T-Cell Research

Gabrielle Salamanca

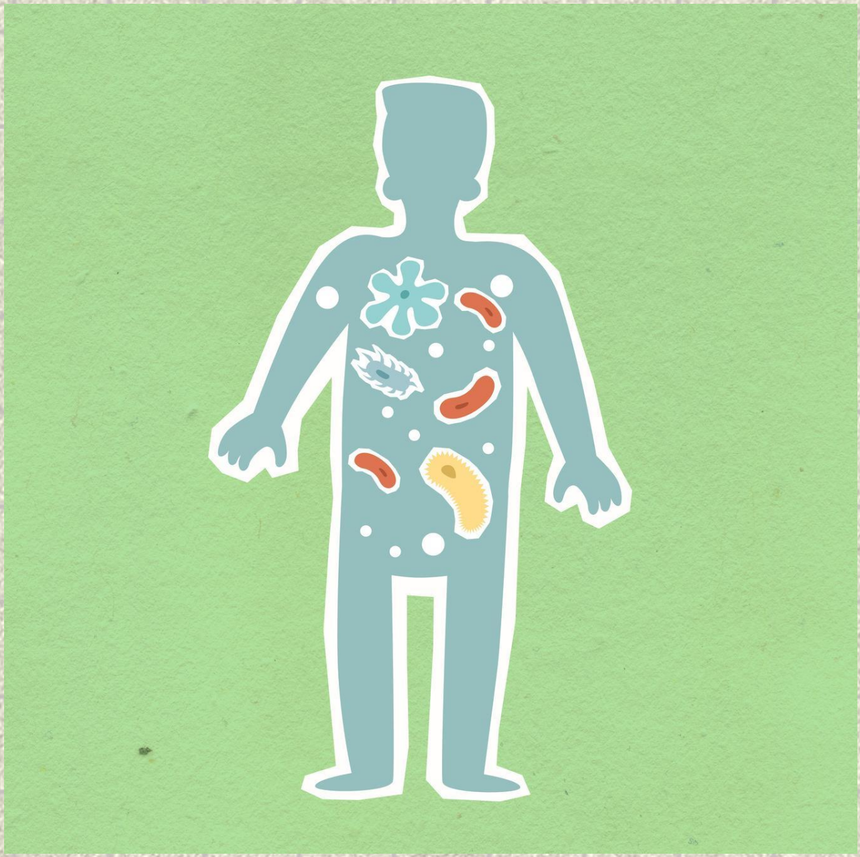
Culminating Experience: Math 895 Project

Table of Contents

01

INTRODUCTION

About the Project



The Start

- Spring 2023, Research Assistant
- Fall 2025, Culminating Experience

The Inspiration

Can the TCR/BCR profile to predict the cancer type?

The Project

Can we use the TCR/BCR profile to predict if the patient has covid or not?

The Dataset

Dimensions

- 108 rows
- 679 columns

About

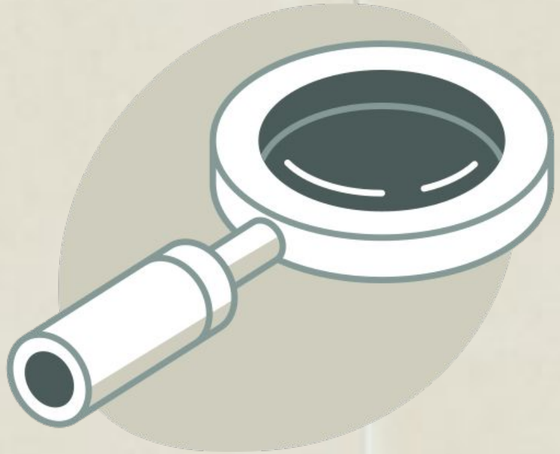
- Combo of 3 datasets

Datasets

- vj gene/patient
- vj gene/healthy patient
- patient info

Sample.ID	TRBV7-9_TRBJ1-1	. . .	Y
1_1	1.34581656747574	. . .	disease
1_2	1.34581656747574	. . .	disease
.	.	.	.
.	.	.	.
.	.	.	.

