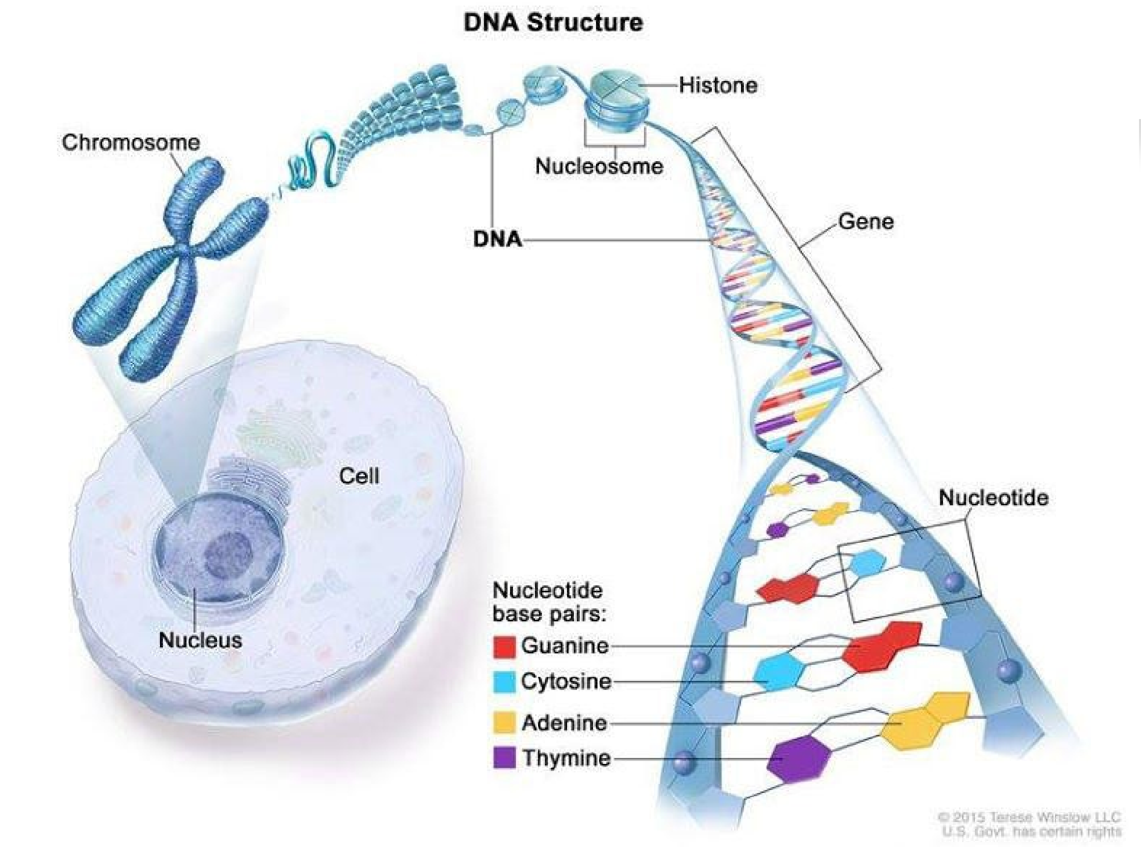
# AI for Precision Oncology

# Project Aim

- The aim of my project is to create a **Machine Learning (ML) model** that can **predict** a patient's Immunotherapy **response** using transcriptomic data.

# Cancer

What is cancer?



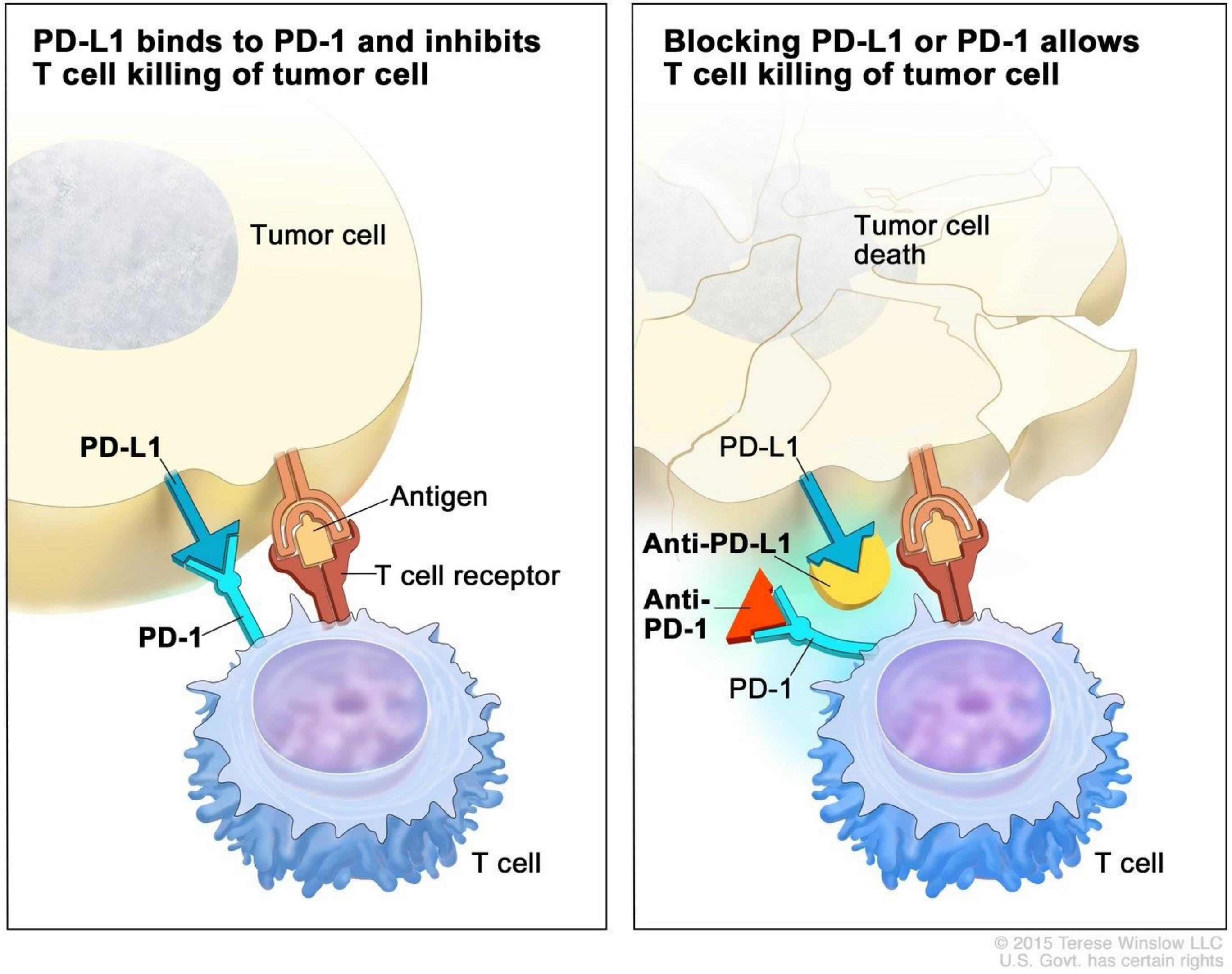
© 2015 Terese Winslow LLC
U.S. Govt. has certain rights

# Cancer Immunotherapy

Immunotherapy leverages the body’s immune system to combat cancer.

Immune Checkpoint Inhibitors

- Anti-PD1
  - Pembrolizumab (Keytruda)
  - Nivolumab (Opdivo)

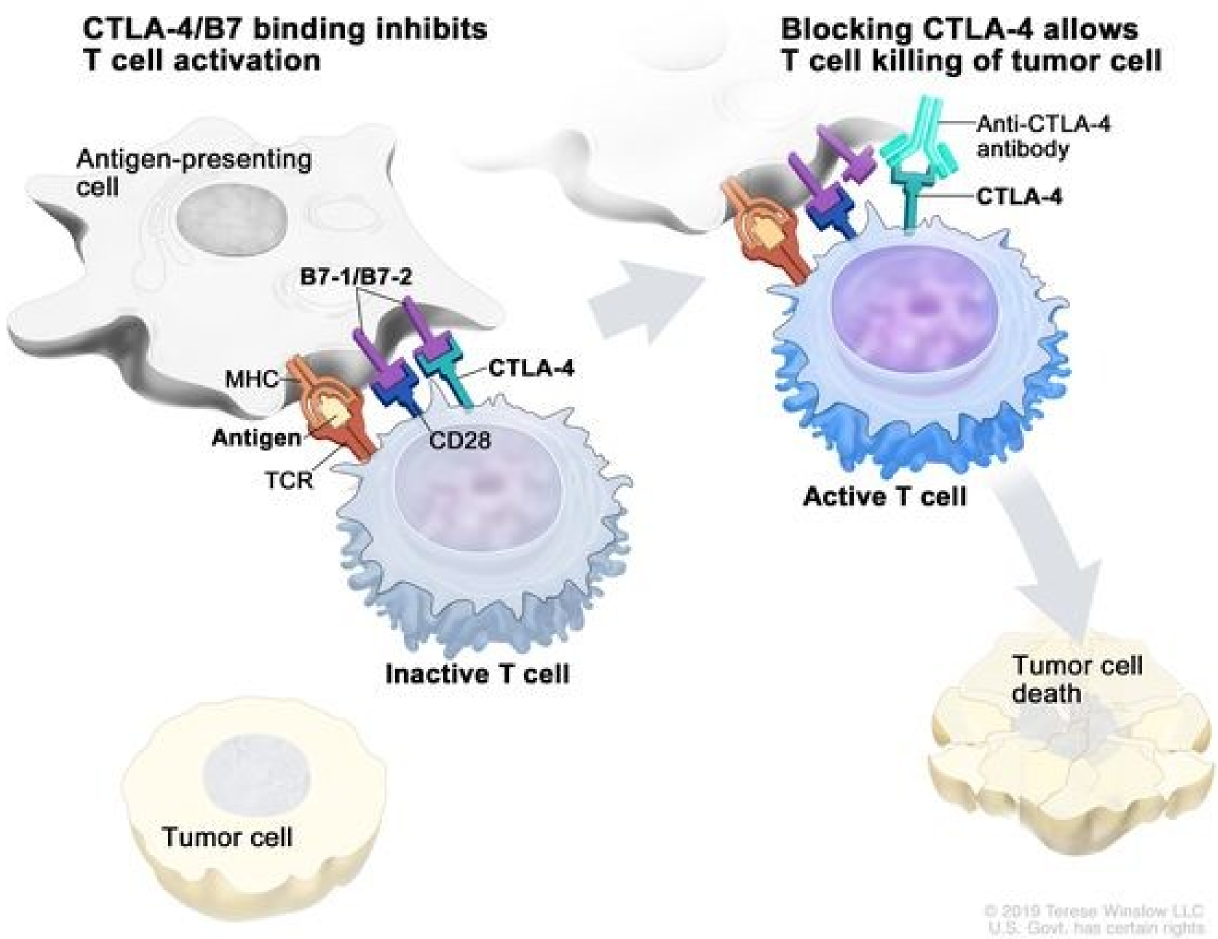

PD-L1 binds to PD-1 and inhibits T cell killing of tumor cell
Tumor cell
PD-L1
Antigen
T cell receptor
PD-1
T cell
Blocking PD-L1 or PD-1 allows T cell killing of tumor cell
Tumor cell death
PD-L1
Anti-PD-L1
Anti-PD-1
PD-1
T cell

© 2015 Terese Winslow LLC
U.S. Govt. has certain rights

[1]

# Cancer Immunotherapy

- Anti-CTLA-4
  - Tremelimumab
  - ipilimumab



© 2019 Terese Winslow LLC
U.S. Govt. has certain rights

[1]

# The problem

20-40% of patients [2]. On average, price patient annually is $150,000 in the USA

Side effects of Immunotherapy can be; diarrhoea, fatigue, nausea or even an autoimmune response.

