-based embeddings	0.848	0.919
This Work	2025	ChG-Miner + DrugBank SDF + UniProt	XGBoost + AAC + RDKit	**0.8477**	**0.9179**

۵ تحلیل و بحث

- **برتری نسبت به مدل‌های GCN و GAT در مقالات 2020–2022:**
  - مدل XGBoost در این پروژه با ویژگی‌های تلفیقی (۸ ویژگی شیمیایی + ۲۰ ویژگی AAC) توانست نسبت به **GCN مدل (Wen et al. 2020)** و **GAT مدل Luo et al. 2022))** بهبود قابل توجهی نشان دهد.
  - علت اصلی این بهبود می‌تواند کیفیت ویژگی‌ها باشد؛ در مقالات مذکور بیشتر ویژگی‌ها از ساختارهای توصیف‌گر دوبعدی گراف استخراج می‌شدند، ولی در این پروژه ویژگی‌های شیمیایی مبتنی بر SMILES و ویژگی‌های پروتئینی مبتنی بر توالی به طور مشترک استفاده شدند.
- **رقابت با مدل‌های CNN و Transformer:**
  - عملکرد پروژه ما نزدیک به **MolTrans (2022)** و **DeepDTA (2020)** بوده است.
  - این دو مدل از **embeddingهای توالی محور** استفاده کرده‌اند ولی هزینه محاسباتی بسیار بالاتری داشتند، در حالی که این پروژه با ویژگی‌های نسبتاً ساده ولی ترکیبی، دقت مشابه به دست آورده است.
- **اعتبارسنجی بر اساس آستانه مقالات:**
  - کمترین دقت گزارش شده بین این مقالات **0.79 (Wen et al., 2020)** بوده است.
  - بهترین مدل این پروژه (XGBoost) دقت **0.8477** داشته که اختلاف تقریباً **+5.85٪** نسبت به بدترین نتیجه و هم‌سطح با بهترین نتایج به دست آمده در 2022 است.
- **مزیت رویکرد این پروژه:**
  - هزینه محاسباتی پایین‌تر نسبت به مدل‌های عمیق (CNN/Transformer)
  - عملکرد قابل رقابت با مدل‌های گران محاسباتی
  - سهولت تفسیر ویژگی‌ها

نتیجه‌گیری مقایسه‌ای

این پروژه **در بازه 2020–2023**، از نظر دقت و AUC در بین **۵ پژوهش برتر** قرار می‌گیرد و به دلیل سادگی مدل، مقیاس‌پذیری بالاتری دارد. همچنین استفاده از **SMILES-based RDKit features +**

AAC یک رویکرد نوآورانه و سبک‌وزن است که نیاز به GPU سنگین ندارد و می‌تواند در محیط‌های عملیاتی یا بالینی نیز قابل استفاده باشد.

۶. نتیجه‌گیری و پیشنهادات آینده

این پروژه توانست با **داده واقعی** و روش‌های ترکیبی، پیش‌بینی DTI را با دقتی فراتر از حداقل مقالات اخیر انجام دهد.

خروجی‌های پروژه:

- مدل‌های آموزش‌دیده (RF، XGBoost، GCN، GAT)
- مقایسه کامل با مقالات علمی
- آماده‌سازی ویژگی‌ها و داده‌های پردازش‌شده

نوآوری‌ها و ایده‌های خلاقانه پروژه

۱. تلفیق ویژگی‌های شیمیایی و زیستی

- از SMILES داروها، ۸ ویژگی شیمیایی با استفاده از کتابخانه RDKit استخراج شد.
- از توالی پروتئین‌ها، ۲۰ ویژگی **Amino Acid Composition** استخراج شد.
- این دو نوع ویژگی برای هر نود در گراف تلفیق و به عنوان **Node Embedding** استفاده شد.
- این موضوع باعث پوشش هر دو جنبه شیمیایی–زیستی تعامل دارو–پروتئین شد که در بسیاری از کارهای قبلی ساده‌تر در نظر گرفته می‌شد.