Dataset Prep

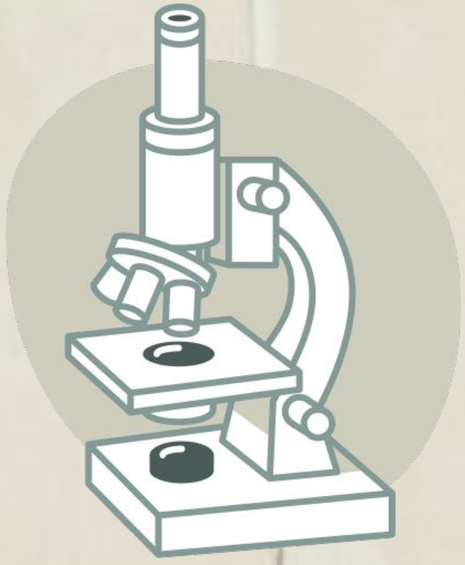


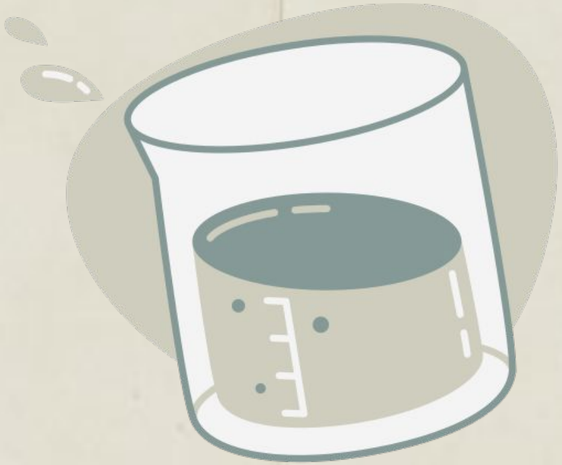
Fully Combined Dataset

- Y = 92 NAs
- Row 22 = no Y

Edits

- Checked prev datasets
  - Row 22 = removed
  - Rest of NAs were healthy

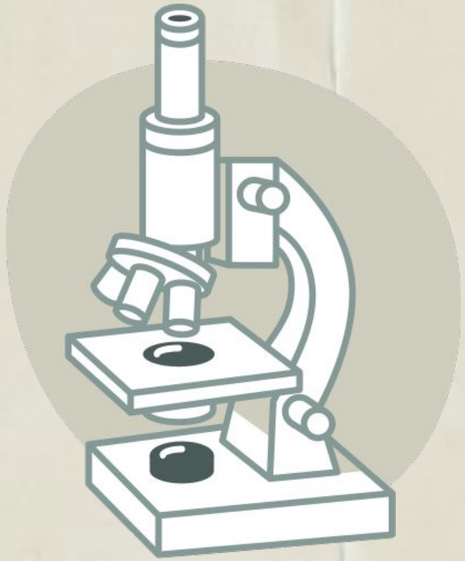




Splitting the data

- Training set
  - 86 rows x 679 cols
- Test set
  - 22 rows x 679 cols

Dataset Prep

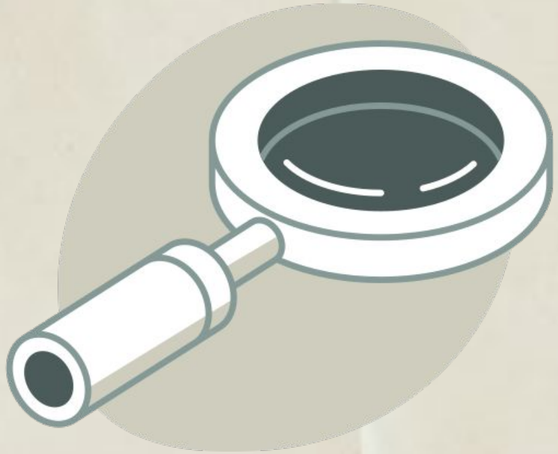


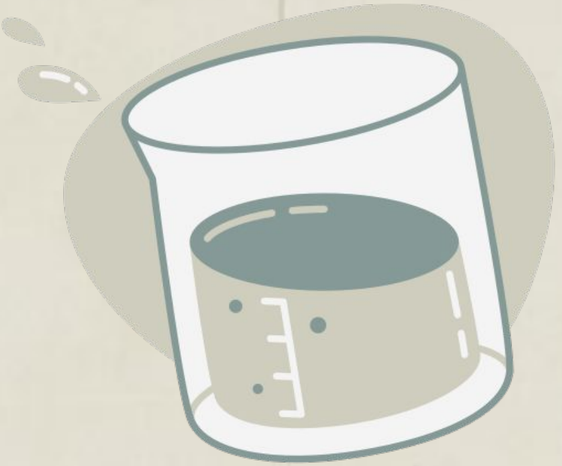
Significant Genes

Because we have 679 vj genes, we want to find the ones that contribute the most. Store those that do and make sure it's in both training and test set.

Prep

- Run a for loop
  - Run glm() with all the genes
  - store their p-values
    - Only those with p-value < 0.05 are kept





Dataset

37/679 significant genes

- Will be used for every method
- Will be tested in batches
  - 5 → 10 → ... → 35 →37

02

Applications

Note: Method Measure

We’ll be using test accuracy to measure our methods

- The percentage of correct predictions our model makes on our test data

Part 1: The Basics

Logistic Regression

- Probabilistic/parametric classifier
- Decision boundary-based predictions

K-Nearest Neighbor

- k = 3,5,7
- Generative, nonlinear classifier
- Proximity-based predictions

