CSC 116

How to build AI models using medical datasets?

Features

57 features:

COHORT – Group or study population identifier.
age – Age of the participant.
fampd – Family history of Parkinson's disease (yes/no).
age_datscan – Age when DAT scan was performed (dopamine transporter scan).
👃 **Olfactory (Smell) Tests**
upsit – University of Pennsylvania Smell Identification Test score.
upsit_pctl – Percentile rank of UPSIT score.
upsit_pctl15 – Whether UPSIT is ≤15th percentile (yes/no).
🧠 **Cognitive Assessments**
moca – Montreal Cognitive Assessment (general cognitive screening).
bjlot – Benton Judgment of Line Orientation Test (spatial judgment).
DVS_JLO_MSSA – Digitally scored version of JLO (accuracy score).
DVS_JLO_MSSAE – JLO efficiency score (accuracy per time).
⏰ **Memory Tests (HVLT, DVT)**
clockdraw – Clock Drawing Test (visuospatial and executive function).
hvlt_discrimination – Hopkins Verbal Learning Test (discrimination index).
hvlt_immediaterecall – HVLT immediate recall score.
hvlt_retention – HVLT retention score.
HVLTFPRL – HVLT false positive recognition.
HVLTRDLY – HVLT delayed recall.
HVLTREC – HVLT recognition score.
DVT_TOTAL_RECALL – Digital Verbal Test (total recall).
DVT_DELAYED_RECALL – DVT delayed recall.
DVT_RETENTION – DVT retention rate.
DVT_RECOG_DISC_INDEX – DVT recognition discrimination index.
💬 **Verbal Fluency & Language**
lexical – Lexical fluency (word generation).
DVT_FAS – FAS test (letters-based fluency).
DVS_FAS – Digital version of FAS test.
lns – Letter-Number Sequencing (working memory).
DVS_LNS – Digital version of LNS.
MODBNT – Modified Boston Naming Test (object naming).
DVS_BNT – Digital BNT.
PCTL_BNT – BNT percentile.
🚀 **Processing Speed & Executive Function**
SDMTOTAL – Symbol Digit Modalities Test (processing speed).
DVT_SDM – Digital SDM score.
DVSD_SDM – Digital SDM duration.
TMT_A – Trail Making Test Part A (attention).
TMT_B – Trail Making Test Part B (executive function).
DVZ_TMTA – Digital TMT-A z-score.
DVZ_TMTB – Digital TMT-B z-score.
🐶 **Semantic Fluency**
VLTANIM – Animal fluency (number of animals named).
DVT_SFTANIM – Digital semantic fluency test for animals.
DVS_SFTANIM – Digital semantic fluency efficiency.
⚠️ **Cognitive Diagnosis**
MCI_testscores – Mild Cognitive Impairment (MCI) diagnosis from test scores.
cogstate – Cognitive status category.
🏠 **Daily Living**
MSEADLG – Modified Schwab and England Activities of Daily Living scale.
🏃 **Behavioral Symptoms**
quip_any – Presence of any impulse control disorder.
quip_walk – Walking-based impulse control issues.
💤 **Sleep**
ess – Epworth Sleepiness Scale (daytime sleepiness).
rem – REM sleep behavior disorder status.
🙂 **Mood**
gds – Geriatric Depression Scale.
stai – State-Trait Anxiety Inventory total score.
stai_state – STAI state anxiety (current).
stai_trait – STAI trait anxiety (general tendency).
🩺 **Other Clinical Measures**
orthostasis – Presence of orthostatic hypotension.
NP1DPRS – Depression rating from MDS-UPDRS Part I.
🧪 **Biomarkers**
abeta – Amyloid-beta levels.
tau – Tau protein levels.
ptau – Phosphorylated tau levels.
urate – Uric acid level (sometimes linked to neuroprotection).