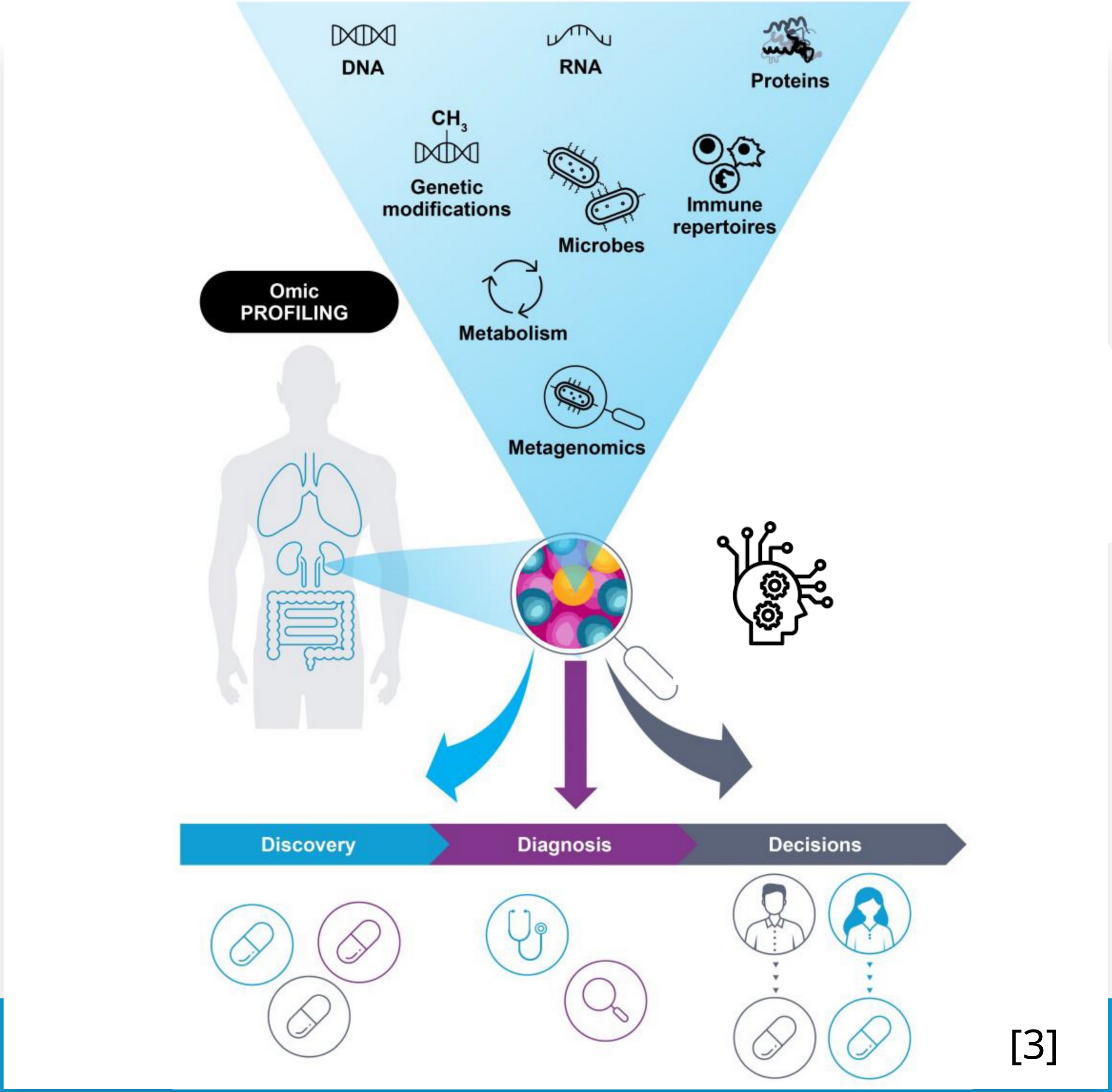
Time waste and disease progression.

# The Current Solution

- **Precision Oncology**: Utilises genetic profiling and biomarkers to customise treatments for individuals.
- **Current Biomarker in Clinical Practice**: Tumour mutational burden, PD-L1 expression, etc.
- **Limitations**:
  - Costly
  - Results different across labs

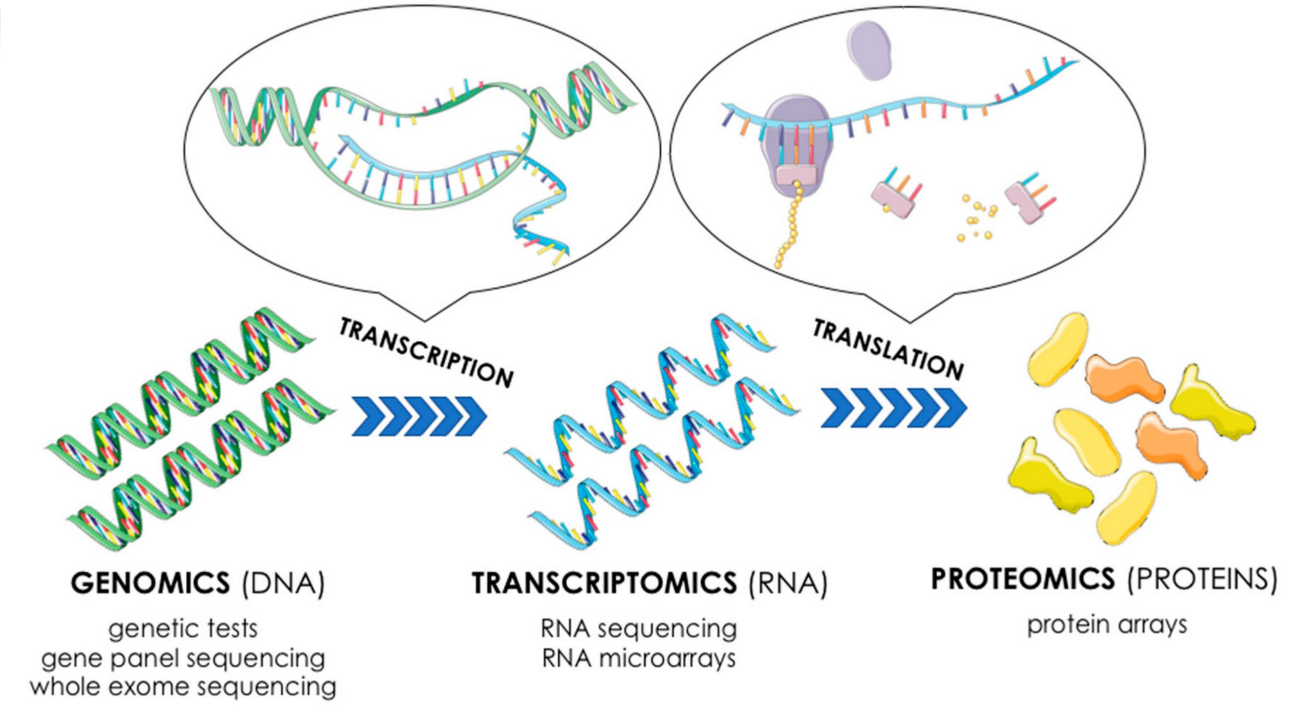
# The Proposed Solution

Combine **ML with precision oncology** and train ML models on pre-treatment gene expression data from next-generation sequencing to **predict patient outcomes**.

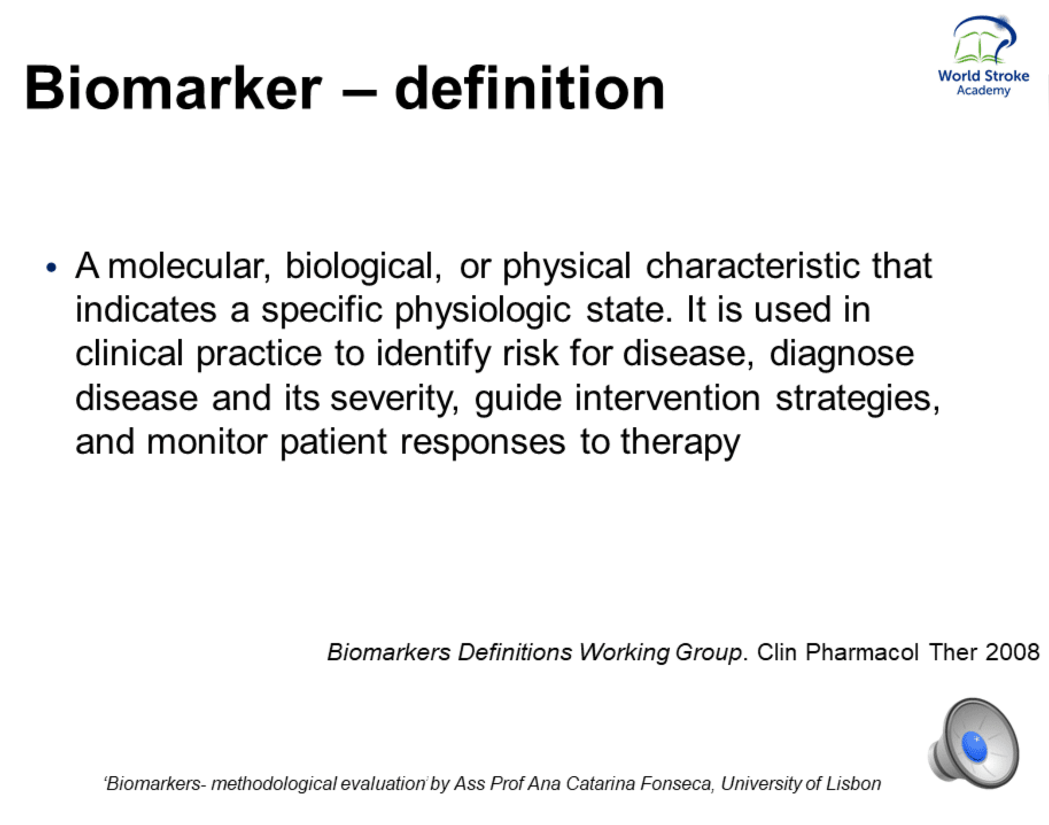


[3]

# Transcriptomics



# Biomarker



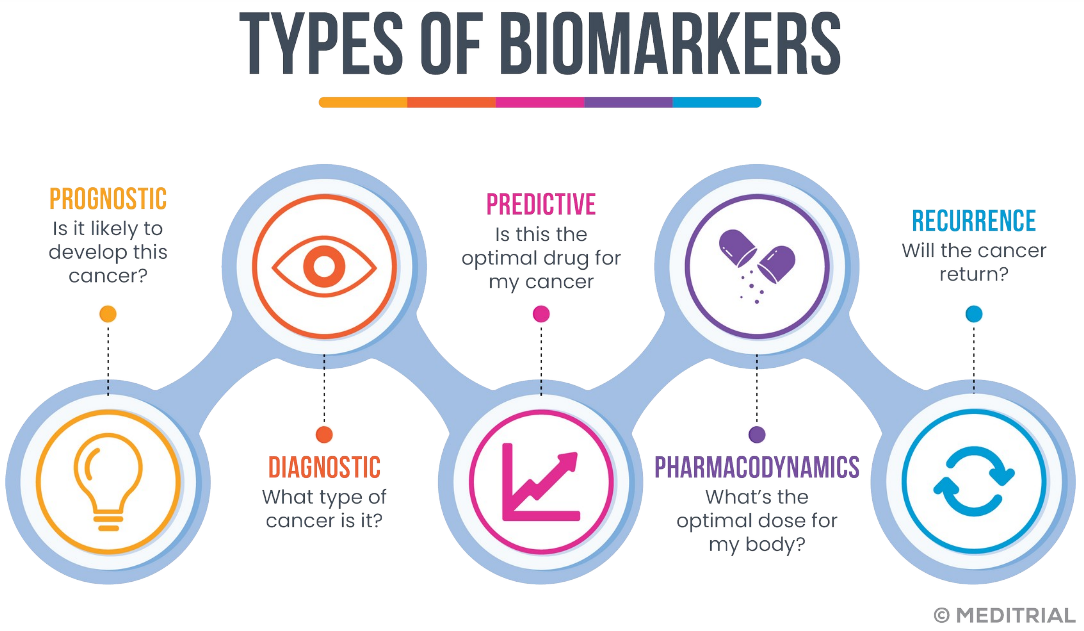
## Biomarker – definition

World Stroke Academy

- A molecular, biological, or physical characteristic that indicates a specific physiologic state. It is used in clinical practice to identify risk for disease, diagnose disease and its severity, guide intervention strategies, and monitor patient responses to therapy

*Biomarkers Definitions Working Group*. Clin Pharmacol Ther 2008

*'Biomarkers- methodological evaluation' by Ass Prof Ana Catarina Fonseca, University of Lisbon*



TYPES OF BIOMARKERS

PROGNOSTIC
Is it likely to develop this cancer?

DIAGNOSTIC
What type of cancer is it?

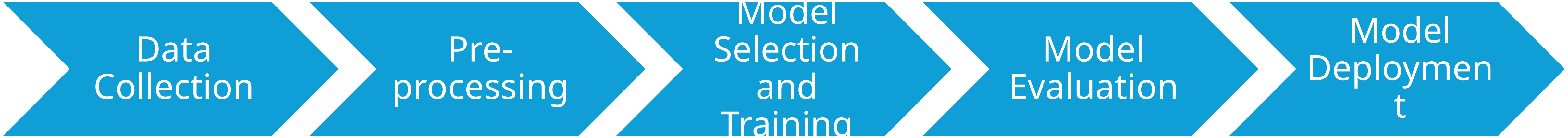
PREDICTIVE
Is this the optimal drug for my cancer

PHARMACODYNAMICS
What's the optimal dose for my body?