۲. ایجاد نمونه‌های منفی مصنوعی

- برای آموزش بهتر مدل، داده‌هایی با برچسب عدم تعامل به صورت مصنوعی ایجاد شد (Negative Sampling).
- استفاده از روش‌های مصنوعی باعث جلوگیری از **Bias** ناشی از فقدان تعادل نمونه‌های مثبت–منفی شد.

۳. متعادل‌سازی داده‌ها با SMOTE

- به جای حذف داده‌های اضافی یا Oversampling ساده، از **Synthetic Minority Oversampling Technique** استفاده شد.
- این روش نمونه‌های اقلیت را به صورت مصنوعی ولی غیریکسان ایجاد کرد و کیفیت تعادل داده را بالا برد.

۴. ارزیابی چند سناریو PCA برای کاهش بعد

- اجرای سه سناریو:
  - بدون PCA
  - PCA با حفظ ۹۵٪ واریانس
  - PCA با ۱۰ مؤلفه

- مقایسه نتایج این سه حالت و تحلیل تأثیر کاهش بعد بر عملکرد مدل‌ها.

۵. ترکیب مدل‌های کلاسیک و مدل‌های گراف

- پیاده‌سازی و مقایسه بین مدل‌های **RandomForest, XGBoost** و مدل‌های گراف مثل **SimpleGCN, SimpleGAT**.
- بررسی اختلاف عملکرد بین این دو خانواده مدل و مقایسه با مقالات علمی.

۶. جلوگیری از Data Leakage در ساخت گراف

- بخش ساخت گراف به گونه‌ای طراحی شد که داده‌های تست وارد ساختار گراف تمرین نشود.
- این کار مانع از نشت اطلاعات و بهبود مصنوعی نتایج مدل شد.

دیتاست دوم :

هدف پروژه

هدف این پروژه توسعه یک **سیستم یادگیری ماشین** برای پیش‌بینی تعامل بین مولکول‌های شیمیایی و پروتئین‌هاست.

این پروژه بر مبنای داده‌های **SMILES داروها** و **توالی پروتئین‌ها** کار می‌کند و با استخراج ویژگی از هر دو بخش، مدل‌هایی مانند **Logistic Regression، Random Forest، MLP و Node2Vec** را برای پیش‌بینی برچسب تعامل (0 یا 1) آموزش می‌دهد.

داده‌ها و ورودی

- فایل CSV شامل ستون‌های الزامی:
  - `SMILES` → نمایش متنی ساختار مولکول
  - `Protein` → توالی آمینواسیدی پروتئین
  - `Y` → برچسب تعامل (1 = تعامل وجود دارد، 0 = ندارد)

پیش‌پردازش اولیه:

- حذف ردیف‌های ناقص (`dropna`)
- اطمینان از نوع داده صحیح (`int` برای `Y`)
- گزارش توازن کلاس‌ها
- استفاده از `Stratified Train/Test Split` برای تقسیم داده‌ها (پیش‌فرض: 80% آموزش، 20% تست)

استخراج ویژگی‌ها (Feature Engineering)

A. ویژگی‌های دارویی (Drug Features)

- **SMILES N-gram Hashing**:
  - حذف کاراکترهای غیرمجاز و استخراج n-gram در اندازه‌های `2` و `3`
  - هش کردن هر n-gram به فضای `dim=4096`
  - نرمال‌سازی فراوانی توالی‌ها
- خروجی: ماتریس Sparse با ابعاد `(n_samples, 4096)`

B. ویژگی‌های پروتئینی (Protein Features)

- **Amino Acid Composition (AAC)**: فراوانی 20 آمینواسید اصلی (نرمال‌سازی‌شده)