Quadratic Discriminant Analysis

- Generative, nonlinear classifier
- Distribution-based predictions
- Assumes each class has a different coV matrix

Linear Discriminant Analysis

- Generative, linear classifier
- Distribution-based predictions
- Assumes the classes has same coV matrix

Logistic Regression

Groups	Test Accuracy
Top 5	0.6818
Top 10	0.7273
Top 15	0.6818
Top 20	0.8182
Top 25	0.7727
Top 30	0.7727
Top 35	0.7727
Top 37	0.7727

Important note!

Warning for all non-integer #successes in binomial glm

When originally ran, the algorithm didn't converge, where fitted probabilities numerically 0 or 1 occurred; and we also had the warning above

We used a separation-safe glm to address the first issue, but even when we converted the Y from categorical to numerical, we still get the warning above

QDA

Groups	Test Accuracy
Top 5	0.6818
Top 10	0.7273
Top 15	0.7273
Top 20	0.7273
Top 25	0.6364
Top 30	NA
Top 35	NA
Top 37	NA

Important note!

Warning
some group too small for qda

This means at least one class doesn't have enough samples to estimate its covariance matrix

This is why the bottom three have no test accuracies

LDA

Groups	Test Accuracy
Top 5	0.6818
Top 10	0.7727
Top 15	0.7727
Top 20	0.7727
Top 25	0.8182
Top 30	0.7727
Top 35	0.7727
Top 37	0.7273