RECURRENCE
Will the cancer return?

© MEDITRIAL

# Machine Learning Pipeline



Data Collection

Pre-processing

Model Selection and Training

Model Evaluation

Model Deployment

# Machine Learning Pipeline: Data

- **Transcriptomic** data from pre-treatment tumour biopsies and blood samples.

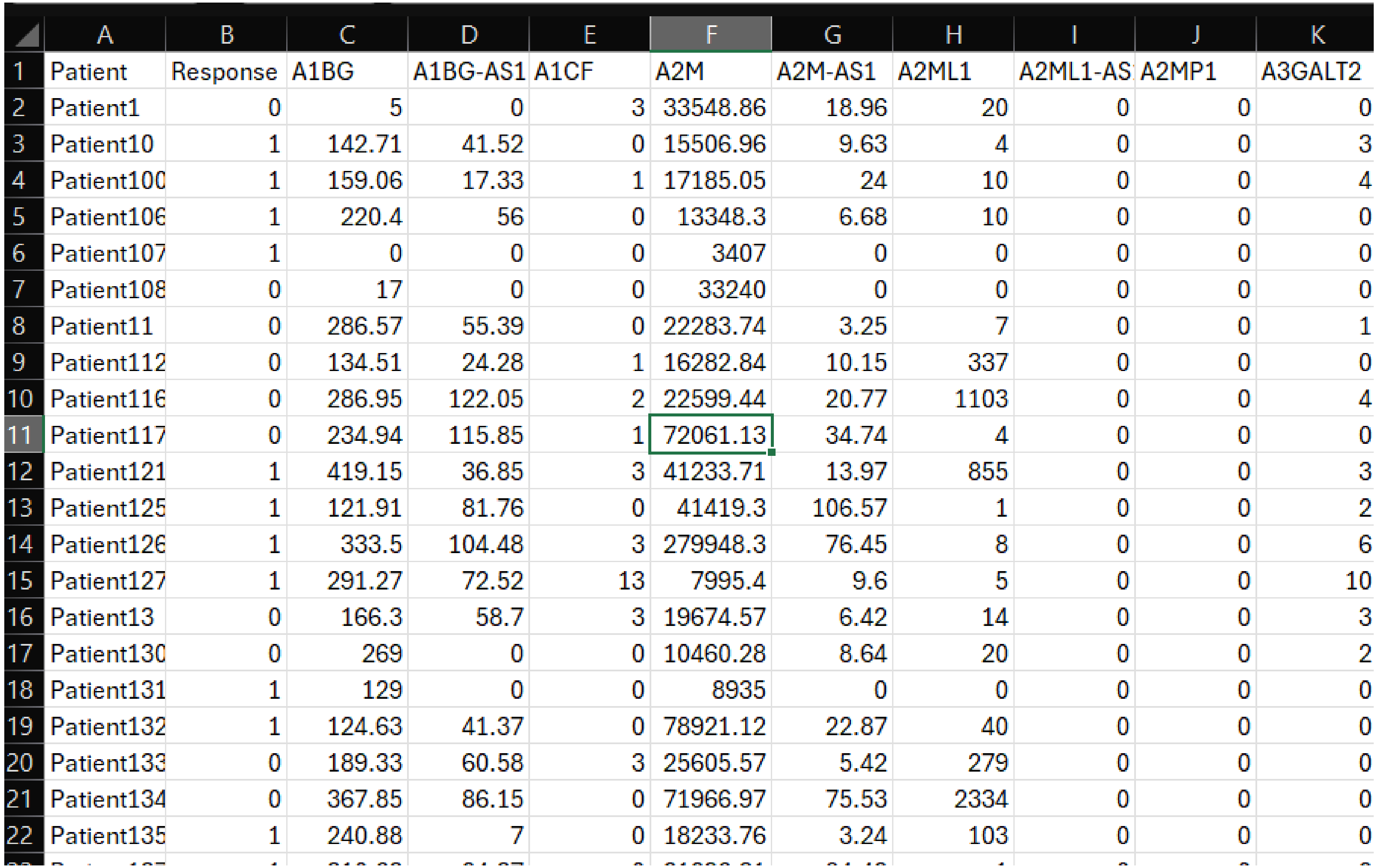
- Collected through **extensive literature reviews**, **public databases**, and **clinical trials.**

- The pre-treatment tumour provides a baseline of the patient's disease before treatment.

# Biopsy Data

| Characteristics | I01 (N = 119) | I14 (N = 56) | I09 (N = 41) | I15 (N = 28) |
|---|---|---|---|---|
| **Studies cohort** | Liu D, 2019 | Riaz N, 2017 | Glide TN, 2019 | Hugo W, 2016 |
| **Data source** | dbGaP (accession number phs000452.v3.p1) | GSE91061 | ENA: PRJEB23709 | GSE78220 |
| **Cancer type** | Melanoma: 121 (100%) | Melanoma: 56 (100%) | Melanoma: 41 (100%) | Melanoma: 28 (100%) |
| **Anti-PD1 received for Melanoma Cancer** | Nivolumab or Pembrolizumab (not identified in each patient) | Nivolumab: 56 (100%) | Nivolumab: 9 (22%)<br>Pembrolizumab: 32 (78%) | Pembrolizumab: 28 (100%) |
| **Number of genes in common** | 7440 | 7440 | 7440 | 7440 |
| **Drug response (RECIST)** | | | | |
| **CR (4)** | 16 (13%) | 3 (5%) | 4 (10%) | 5 (18%) |
| **PR (3)** | 31 (26%) | 8 (14%) | 15 (36%) | 10 (36%) |
| **SD (2)** | 16 (13%) | 19 (34%) | 6 (15%) | 0 (0%) |
| **PD (1)** | 57 (47%) | 26 (46%) | 16 (39%) | 13 (46%) |
| **Drug response** | | | | |
| **Responder** | 47 (39%) | 11 (20%) | 19 (46%) | 15 (54%) |

# Machine Learning Pipeline: Data

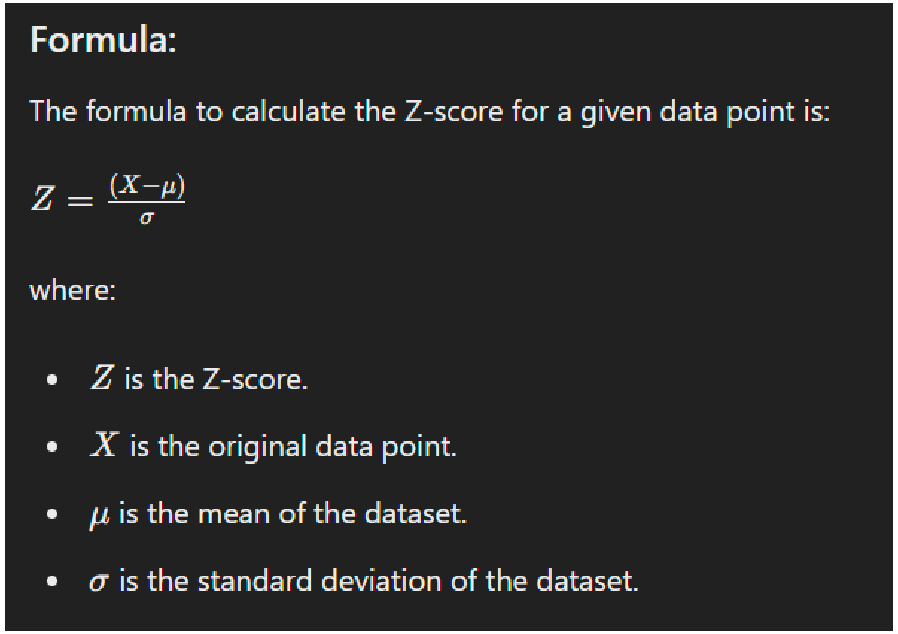


|  | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Patient | Response | A1BG | A1BG-AS1 | A1CF | A2M | A2M-AS1 | A2ML1 | A2ML1-AS | A2MP1 | A3GALT2 |
| 2 | Patient1 | 0 | 5 | 0 | 3 | 33548.86 | 18.96 | 20 | 0 | 0 | 0 |
| 3 | Patient10 | 1 | 142.71 | 41.52 | 0 | 15506.96 | 9.63 | 4 | 0 | 0 | 3 |
| 4 | Patient100 | 1 | 159.06 | 17.33 | 1 | 17185.05 | 24 | 10 | 0 | 0 | 4 |
| 5 | Patient106 | 1 | 220.4 | 56 | 0 | 13348.3 | 6.68 | 10 | 0 | 0 | 0 |
| 6 | Patient107 | 1 | 0 | 0 | 0 | 3407 | 0 | 0 | 0 | 0 | 0 |
| 7 | Patient108 | 0 | 17 | 0 | 0 | 33240 | 0 | 0 | 0 | 0 | 0 |
| 8 | Patient11 | 0 | 286.57 | 55.39 | 0 | 22283.74 | 3.25 | 7 | 0 | 0 | 1 |
| 9 | Patient112 | 0 | 134.51 | 24.28 | 1 | 16282.84 | 10.15 | 337 | 0 | 0 | 0 |
| 10 | Patient116 | 0 | 286.95 | 122.05 | 2 | 22599.44 | 20.77 | 1103 | 0 | 0 | 4 |
| 11 | Patient117 | 0 | 234.94 | 115.85 | 1 | 72061.13 | 34.74 | 4 | 0 | 0 | 0 |
| 12 | Patient121 | 1 | 419.15 | 36.85 | 3 | 41233.71 | 13.97 | 855 | 0 | 0 | 3 |
| 13 | Patient125 | 1 | 121.91 | 81.76 | 0 | 41419.3 | 106.57 | 1 | 0 | 0 | 2 |
| 14 | Patient126 | 1 | 333.5 | 104.48 | 3 | 279948.3 | 76.45 | 8 | 0 | 0 | 6 |
| 15 | Patient127 | 1 | 291.27 | 72.52 | 13 | 7995.4 | 9.6 | 5 | 0 | 0 | 10 |
| 16 | Patient13 | 0 | 166.3 | 58.7 | 3 | 19674.57 | 6.42 | 14 | 0 | 0 | 3 |
| 17 | Patient130 | 0 | 269 | 0 | 0 | 10460.28 | 8.64 | 20 | 0 | 0 | 2 |
| 18 | Patient131 | 1 | 129 | 0 | 0 | 8935 | 0 | 0 | 0 | 0 | 0 |
| 19 | Patient132 | 1 | 124.63 | 41.37 | 0 | 78921.12 | 22.87 | 40 | 0 | 0 | 0 |
| 20 | Patient133 | 0 | 189.33 | 60.58 | 3 | 25605.57 | 5.42 | 279 | 0 | 0 | 0 |
| 21 | Patient134 | 0 | 367.85 | 86.15 | 0 | 71966.97 | 75.53 | 2334 | 0 | 0 | 0 |
| 22 | Patient135 | 1 | 240.88 | 7 | 0 | 18233.76 | 3.24 | 103 | 0 | 0 | 0 |

# Machine Learning Pipeline: Data Preprocessing

# Z - score Normalization

- The purpose of Z-score normalization is to scale and centre the data such that it has a **mean of zero** and a **standard deviation of one**.
- This makes it easier to compare data that are on different scales or have different units.



**Formula:**

The formula to calculate the Z-score for a given data point is:

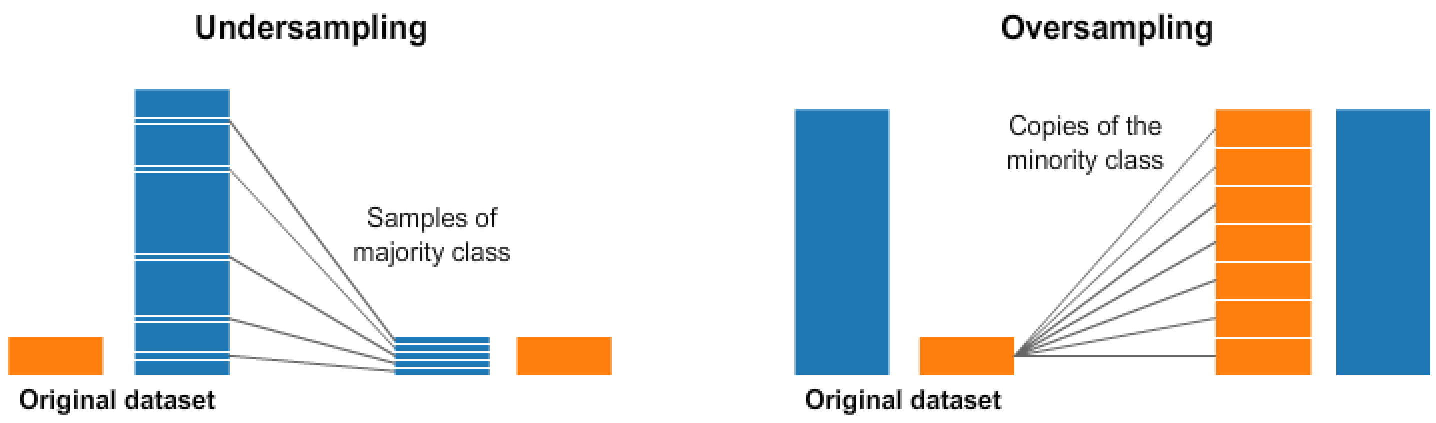
$$Z = \frac{(X - \mu)}{\sigma}$$

where:

- $Z$ is the Z-score.
- $X$ is the original data point.
- $\mu$ is the mean of the dataset.
- $\sigma$ is the standard deviation of the dataset.

