AI Driven Analysis Of Drug Side Effect Using DNA Profiles

predict possible side effects of a drug for a person **using their genetic information**

Data Sources Used

Dataset	What it contains	Why we use it
1000 Genomes	Human DNA (SNP's)	To get genetic variation
PharmGKB	Gene–Drug relationships	To know which genes affect which drugs
SIDER	Drug ID + Side effect	To know side effects of drugs
PubChem Web	Drug identifiers (CID → name)	To get drug name using ID

STEP 1: Raw DNA Data (VCF files)

Sources : https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/

STEP 2: Convert VCF → CSV

VCF files are very large and hard to use directly.

So we Convert them into CSV format

	A	B	C	D	E	F
1	Person	SNP_6_63979	SNP_6_63980	SNP_6_89506	SNP_6_100112	SNP_6_100114
2	1	0	0	0	0	0
3	2	0	0	0	0	0
4	3	0	0	0	0	0
5	4	0	0	0	0	0
6	5	0	0	0	0	0
7	6	0	0	0	0	0
8	7	0	0	0	0	0
9	8	0	0	0	0	0
10	9	0	0	0	0	0
11	10	0	0	0	0	0
12	11	0	0	0	0	0
13	12	0	0	0	0	0
14	13	1	1	0	0	0
15	14	0	0	0	0	0
16	15	2	2	0	0	0

Here we already encode SNP's

0|0 = 0 , Normal

0|1, 1|0 = 1, Normal + Chnaged = One Changed

1|1 = 2 , Changed + Changed = Full changed

Exa :

Chromosome	SNP_Position	Ref	ALT
6	63979	CAG	C

At chrom 6, Position **63979**

Most people have CAG , but here this person is missing AG

- DNA is long text -> A T C G
- Sometimes one letter is different
  exa = ATCG -> ATCC -> 1 letter change SNP
  SNP = Single nucleotide polymorephism

STEP 4: Filter SNPs

We remove:

- SNPs that are same for everyone
- This keeps only **important genetic variations**

STEP 5: SNP → Gene Mapping and Genetic Score Calculation

Sources: genecode annotation
https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz

	A	B	C	D
1	gene	chromosome	start	end
2	RP1-24O22.5	chr6	95124	95454
3	OR4F1P	chr6	105919	106856
4	LINC00266-3	chr6	131910	144885
5	CICP18	chr6	142272	145083
6	RP3-416J7.1	chr6	167607	170631
7	RP3-416J7.4	chr6	181466	205484
8	RP3-416J7.5	chr6	203313	206392
9	DUSP22	chr6	291630	351355
10	IRF4	chr6	391739	411447
11	EXOC2	chr6	485133	693117

Exa.

SNP_6_95224 ----> RP1

SNP_6_95454 ----> RP1

SNP_6_181466 ----> RP3

For person

RP1 genes = SNP_6_95224 + SNP_6_95454 = 0+1 = 1

RP3 genes = SNP_6_181466 = 2

Does 95224 fall inside any gene range

Yes = assign SNP to that gene

No = ignore SNP

After this we get :

	A	B	C	D	E	F	G	H	I
1	Person	LINC00266-3	CICP18	RP3-416J7.1	RP3-416J7.4	RP3-416J7.5	DUSP22	IRF4	EXOC2
2	1	0	0	0	2	0	152	50	576
3	2	0	0	0	3	1	129	34	519
4	3	0	0	0	9	4	137	39	510
5	4	0	0	0	2	4	145	37	420
6	5	0	0	0	4	3	138	40	492
7	6	0	0	0	4	4	147	46	554
8	7	0	0	0	4	2	130	42	540
9	8	0	0	0	4	0	137	47	579
10	9	0	0	0	2	3	135	48	591
11	10	0	0	0	9	7	144	47	526
12	11	0	0	0	13	12	131	30	537
13	12	0	0	0	3	1	131	40	571
14	13	0	0	0	4	6	132	60	397
15	14	0	0	0	10	3	121	34	534
16	15	0	0	0	2	0	133	62	512

STEP 6

Gene → Drug Mapping and Drug-Level Genetic Score Calculation

	A	B
1	gene	drug
2	NQO1	antiinflammatory and antirheumatic products, non-steroids
3	NQO1	imatinib
4	NQO1	corticosteroids
5	NQO1	fluorouracil
6	NQO1	anthracyclines and related substances
7	NQO1	oxaliplatin
8	NQO1	antiepileptics
9	NQO1	Analgesics and anesthetics
10	NQO1	Antibiotics
11	NQO1	Antifungals For Systemic Use
12	NQO1	Antimycobacterials
13	NQO1	Antithyroid Preparations
14	NQO1	Drugs For Treatment Of Tuberculosis
15	NQO1	Platinum compounds

This PharmGKb data is used

Suppose person 1

- CYP2C9 score = 3
- VKORC1 score = 0

Warfarin depends on:

- CYP2C9
- VKORC1

Genetic score = 3 + 0 = 3

After this We get

	A	B	C	D
1	person_id	drug_name	side_effect	genetic_score
2	1	doxorubicin	Gastrointestinal pain	118
3	1	doxorubicin	Abdominal pain	118
4	1	doxorubicin	Abscess	118
5	1	doxorubicin	Agranulocytosis	118
6	1	doxorubicin	Albuminuria	118
7	1	doxorubicin	Alopecia	118
8	1	doxorubicin	Amblyopia	118
9	1	doxorubicin	Amenorrhoea	118
10	1	doxorubicin	Anaphylactic shock	118
11	1	doxorubicin	Anaemia	118
12	1	doxorubicin	Hypochromic anaemia	118
13	1	doxorubicin	Angina pectoris	118
14	1	doxorubicin	Decreased appetite	118
15	1	doxorubicin	Anxiety	118

STEP 7: ADR Label

To train a machine learning model, we need a **target label**.

- People with **higher genetic scores** have **higher risk of ADRs**.

Rule used for ADR label creation

For each drug separately:

1. Collect all genetic scores for that drug
2. Find the **75th percentile (top 25%)** of genetic scores
3. Assign labels as:

Condition	ADR Label
Genetic score ≥ threshold	1 (High risk)
Genetic score < threshold	0 (Low risk)

Genetic scores for Warfarin

person_id	genetic_score
P1	5
P2	2
P3	4
P4	1

1, 2, 4, 5

75th percentile ≈ **4**

After this We get

	A	B	C	D	E
1	person_id	drug_name	side_effect	genetic_score	adr_label
2	1	doxorubicin	Gastrointestinal pain	118	1
3	1	doxorubicin	Abdominal pain	118	1
4	1	doxorubicin	Abscess	118	1
5	1	doxorubicin	Agranulocytosis	118	1
6	1	doxorubicin	Albuminuria	118	1
7	1	doxorubicin	Alopecia	118	1
8	1	doxorubicin	Amblyopia	118	1
9	1	doxorubicin	Amenorrhoea	118	1
10	1	doxorubicin	Anaphylactic shock	118	1
11	1	doxorubicin	Anaemia	118	1
12	1	doxorubicin	Hypochromic anaemia	118	1
13	1	doxorubicin	Angina pectoris	118	1
14	1	doxorubicin	Decreased appetite	118	1
15	1	doxorubicin	Anxiety	118	1
16	1	doxorubicin	Arrhythmia	118	1
17	1	doxorubicin	Arthralgia	118	1
18	1	doxorubicin	Musculoskeletal discomfort	118	1
19	1	doxorubicin	Ascites	118	1

Saving the final dataset and freezing it for ML