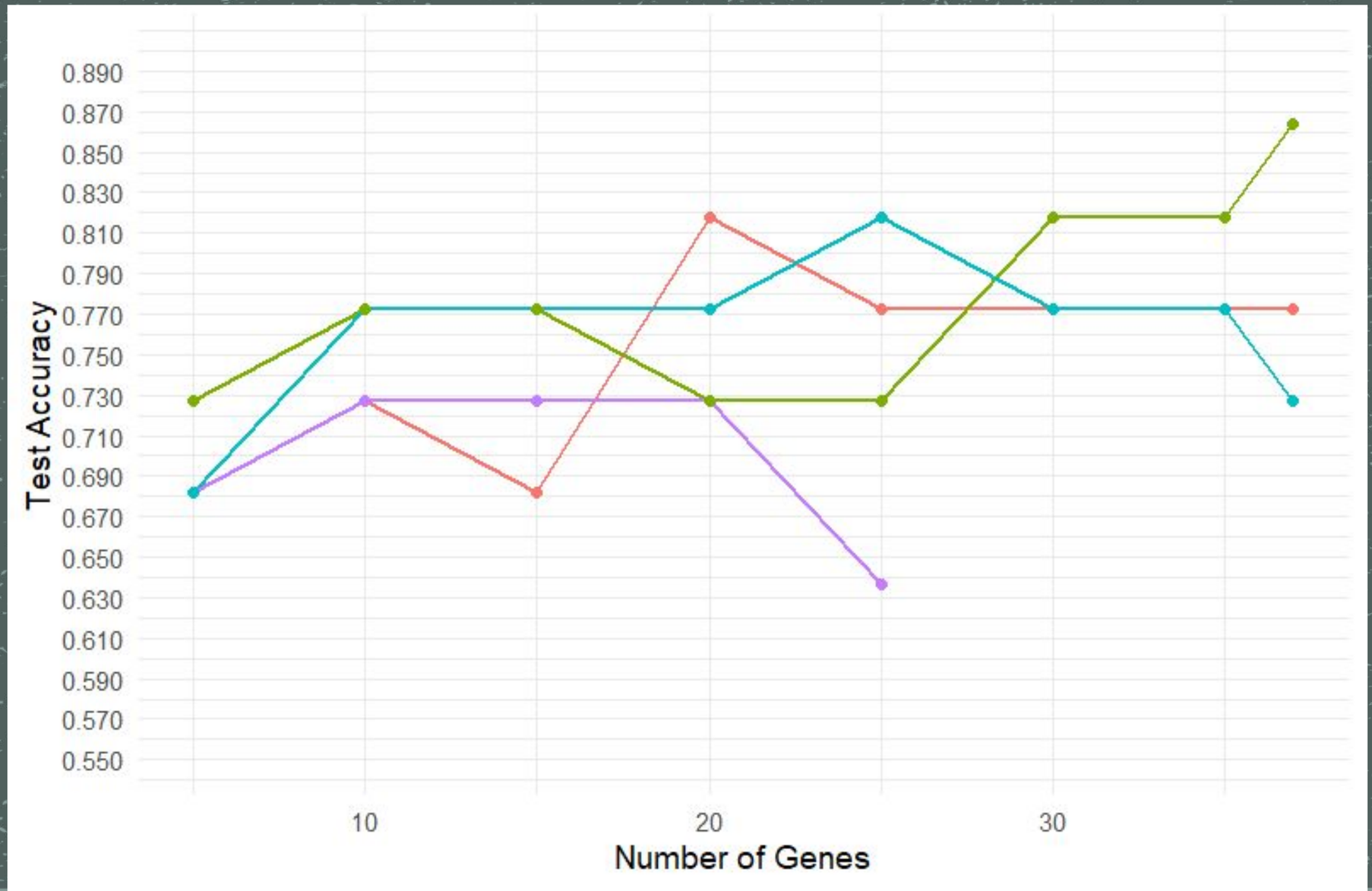
Because LDA assumes shared covariance, it only estimates a single covariance matrix

Therefore, it is able to give us the test accuracy for all the groups compared to QDA