# Handling Imbalanced Datasets

- Re-sampling techniques
  - Oversampling
  - Under-sampling



[8]

# Machine Learning Pipeline: Model Training

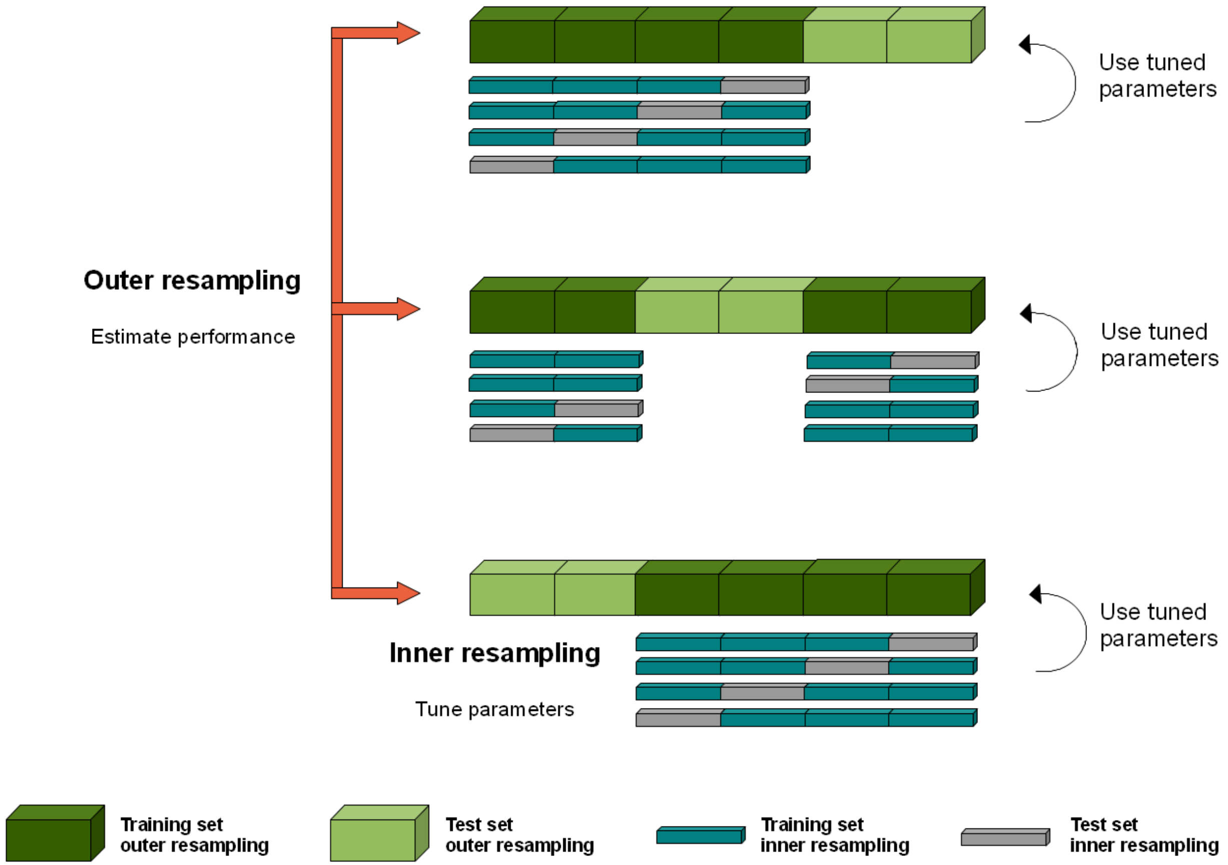
# Machine Learning Pipeline: Model Training and Evaluation

Nested k Fold Cross Validation

- Hyperparameter tuning
- Feature selection

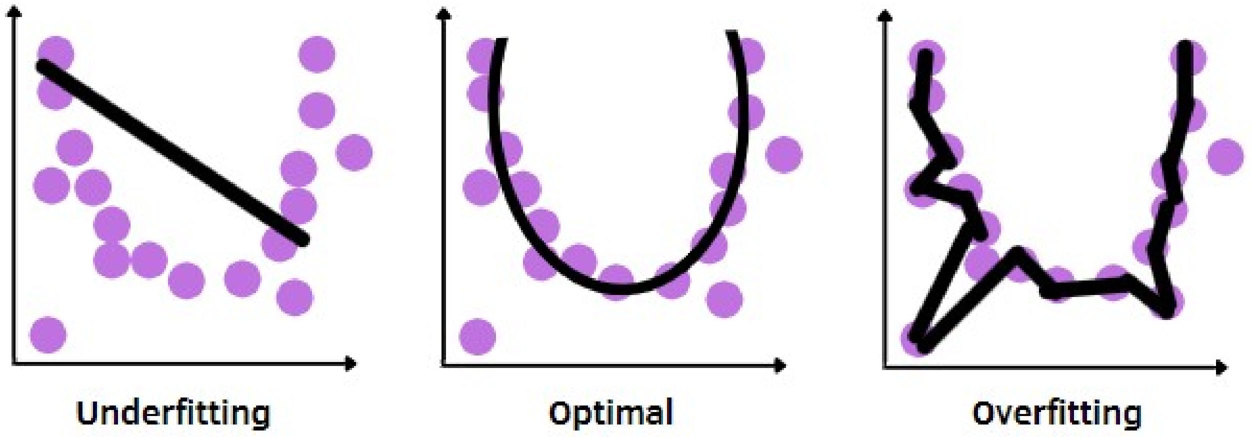
Evaluation Metrics

- Model selection
- **MCC** – Matthews Correlation Coefficient
- **ROC** – Receiver Operating Characteristic



[3]

# Hyperparameter tuning



Underfitting

Optimal

Overfitting

(model is too simple)

(model is too complex and captures even noise in the data)

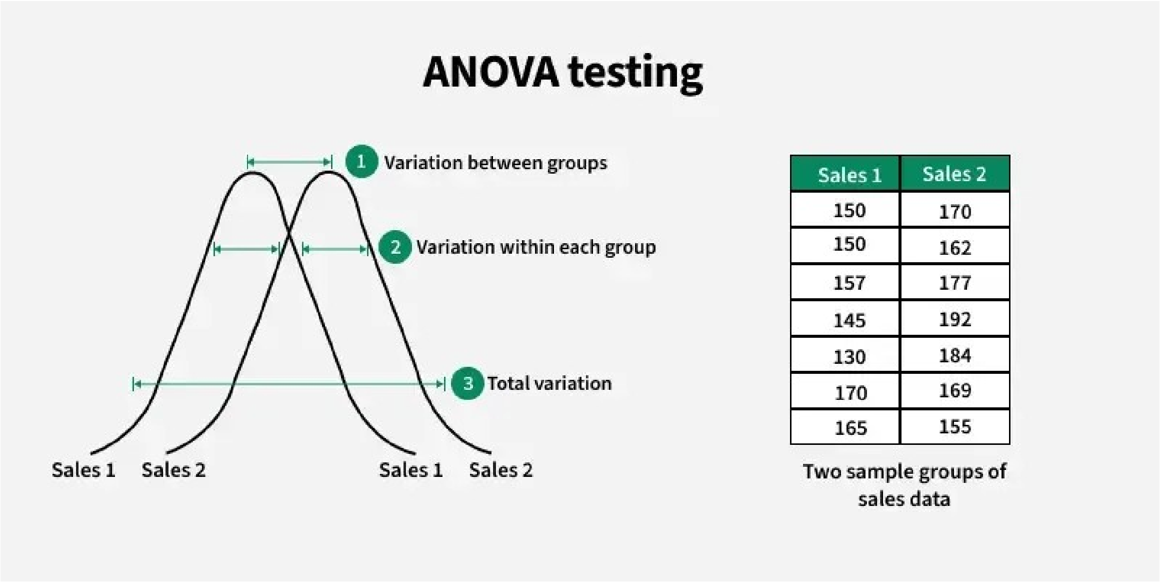
[8]

# Feature Selection

- ANOVA
- Lasso, L1 regularisation
- Mutual Info

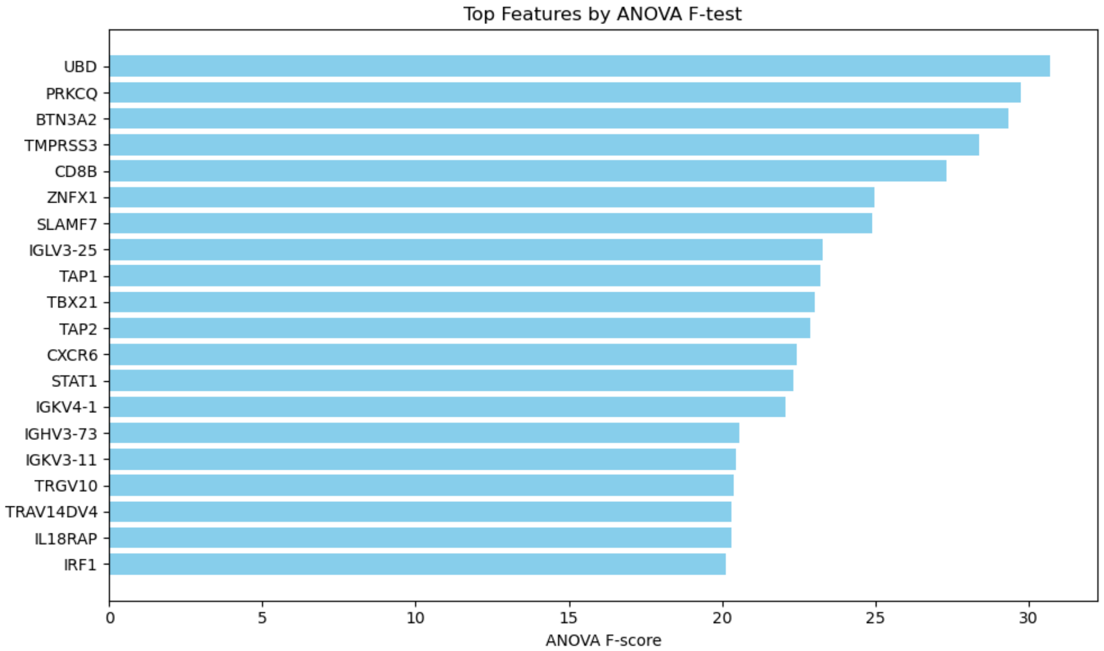
[8]