2	100889	1	Sporadic PD	
2	100890	2	Healthy Control	
2	100890	2	Healthy Control	
2	100890	2	Healthy Control	
2	100890	2	Healthy Control	
3	100891	1	Sporadic PD	

Label: Parkinson or Healthy

Parkinson Feature 1, 2, 3,4,5,6,7,8,9…

Healthy Feature 1, 2, 3,4,5,6,7,8,9…

We l uses the features to train the labels.
Put all the features in the trained model, the model will tell you based-on these features, the patient is likely parkinson or Healthy

4000+ patient records with 57 features

Will you use all the datasets for training??

3200 for training

800 for evaluations

How to find your datasets?

Find your datasets in Kaggle, Hugging face or Github, or others.

GPU version – NVIDIA

Check the memory of your GPU.

4090 4080 GPUs need more $3000. It can be used for train very good models except very LLMs.

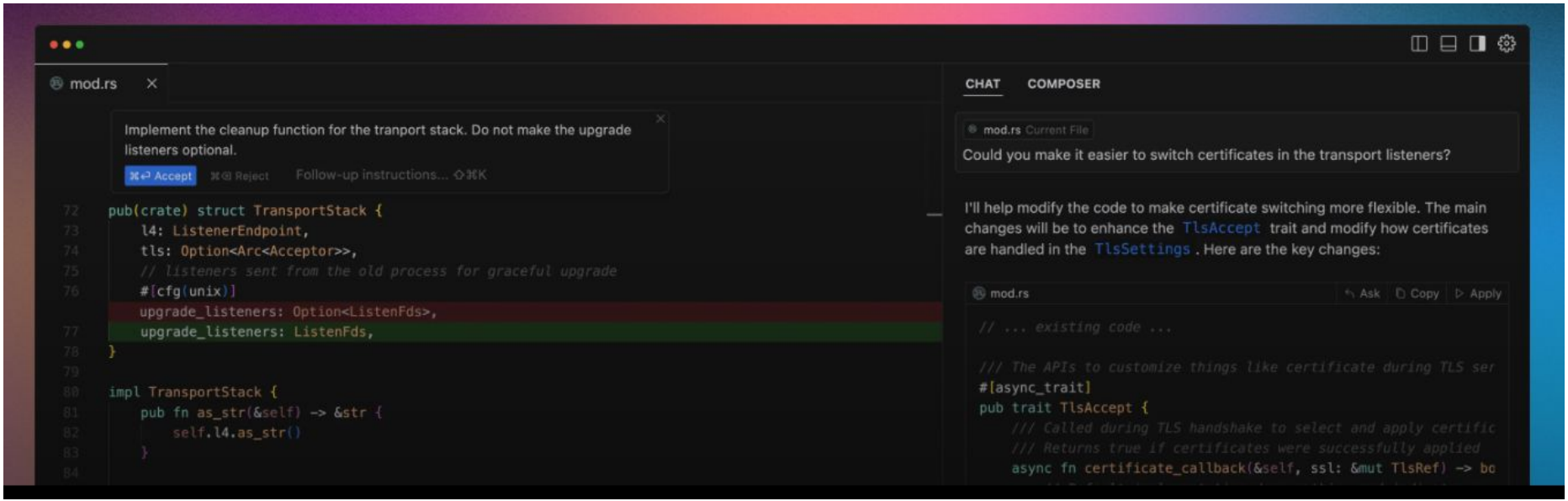
Accelerator	Memory	Memory Type	PyTorch Support	Ideal For	Notes
RTX 4080	16 GB	GDDR6X	✅ Yes	Medium-sized models, inference	Great for developers, high value
RTX 4090	24 GB	GDDR6X	✅ Yes	Large models, high-end training	Very powerful consumer GPU
Tesla P100	16 GB	HBM2	✅ Yes	Traditional model training	Older, but still used in data centers
T4 ×2	2 × 16 GB	GDDR6	✅ Yes	Inference, light training	Energy-efficient, not great for big training
A100 40GB	40 GB	HBM2	✅ Yes	Large-scale training	Data center GPU, expensive
A100 80GB	80 GB	HBM2e	✅ Yes	Extra-large model training (e.g. GPT-3)	Very high memory, powerful
TPU v3-8	8 × 16 GB = 128 GB	HBM	❌ No (only TensorFlow/JAX)	TensorFlow/JAX, massive model training	Requires special setup, not for beginners