KNN

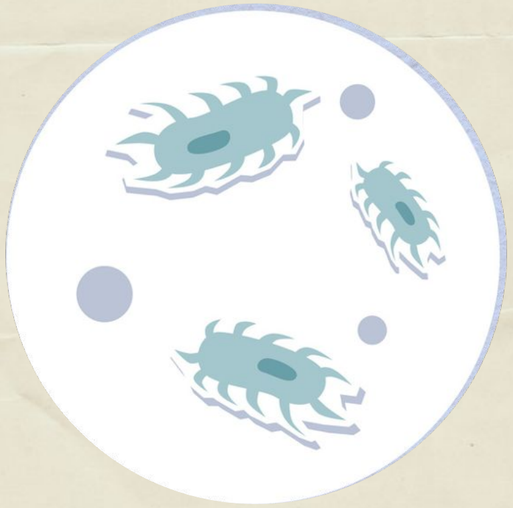
Groups	TA k=3	TA k=5	TA k=7
Top 5	0.7273	0.7273	0.7273
Top 10	0.7273	0.7727	0.7727
Top 15	0.8182	0.7727	0.7727
Top 20	0.7727	0.7727	0.7273
Top 25	0.8182	0.7727	0.7273
Top 30	0.7727	0.7727	0.8182
Top 35	0.8182	0.7727	0.8182
Top 37	0.7273	0.7727	0.8636

Test Accuracy of Basics



Part 2: Advanced Techniques

Classification Tree

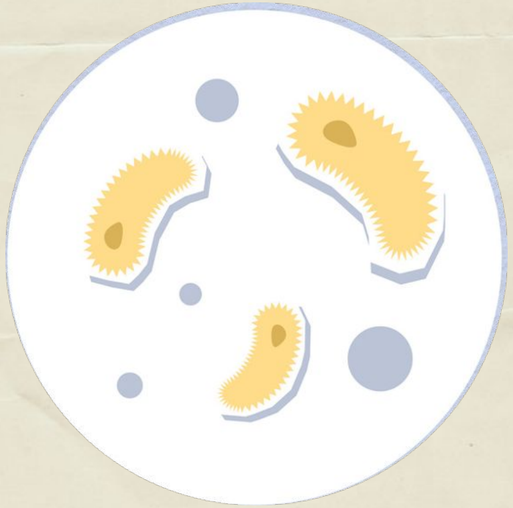


A decision tree used for classification

Boosting

A ML technique that builds a strong model by combining many weak ones

Random Forest



An ensemble of 100s of decision trees, each trained on a random subset of data & features

Classification Tree

Groups	Test Accuracy
Top 5	0.6364
Top 10	0.6818
Top 15	0.6818
Top 20	0.7273
Top 25	0.6818
Top 30	0.5909
Top 35	0.6364
Top 37	0.7273