# ANOVA

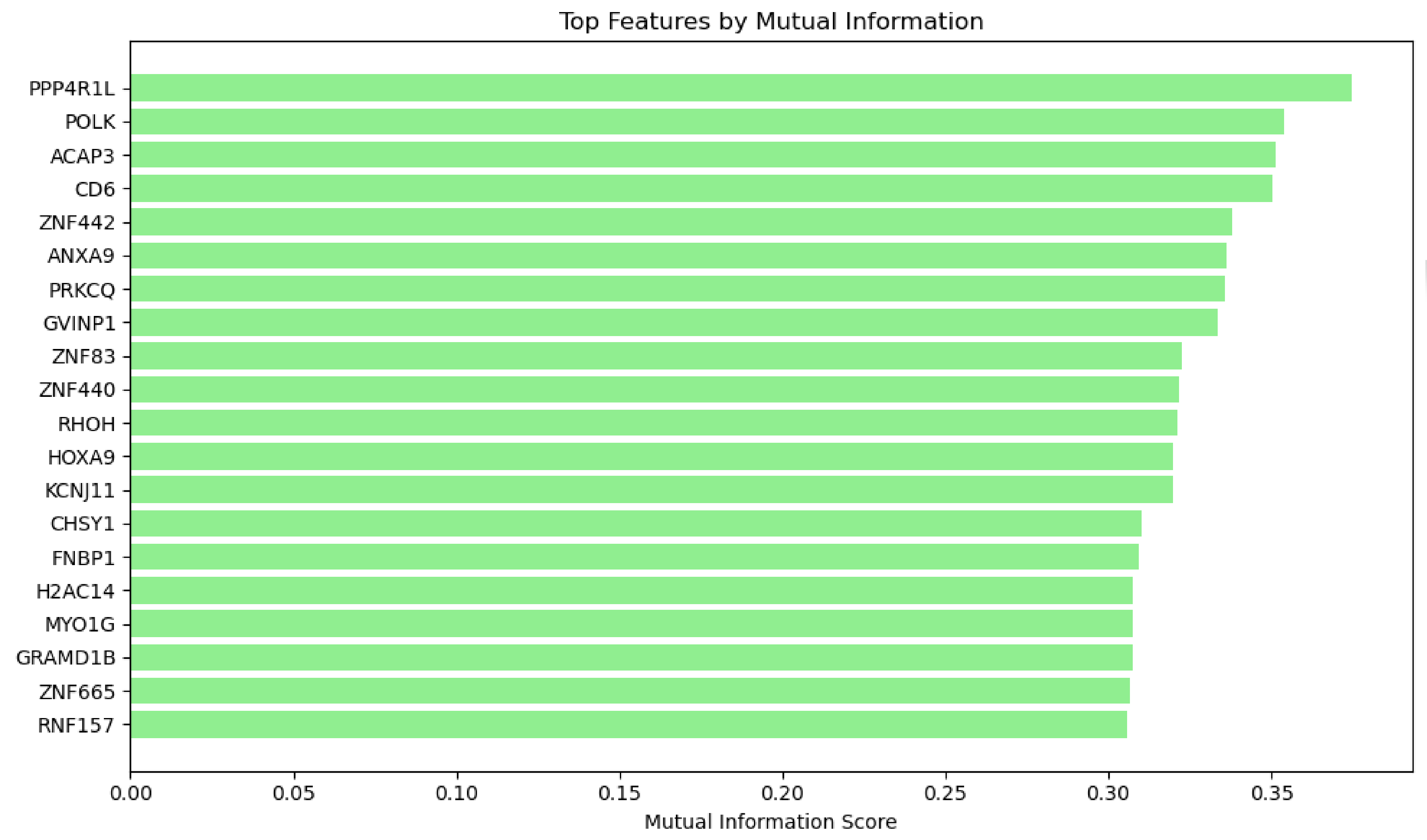


| Sales 1 | Sales 2 |
|---|---|
| 150 | 170 |
| 150 | 162 |
| 157 | 177 |
| 145 | 192 |
| 130 | 184 |
| 170 | 169 |
| 165 | 155 |

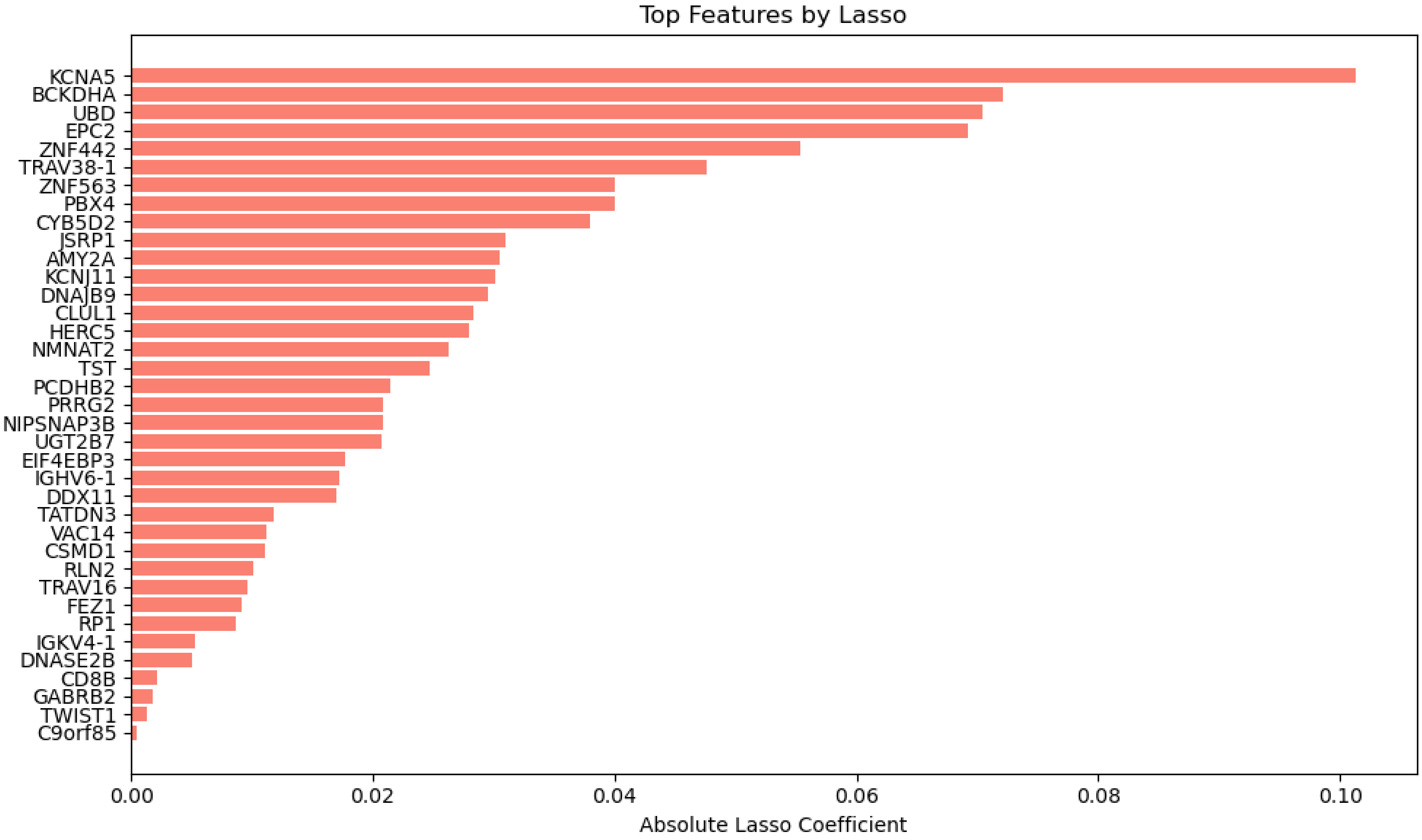
Two sample groups of sales data



# Mutual Info

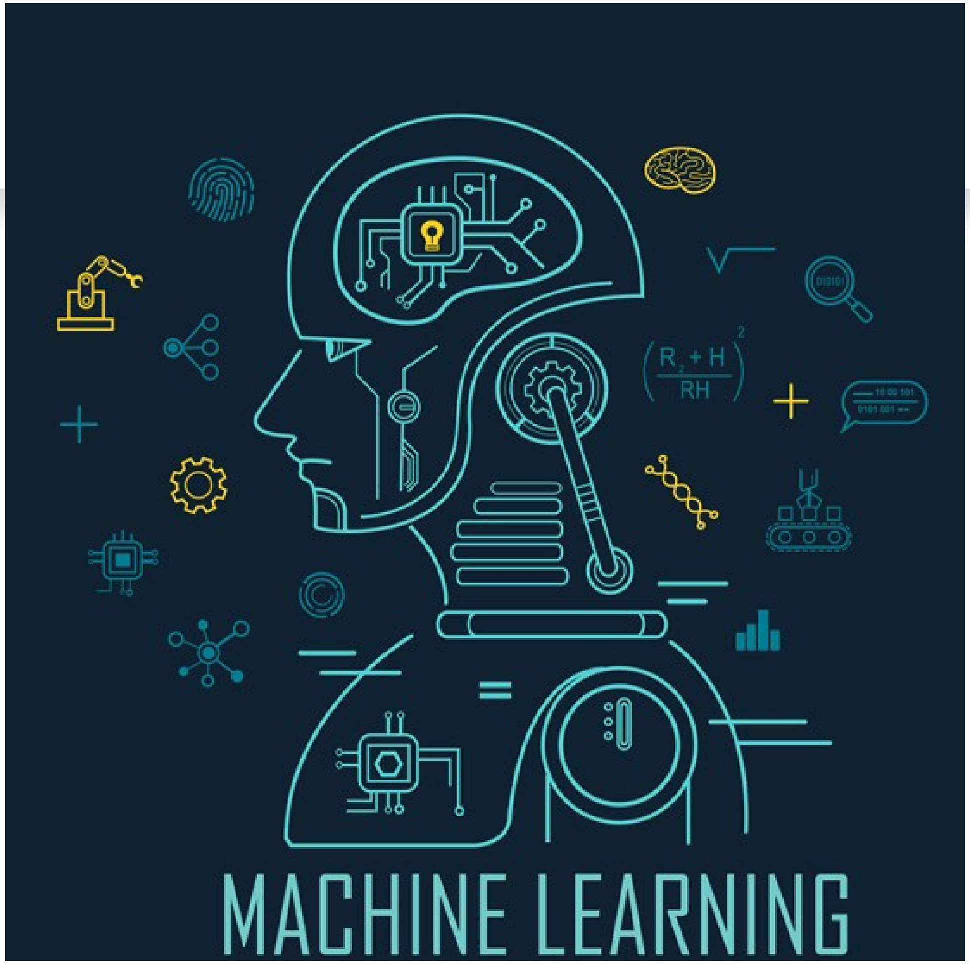

Top Features by Mutual Information
PPP4R1L
POLK
ACAP3
CD6
ZNF442
ANXA9
PRKCQ
GVINP1
ZNF83
ZNF440
RHOH
HOXA9
KCNJ11
CHSY1
FNBP1
H2AC14
MYO1G
GRAMD1B
ZNF665
RNF157
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Mutual Information Score

# Lasso, L1 regularisation



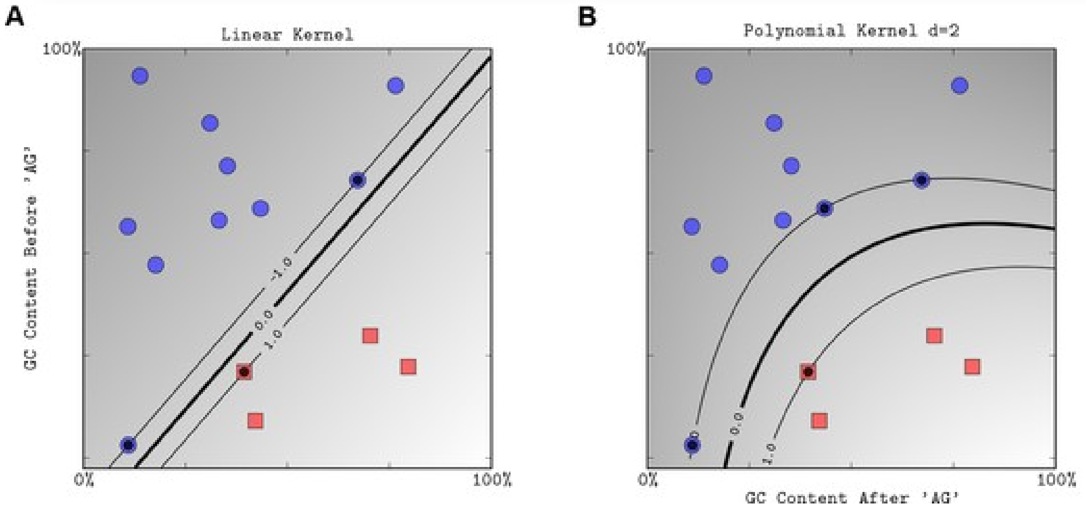
# Algorithms

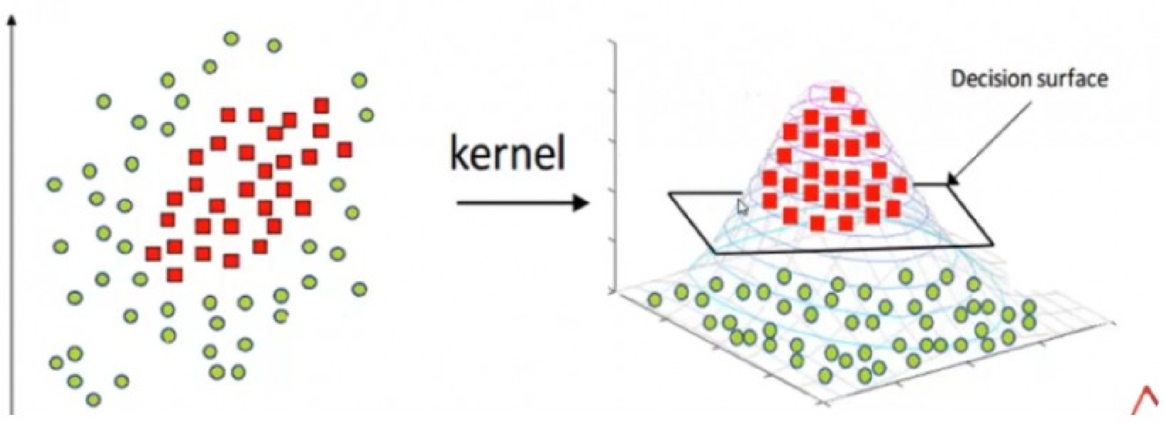
- Support Vector Machine
- Logistic Regression
- Extreme Gradient Boost
- Random Forest
- Classification and Regression Trees