GPU platform for training

Kaggle

Google Colab

Cursor — for local training



Datasets

https://github.com/elastic/ember?utm_source=chatgpt.com

Model **test accuracy** is high in the test datasets.

But it is pretty low in the real cases. What should we do?

Fine-tuning

"Fine-tuning" refers to a **transfer learning** technique where a pre-trained model is further trained on a new, specific dataset to improve its performance on a particular task, rather than training a model from scratch.

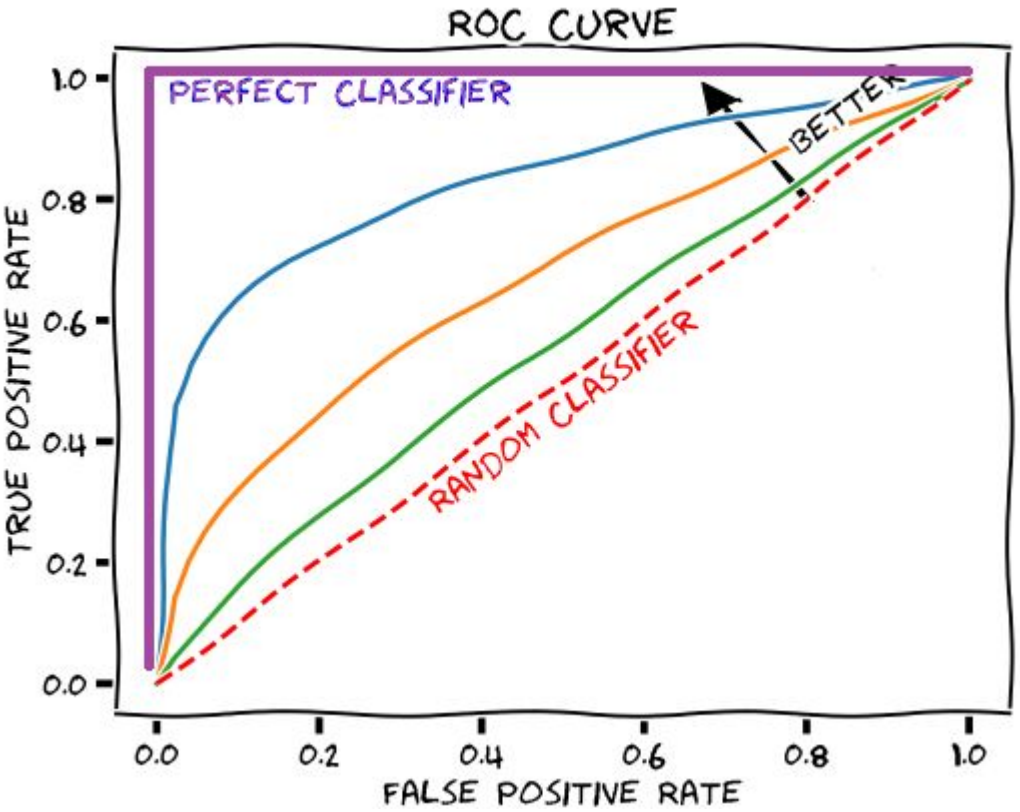
Training Accuracy:

Definition: The percentage of correct predictions made by the model on the **training dataset**, which is the data used to train the model.
Purpose: It tells you how well the model has **learned** from the data it was trained on.

Testing Accuracy:

Definition: The percentage of correct predictions made by the model on the **testing dataset**, which is separate from the training data.
Purpose: It shows how well the model **generalizes** to **new, unseen data**.