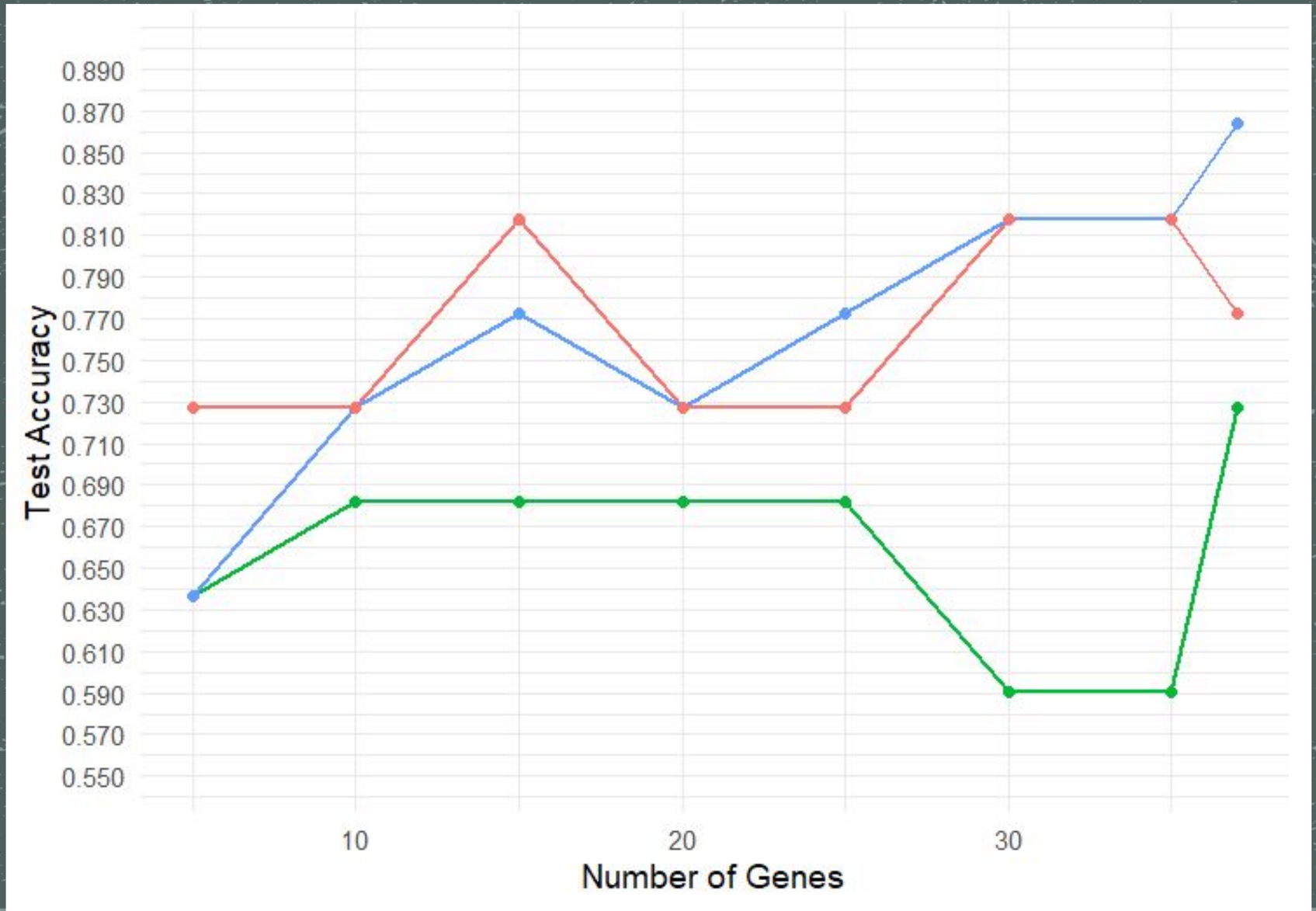
Boosting

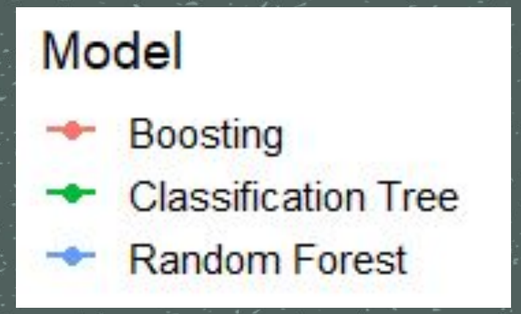
Groups	Test Accuracy	Shrinkage	Depth	Trees
Top 5	0.6818	0.01	3	352
Top 10	0.6818	0.1	4	29
Top 15	0.8182	0.05	1	152
Top 20	0.7273	0.01	1	737
Top 25	0.7273	0.01	1	942
Top 30	0.8182	0.05	1	147
Top 35	0.8182	0.1	2	94
Top 37	0.8182	0.01	1	1358

Random Forest

Groups	Test Accuracy	Best mtry	OOB error
Top 5	0.7273	2	0.2442
Top 10	0.7727	6	0.1512
Top 15	0.8182	2	0.1628
Top 20	0.7273	4	0.1744
Top 25	0.7273	8	0.1512
Top 30	0.8182	30	0.1628
Top 35	0.8182	6	0.1512
Top 37	0.7727	6	0.1744