[8]

# Support Vector Machine

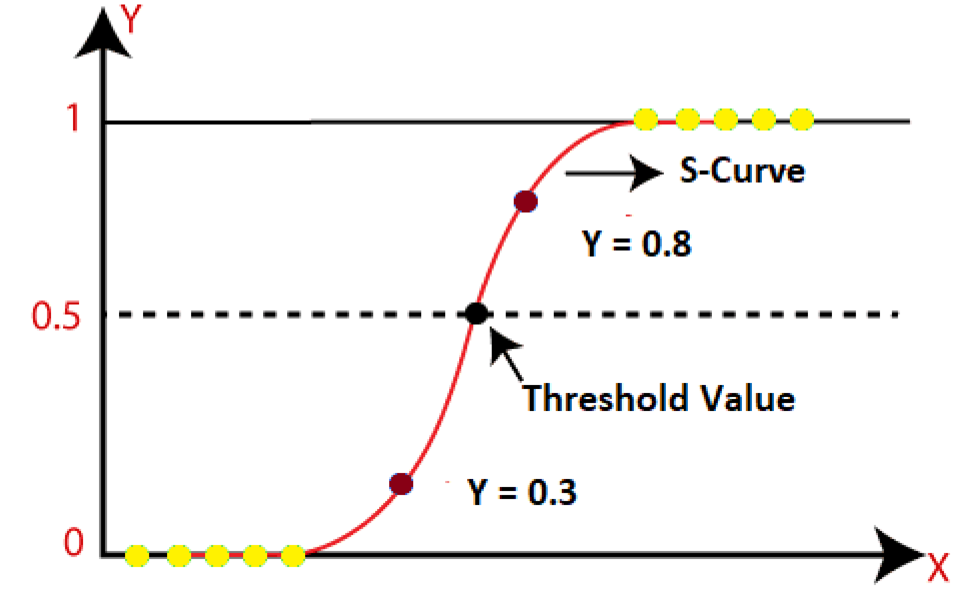




- Hyper Parameters: C, kernel

[8]

# Logistic Regression

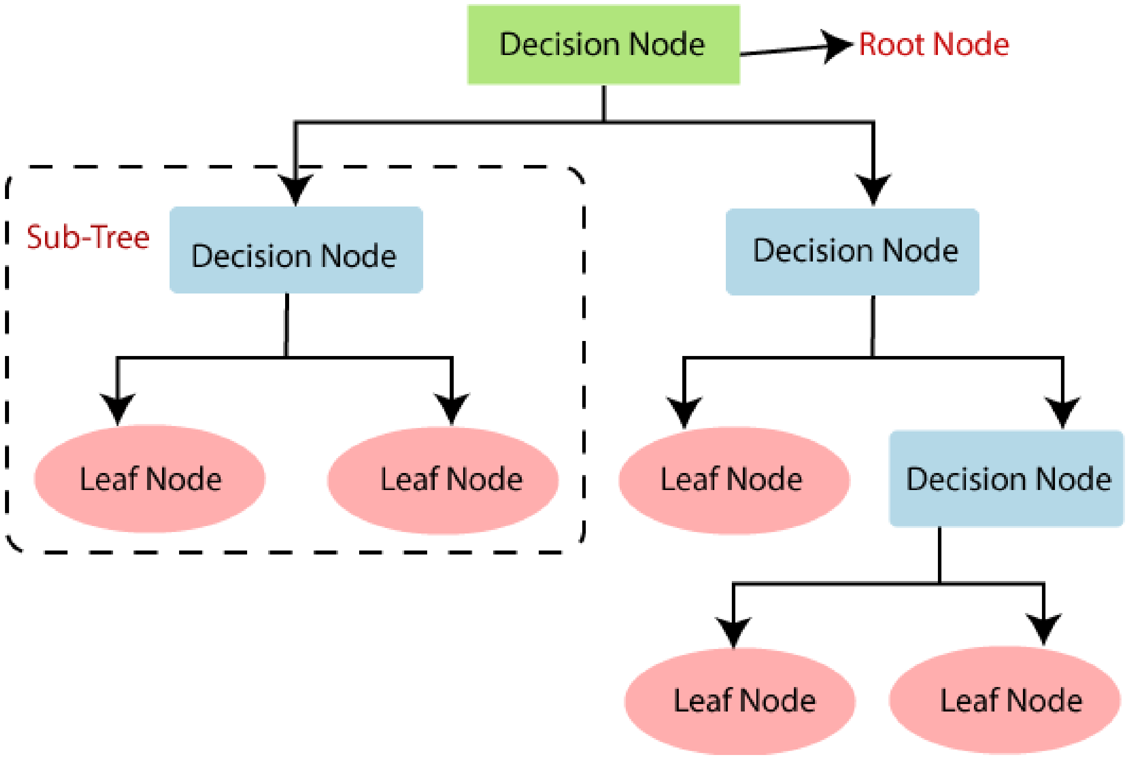


Logistic regression

- Hyper Parameters: C, penalty

[8]

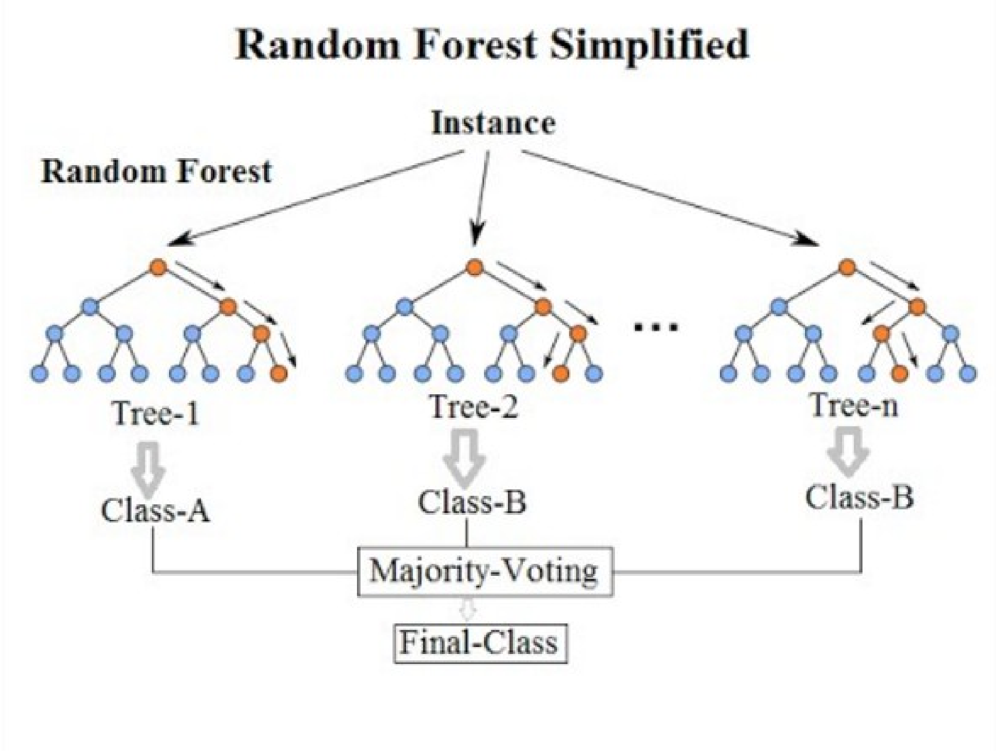
# Classification and Regression Trees



- Hyper Parameters:  max depth, min sample split

[8]

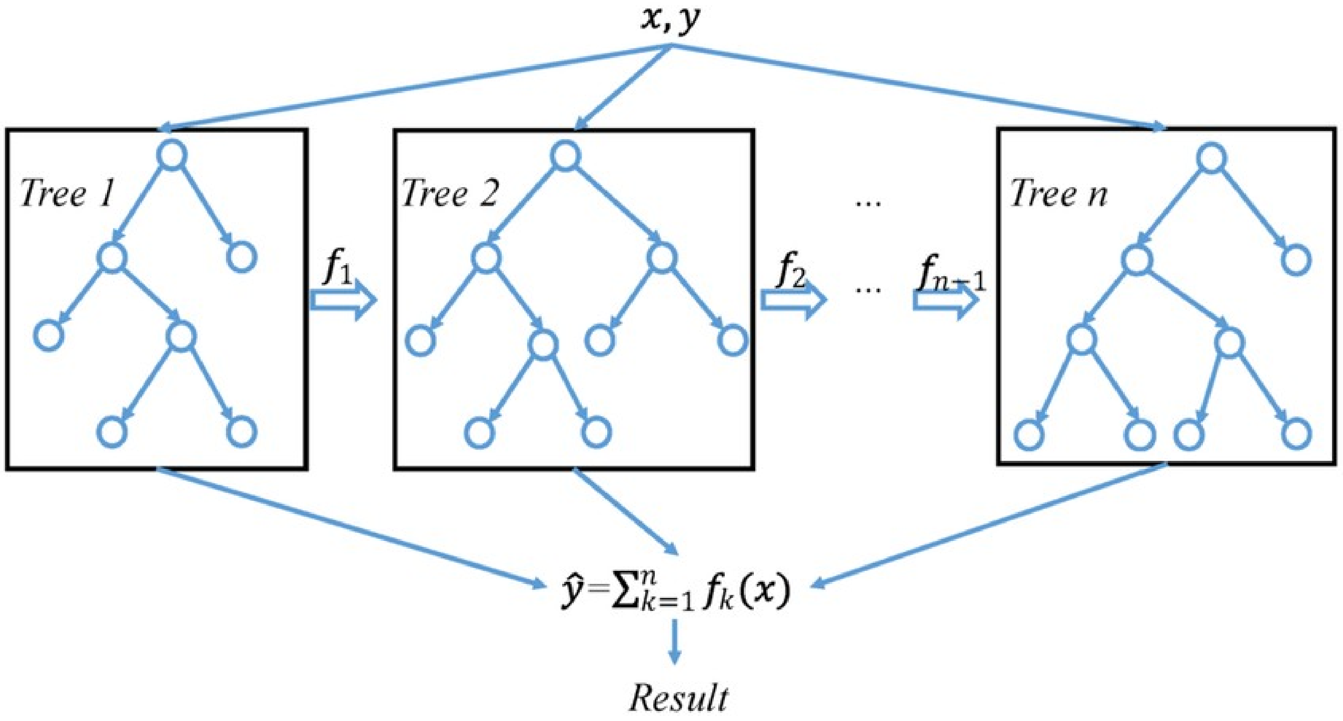
# Random Forest



- Hyper Parameters: n estimator, max depth

[8]

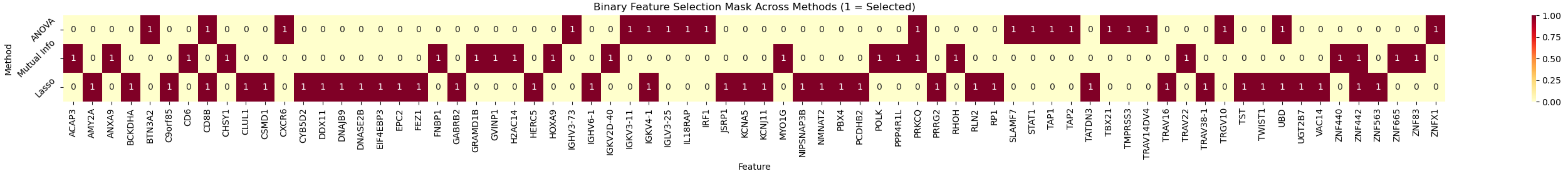
# Extreme Gradient Boost



- Hyper Parameters: n estimators, learning rate, max depth

[8]