AUC/ROC



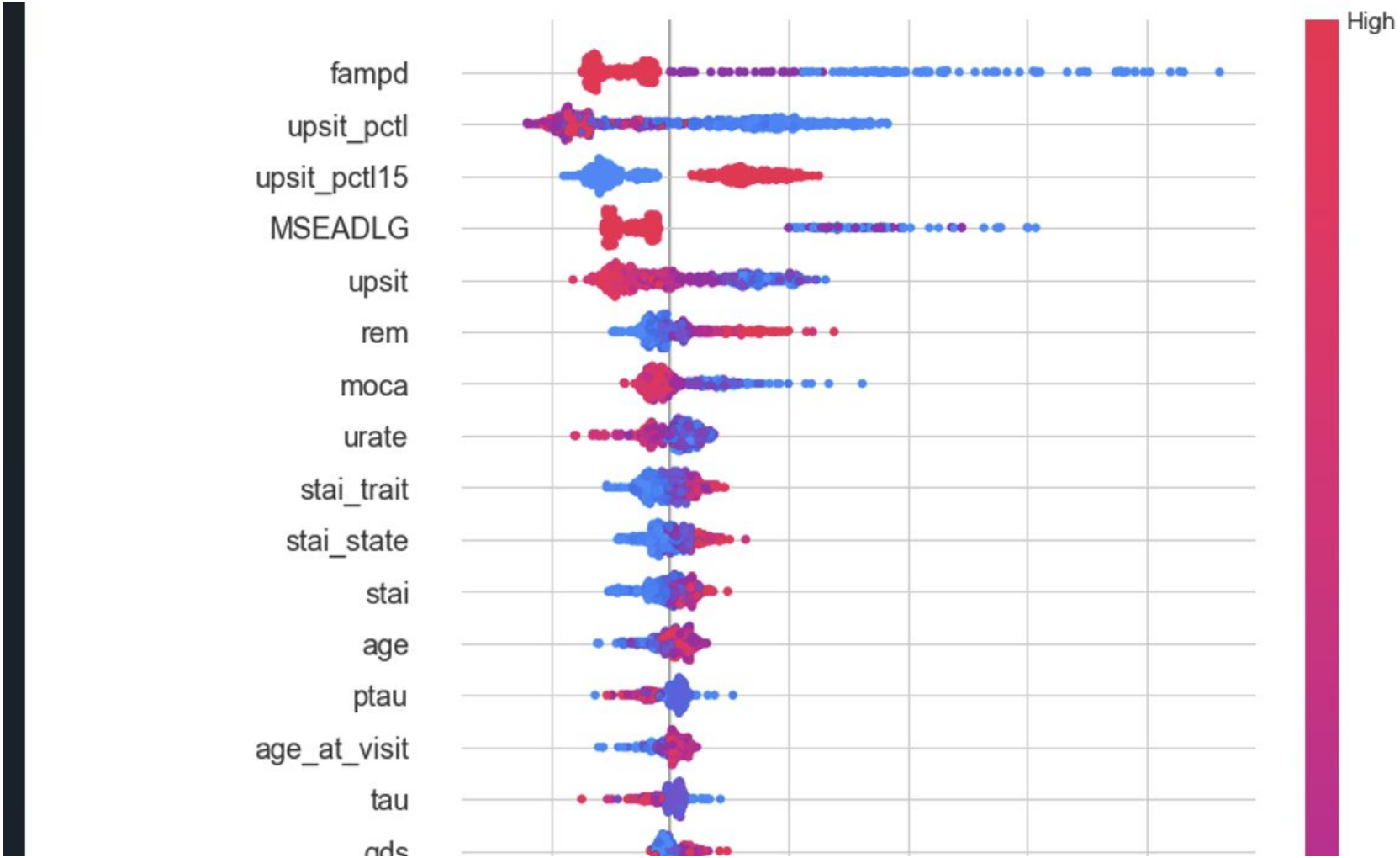
AUC, or Area Under the Curve, is a metric used in machine learning, particularly for evaluating the performance of binary classification models.