Test Accuracy of Advanced

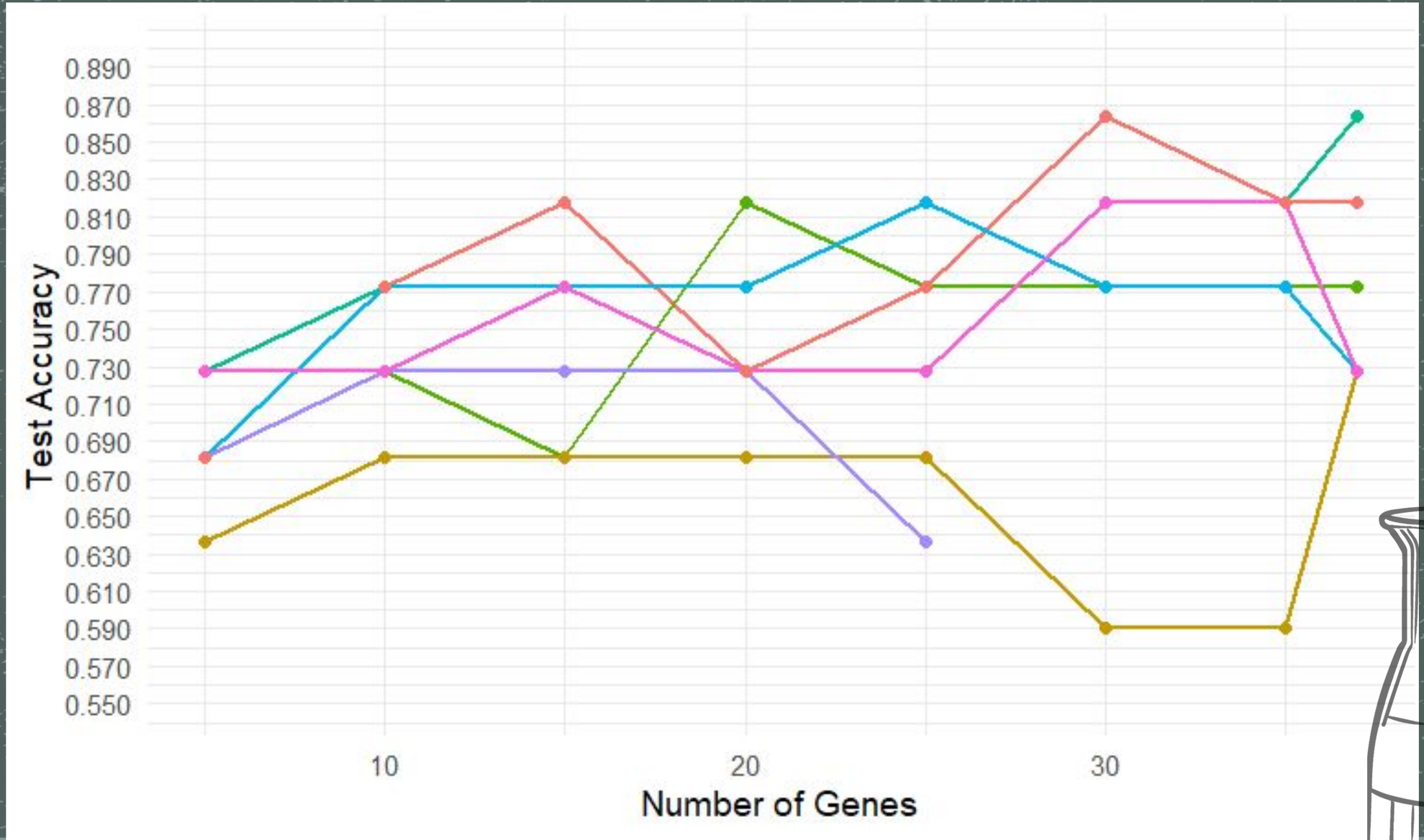


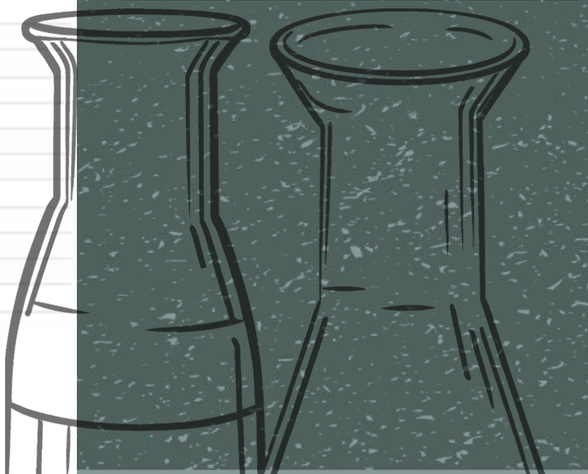


03

Conclusion

Test Accuracy of all Models





Results

The Basics: Rankings

- KNN (k=7)
- Log regress/LDA/KNN (k=3)
- KNN (k=5)
- QDA (incomplete)

Advanced: Rankings

- Boosting/Random Forest
- Classification Tree

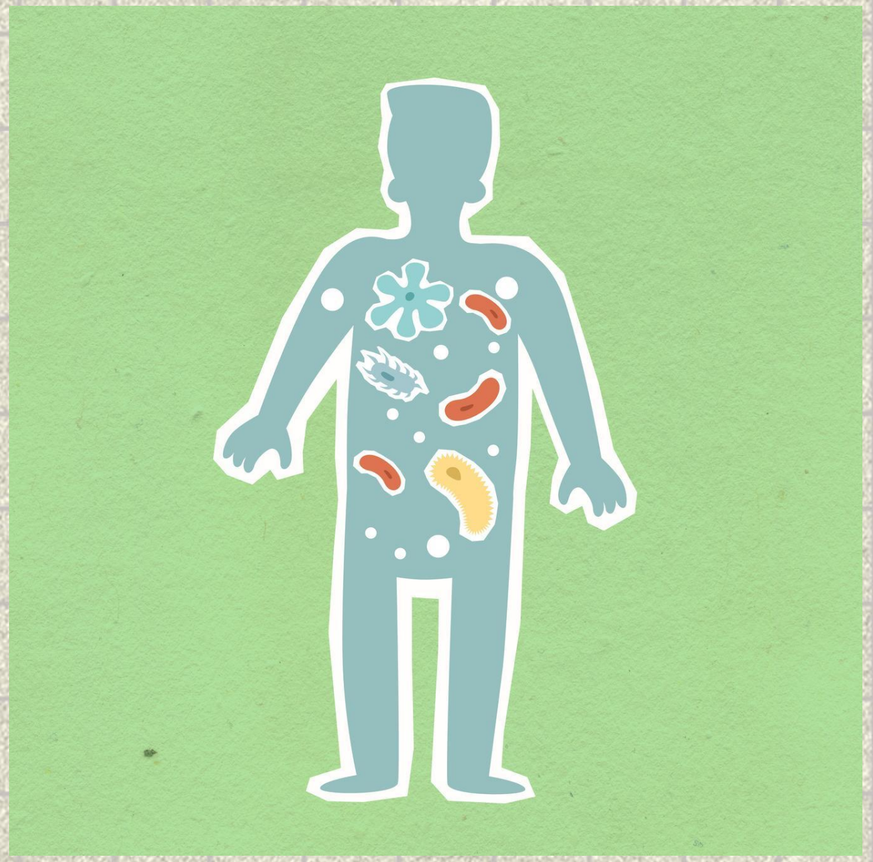
The Basics: The Numbers

- KNN (k=7) → 0.8636
- Log regress → 0.8182
- LDA → 0.8182
- KNN (k=3) → 0.8182
- KNN (k=5) → 0.7727

Advanced: The Numbers

- Boosting → 0.8182
- Random Forest → 0.8182
- Classification Tree → 0.7273

Overall



This project

It was good to revisit an old project and apply techniques that I've done in my undergrad and grad courses

Being a research assistant & this culminating experience has given me invaluable experience

Thank You

Credits

Tao He

Professor, Mentor, Project Advisor

Slides Carnival

for the presentation template

PEXELS & PIXABAY

for the photos