# Binary Feature Selection Mask Across Methods (1 = Selected)



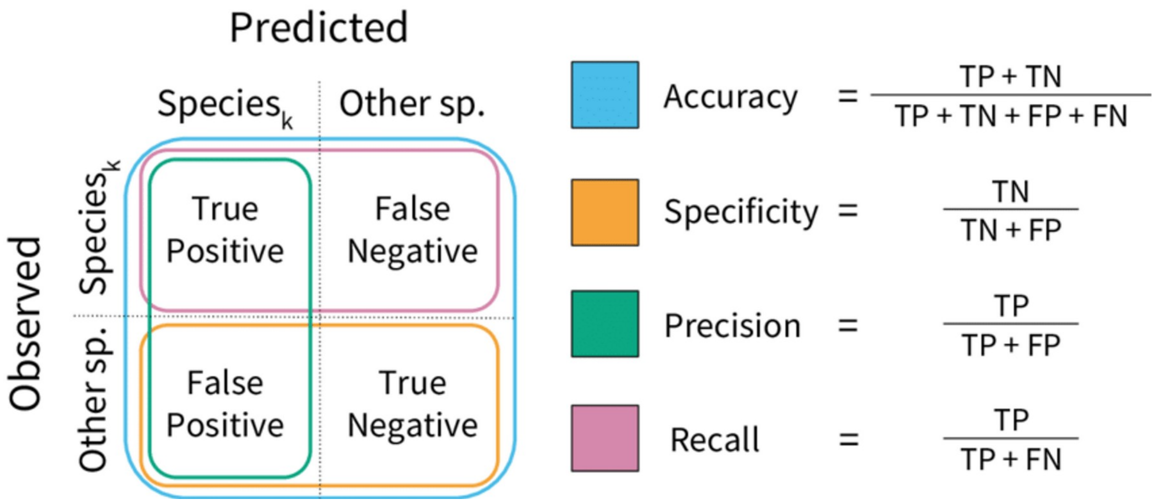
| Method | ACAP3 | AMY2A | ANXA9 | BCKDHA | BTN3A2 | C9orf85 | CD6 | CD8B | CHSY1 | CLUL1 | CSMD1 | CXCR6 | CYB5D2 | DDX11 | DNAJB9 | DNASE2B | EIF4EBP3 | EPC2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ANOVA | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Mutual Info | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Lasso | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |

| Method | FEZ1 | FNBP1 | GABRB2 | GRAMD1B | GVINP1 | H2AC14 | HERC5 | HOXA9 | IGHV3-73 | IGHV6-1 | IGKV2D-40 | IGKV3-11 | IGKV4-1 | IGLV3-25 | IL18RAP | IRF1 | JSRP1 | KCNA5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ANOVA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| Mutual Info | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Lasso | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |

| Method | KCNJ11 | MYO1G | NIPSNAP3B | NMNAT2 | PBX4 | PCDHB2 | POLK | PPP4R1L | PRKCQ | PRRG2 | RHOH | RLN2 | RP1 | SLAMF7 | STAT1 | TAP1 | TAP2 | TATDN3 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ANOVA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| Mutual Info | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Lasso | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |

| Method | TBX21 | TMPRSS3 | TRAV14DV4 | TRAV16 | TRAV22 | TRAV38-1 | TRGV10 | TST | TWIST1 | UBD | UGT2B7 | VAC14 | ZNF440 | ZNF442 | ZNF563 | ZNF665 | ZNF83 | ZNFX1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ANOVA | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| Mutual Info | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| Lasso | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 |

# Evaluation Metrics

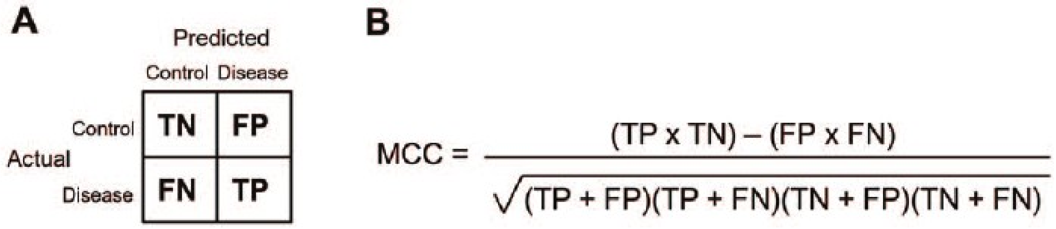


🟦 $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

🟧 $$\text{Specificity} = \frac{TN}{TN + FP}$$

🟩 $$\text{Precision} = \frac{TP}{TP + FP}$$

🟪 $$\text{Recall} = \frac{TP}{TP + FN}$$

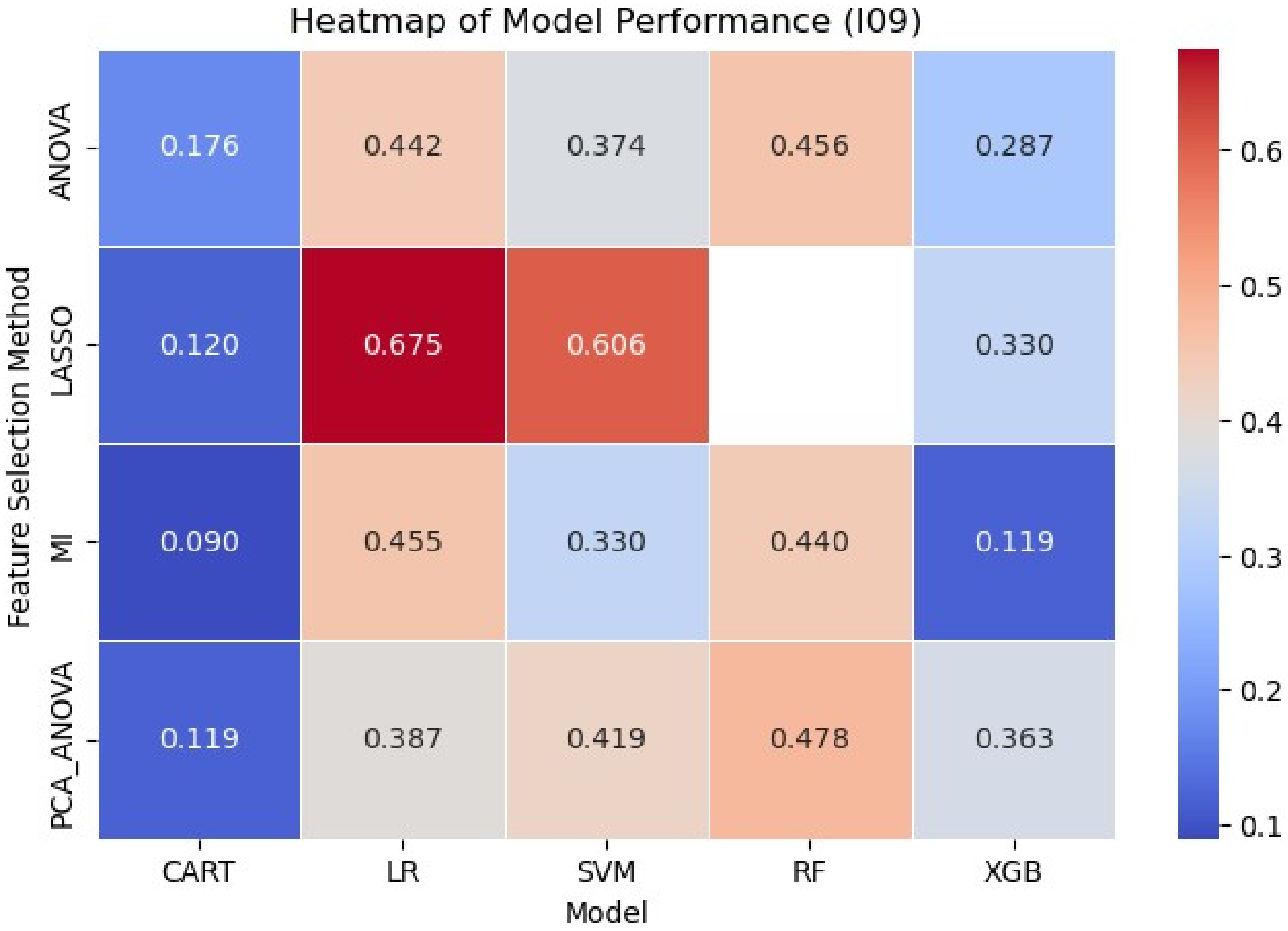


| | | Predicted | |
|---|---|---|---|
| | | Control | Disease |
| Actual | Control | TN | FP |
| | Disease | FN | TP |

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

[8]

# I09 Dataset: Results (MCC)



# Test on I15

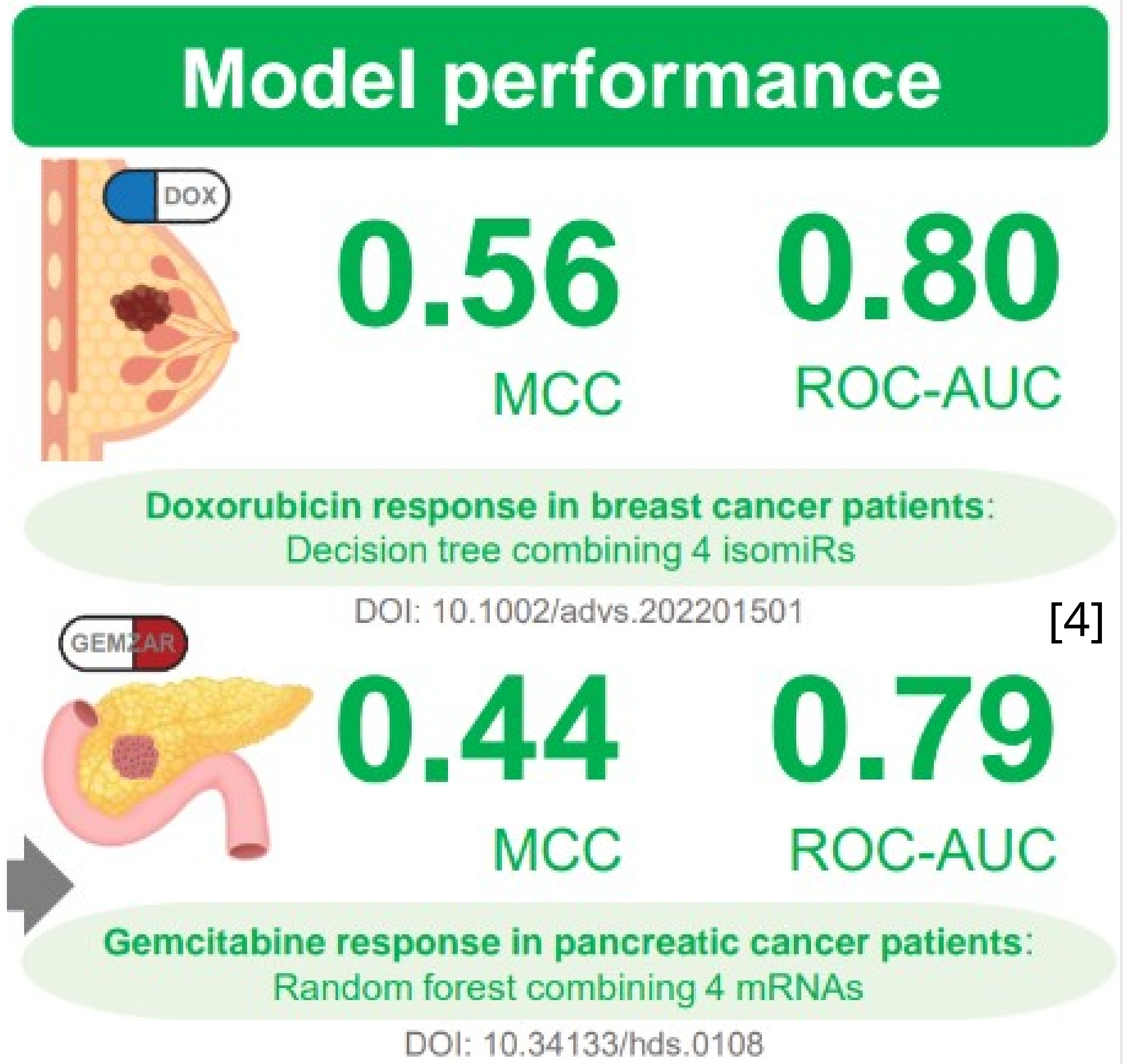
|  | CV MCC | Test MCC |
|---|---|---|
| LR+LASSO | 0.690228055 | -0.13092 |
| SVM+LASSO | 0.671850832 | 0.005 |