How good the model it is?

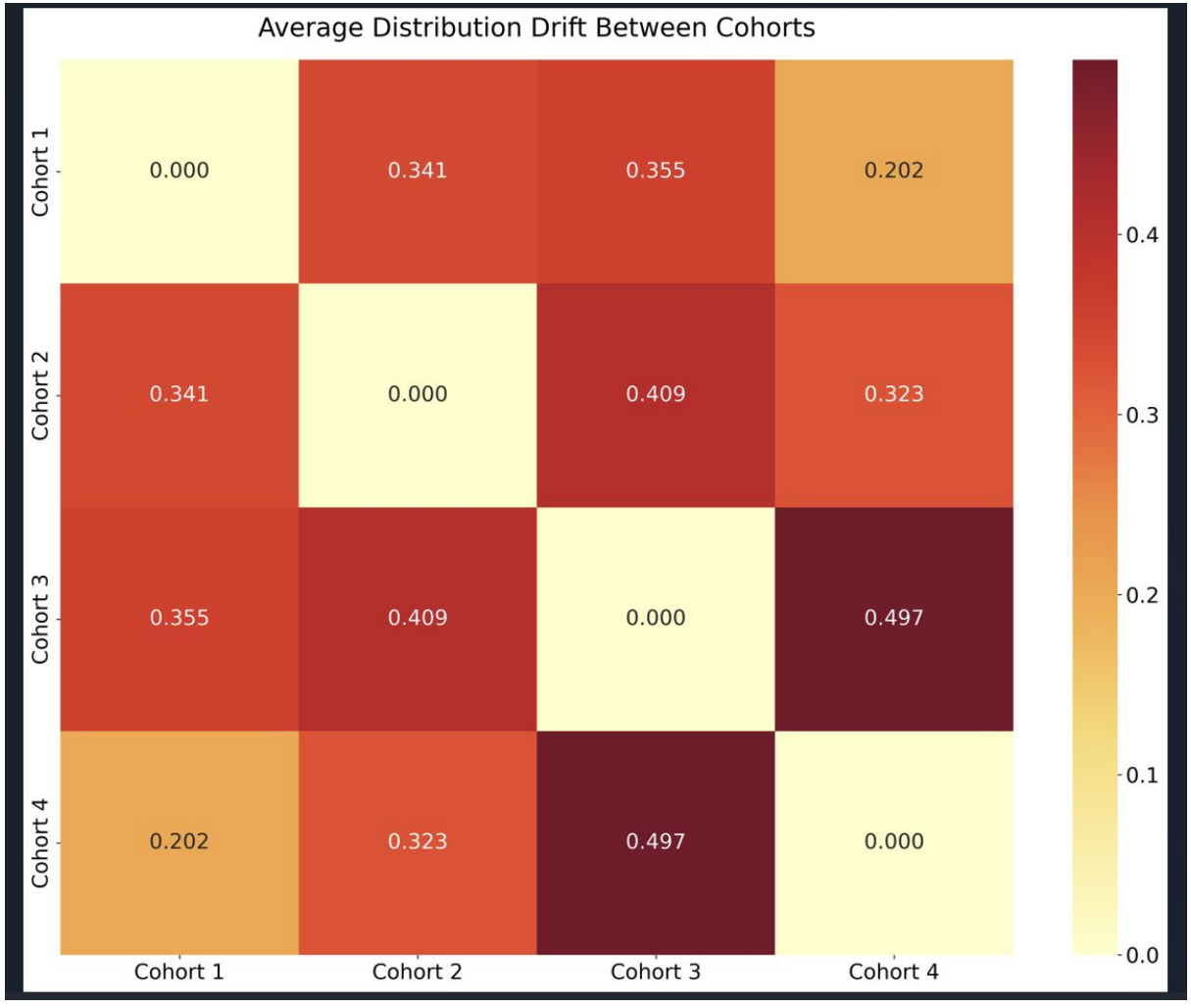
- **AUC = 1.0**: Perfect classifier
- **AUC = 0.5**: No discriminative power (equivalent to random guessing)
- **AUC < 0.5**: Worse than random (model is misclassifying)