- **K-mer Hash Features (k=3)**: مشابه روش SMILES، هش‌کردن توالی‌های سه حرفی آمینواسیدی به فضای `dim=4096`
- خروجی: ماتریس Sparse شامل `(AAC 20-Features + K-mer 4096-Features)`

C. ادغام ویژگی‌ها

- ماتریس نهایی ویژگی پایه: **[Drug Features + Protein Features]**
- فرمت Sparse و ابعاد کلی: `((n_samples, 4096 + (20 + 4096)`

تقسیم داده‌ها (Splitting Modes)

پروژه سه حالت برای تقسیم داده در نظر گرفته:

1. **Random Split (پیش‌فرض)** → حفظ توزیع برچسب‌ها
2. **Cold Drug Split** → تست روی داروهای جدید که در آموزش نبوده‌اند
3. **Cold Target Split** → تست روی پروتئین‌های جدید که در آموزش نبوده‌اند

این حالت‌ها با GroupShuffleSplit پیاده‌سازی شده‌اند.

کاهش ابعاد (Dimensionality Reduction)

- از **TruncatedSVD** برای کاهش Sparse Matrix به فضای 300 بُعدی استفاده شده
- نمایش واریانس توضیح داده‌شده با نمودار
- کاربرد:
  - سرعت بخشیدن به آموزش
  - کاهش نویز

ویژگی‌های گرافی (Graph-based Features)

- ساخت گراف دو بخشی (Bipartite Graph) از زوج‌های مثبت (Y=1)
- نودها = داروها و پروتئین‌ها، یال‌ها = تعامل
- استفاده از **Node2Vec** و **SVD روی ماتریس مجاورت** برای ساخت embedding
- تبدیل embeddingها به ویژگی‌های آموزش با ترکیب:
  - دارو
  - پروتئین
  - قدر مطلق اختلاف
  - ضرب مؤلفه‌ای (Elementwise Product)

مدل‌ها و یادگیری ماشین

مدل‌های تست‌شده برای سه دسته ویژگی:

- **Baseline Features**
- **Graph-SVD Features**
- **Fusion (Baseline + Graph)**
- **Node2Vec Features**

مدل‌ها:

- `LogisticRegression` (solver="saga", max_iter=5000)
- `RandomForestClassifier` (n_estimators=300-350)
- `MLPClassifier` (دو لایه مخفی: 512 و 256 نورون، ReLU)

ارزیابی

- معیارها:
  - **ROC-AUC**
  - **PR-AUC**
  - **F1-score**
  - **Accuracy**
- انتخاب آستانه بهینه با `precision_recall_curve`
- چاپ `classification_report` و ماتریس درهم‌ریختگی (Confusion Matrix)

نتایج خلاصه

- بهترین عملکرد کلی روی **Fusion Features** با Logistic Regression و Random Forest
- وجود بهبود PR-AUC و ROC-AUC نسبت به استفاده از ویژگی‌های تک‌منبعی
- Node2Vec به تنهایی عملکرد متوسطی داشته ولی در حالت ترکیبی کمک کرده

تحلیل اهمیت ویژگی‌ها (Permutation Feature Importance)

- محاسبه اهمیت ویژگی‌ها روی تست‌ست
- نمایش **Top-30 Features**

- رنگ‌ها: BASE (ویژگی‌های اولیه)، BASE_RED (ویژگی‌های SVDشده)، GRAPH (ویژگی‌های گرافی)

پیشنهادات توسعه آینده

1. استفاده از embeddingهای عمیق‌تر:
   - Mol2Vec / ChemBERTa برای داروها
   - ProtBERT / ESM برای پروتئین‌ها
2. تست روی دیتاست‌های بزرگتر مثل **BindingDB** یا **Davis**
3. آزمایش مدل‌های هیبریدی (ML + GNN)
4. استفاده از Multi-Modal Transformers برای ترکیب اطلاعات SMILES + توالی
5. پیاده‌سازی Cross-validation برای تضمین پایداری نتایج