# THANK YOU!
# QUESTIONS

# References

1. National Cancer Institute, Winslow T. Immune Checkpoint Inhibitor (PD-1). NCI Visuals Online. June 8, 2016. Available from: https://visualsonline.cancer.gov/details.cfm?imageid=10396
2. Wolchok JD, et al. Overall survival with combined nivolumab and ipilimumab in advanced melanoma. N Engl J Med. 2017;377(14):1345-1356.
3. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM. mlr: Machine Learning in R. J Mach Learn Res. 2016;17(170):1-5.
4. Ogunleye AZ, Piyawajanusorn C, Gonçalves A, Ghislat G, Ballester PJ. Interpretable machine learning models to predict the resistance of breast cancer patients to doxorubicin from their microRNA profiles. Adv Sci (Weinh). 2022 Aug;9(24). doi: 10.1002/advs.202201501. Epub 2022 Jul 3. PMID: 35785523; PMCID: PMC9403644.
5. Ogunleye A, Piyawajanusorn C, Ghislat G, Ballester PJ. Large-Scale Machine Learning Analysis Reveals DNA Methylation and Gene Expression Response Signatures for Gemcitabine-Treated Pancreatic Cancer. Health Data Sci. 2024;4:0108. doi: 10.34133/hds.0108. PMID: 38486621. Available from: https://pubmed.ncbi.nlm.nih.gov/38486621/
6. Tsamardinos I, Lagani V, Papagregoriou G, Tsagris M, Borboudakis G, Zenker M, et al. A perspective on automated machine learning in bioinformatics. Brief Bioinform. 2022;23(1). doi: 10.1093/bib/bbac059. Available from: https://pubmed.ncbi.nlm.nih.gov/35382509/
7. Tanner G. Kernel PCA Explained. ML Explained. January 2, 2022. Available from: https://ml-explained.com/blog/kernel-pca-explained
8. Jappinen R. Resampling strategies for imbalanced datasets. Kaggle. Available from: https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets

# AI in Precision Oncology Case Studies



Model performance

Doxorubicin response in breast cancer patients:
Decision tree combining 4 isomiRs

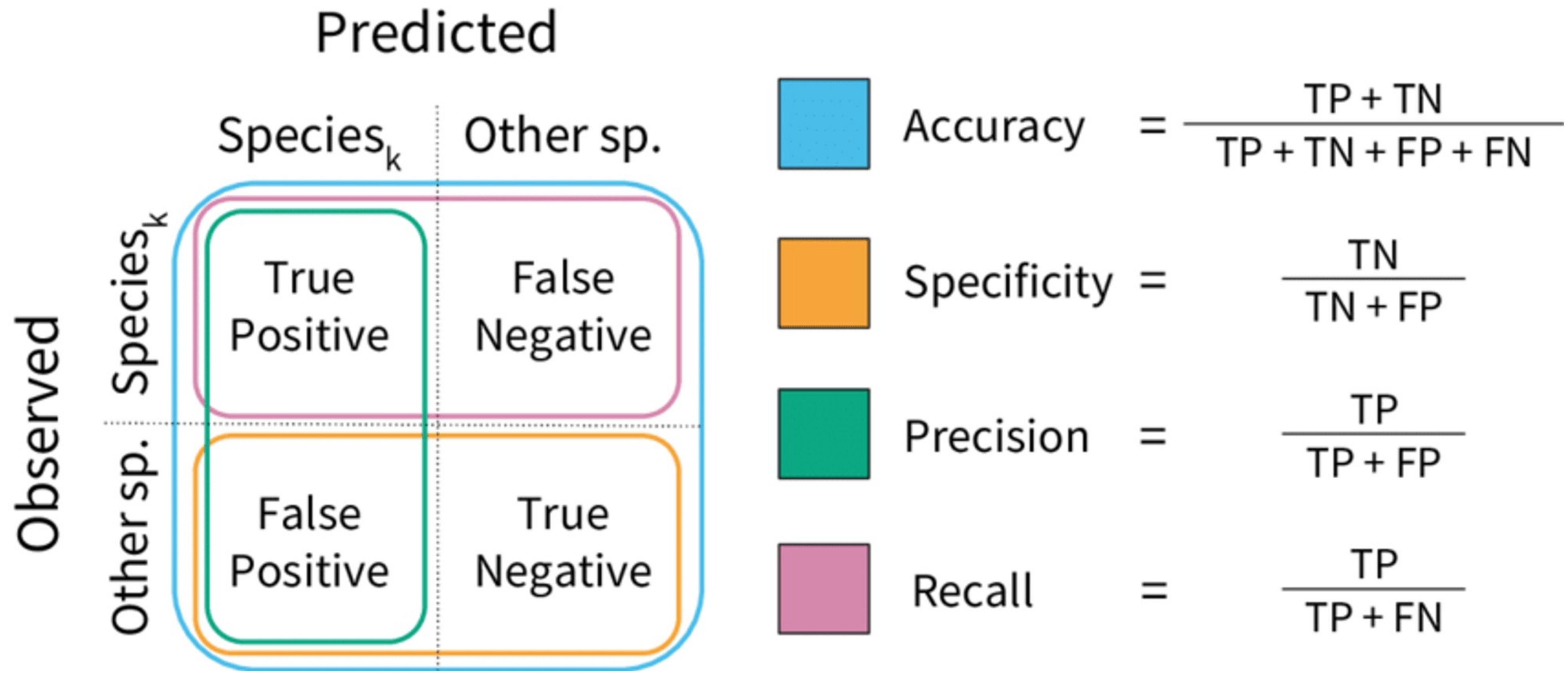
DOI: 10.1002/advs.202201501

[4]

Gemcitabine response in pancreatic cancer patients:
Random forest combining 4 mRNAs

DOI: 10.34133/hds.0108

[5]



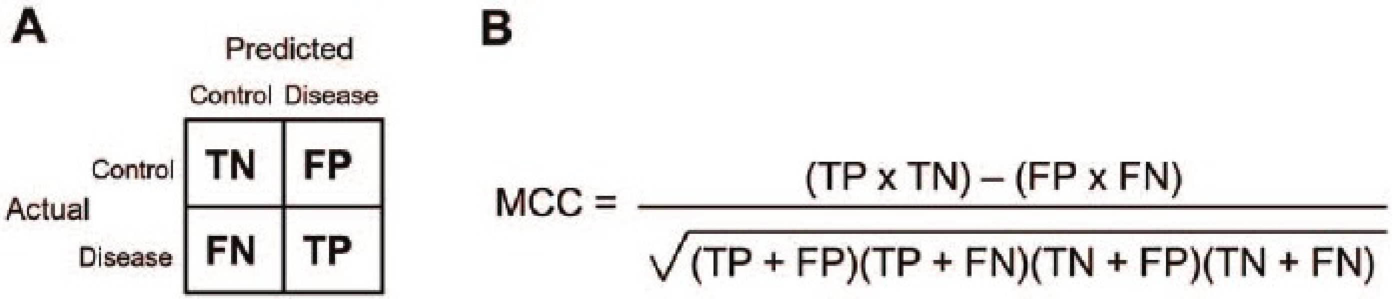
| | | Predicted | |
|---|---|---|---|
| | | $\text{Species}_k$ | Other sp. |
| Observed | $\text{Species}_k$ | True Positive | False Negative |
| | Other sp. | False Positive | True Negative |

🟦 Accuracy $= \dfrac{TP + TN}{TP + TN + FP + FN}$

🟧 Specificity $= \dfrac{TN}{TN + FP}$

🟩 Precision $= \dfrac{TP}{TP + FP}$

🟪 Recall $= \dfrac{TP}{TP + FN}$



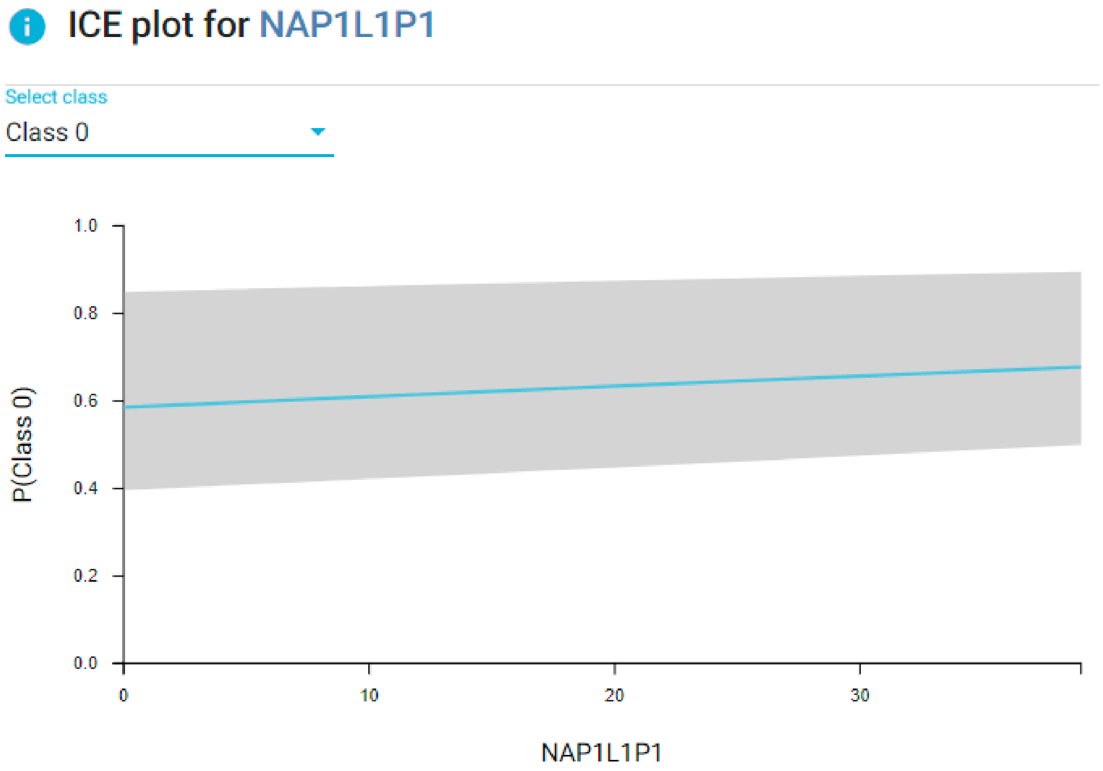
A

| | | Predicted | |
|---|---|---|---|
| | | Control | Disease |
| Actual | Control | TN | FP |
| | Disease | FN | TP |

B

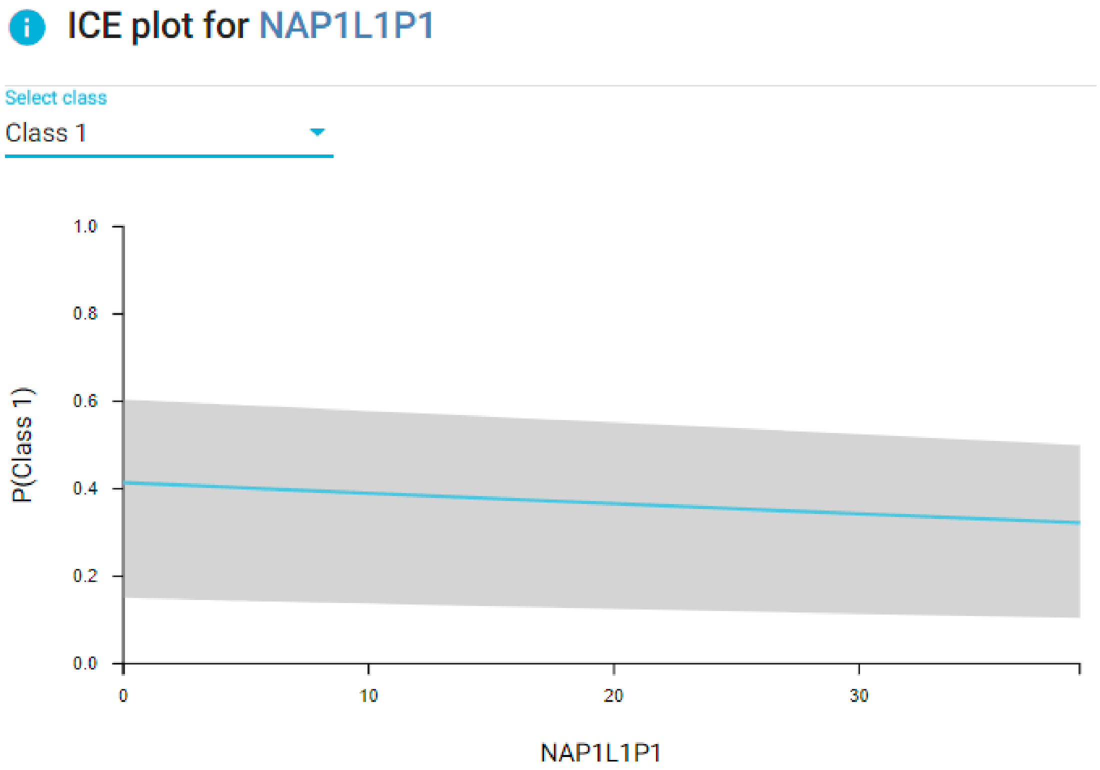
$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

# I01 Dataset Feature ICE Plot (Individual Conditional Expectation Plot)



ICE plot for NAP1L1P1

Select class
Class 0



ICE plot for NAP1L1P1

Select class
Class 1

# What is transcriptomics

**Transcriptomics** is the study of the complete set of RNA transcripts produced by the genome in a specific cell, tissue, or organism.

**TPM – Transcripts per Kilobase Million**

Normalize for gene length first, and then normalize for sequencing depth second