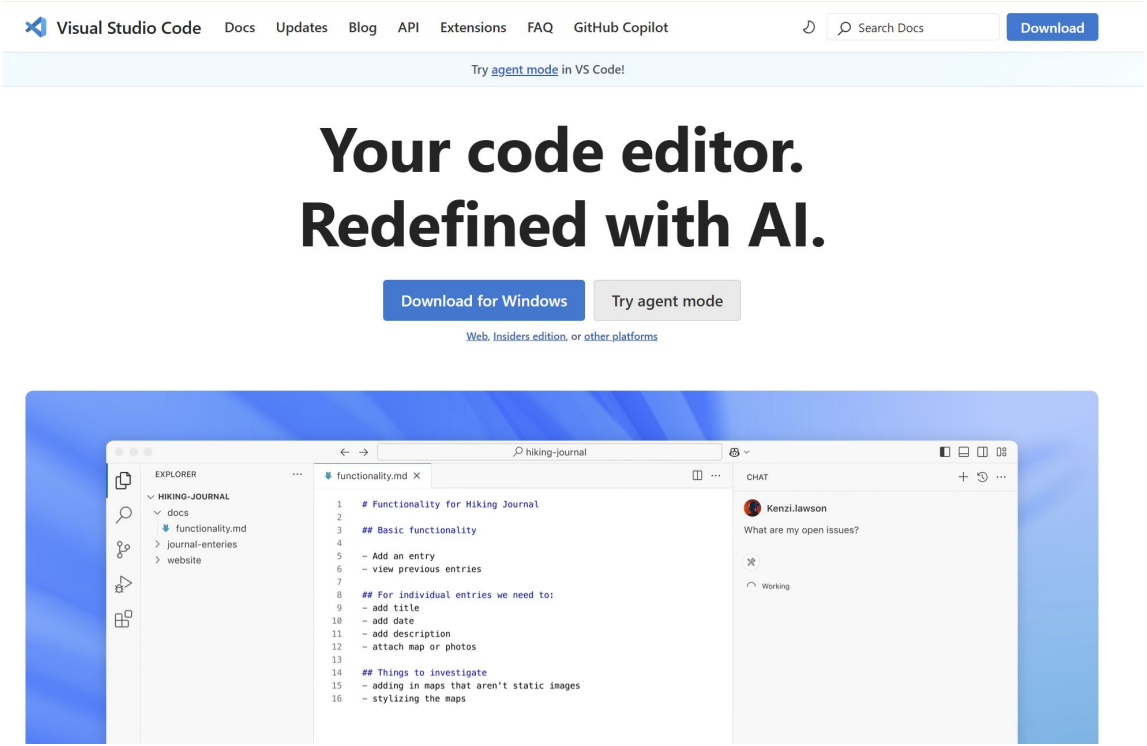
Feature Importances

Explainable AI solutions:
SHAP


fampd
upsit_pctl
upsit_pctl15
MSEADLG
upsit
rem
moca
urate
stai_trait
stai_state
stai
age
ptau
age_at_visit
tau
High

Drift Detection Solution


Average Distribution Drift Between Cohorts
Cohort 1
Cohort 2
Cohort 3
Cohort 4
0.000
0.341
0.355
0.202
0.341
0.000
0.409
0.323
0.355
0.409
0.000
0.497
0.202
0.323
0.497
0.000
Cohort 1
Cohort 2
Cohort 3
Cohort 4
0.4
0.3
0.2
0.1
0.0



https://code.visualstudio.com/

Kaggle

Lung cancer：

https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction

Thanks