Colorectal Cancer

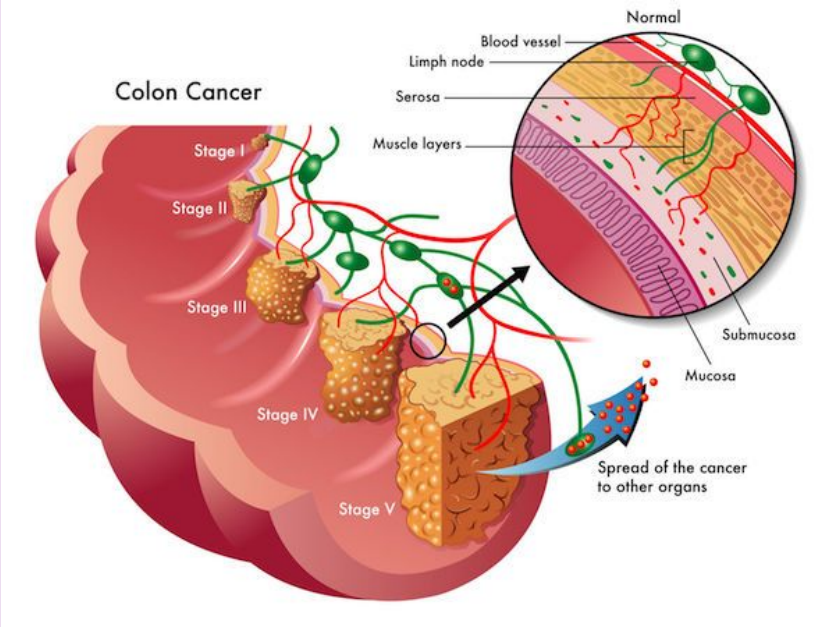
Predicting Presence of Colorectal Cancer using Metabolites

Luis Aguilar, Claire Hsieh, Michelle Tekawy, Daisy Wang

Background

- 3rd most common cancer in the US
- Second leading cause of cancer death according to the AMA
- Early detection leads to more successful treatments

SEER Stage	5 year Survival rate
Localized	90%
Regional	74%
Distant	17%

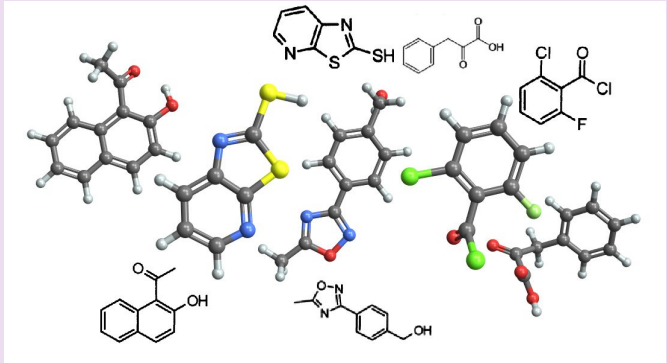


Background

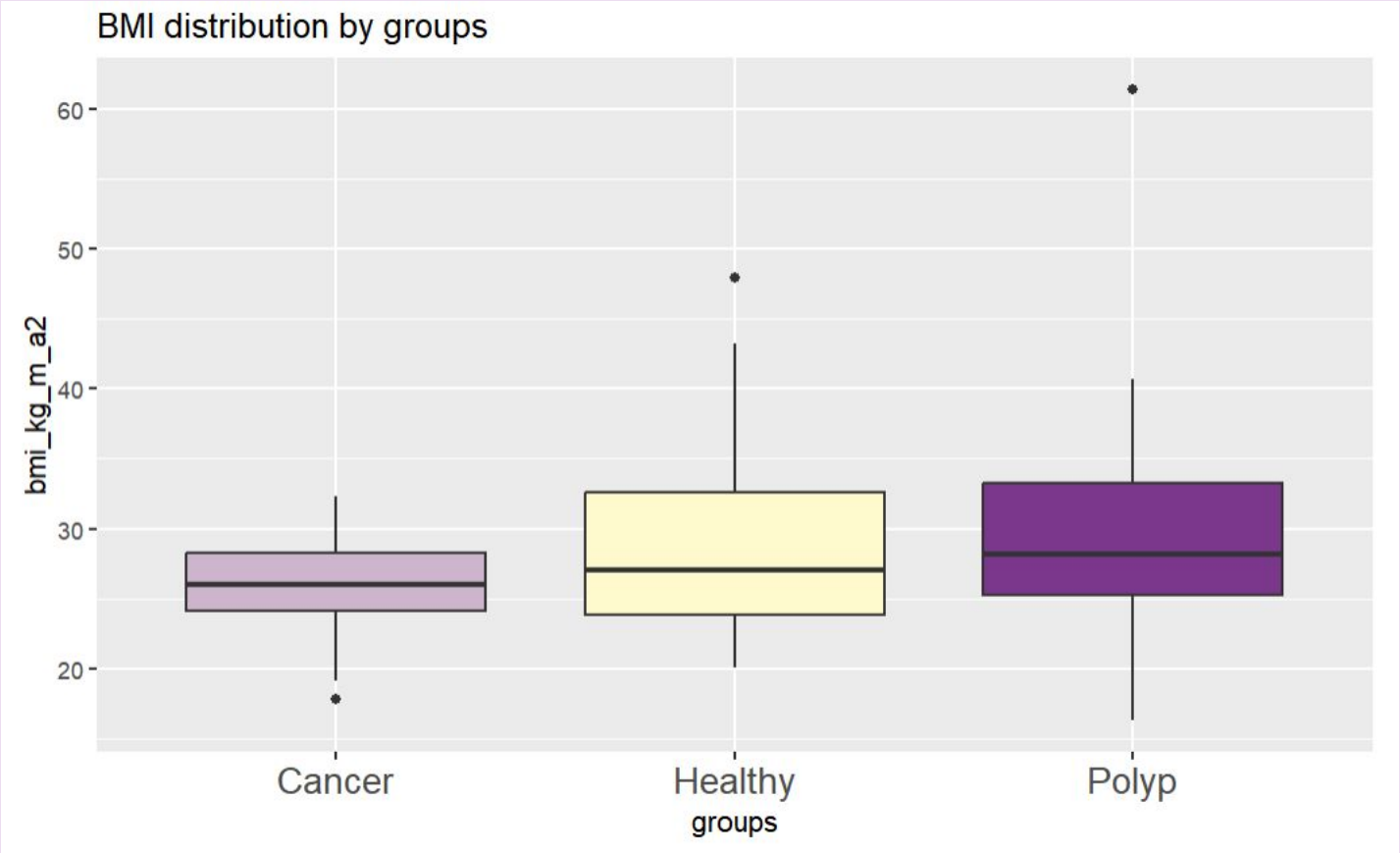
1. Fecal Immunochemical Test (FIT)
2. Guaiac Fecal Occult Blood Test (gFOBT)
3. Fecal DNA testing
4. CT-Colonography
5. Colonoscopy

Study Design

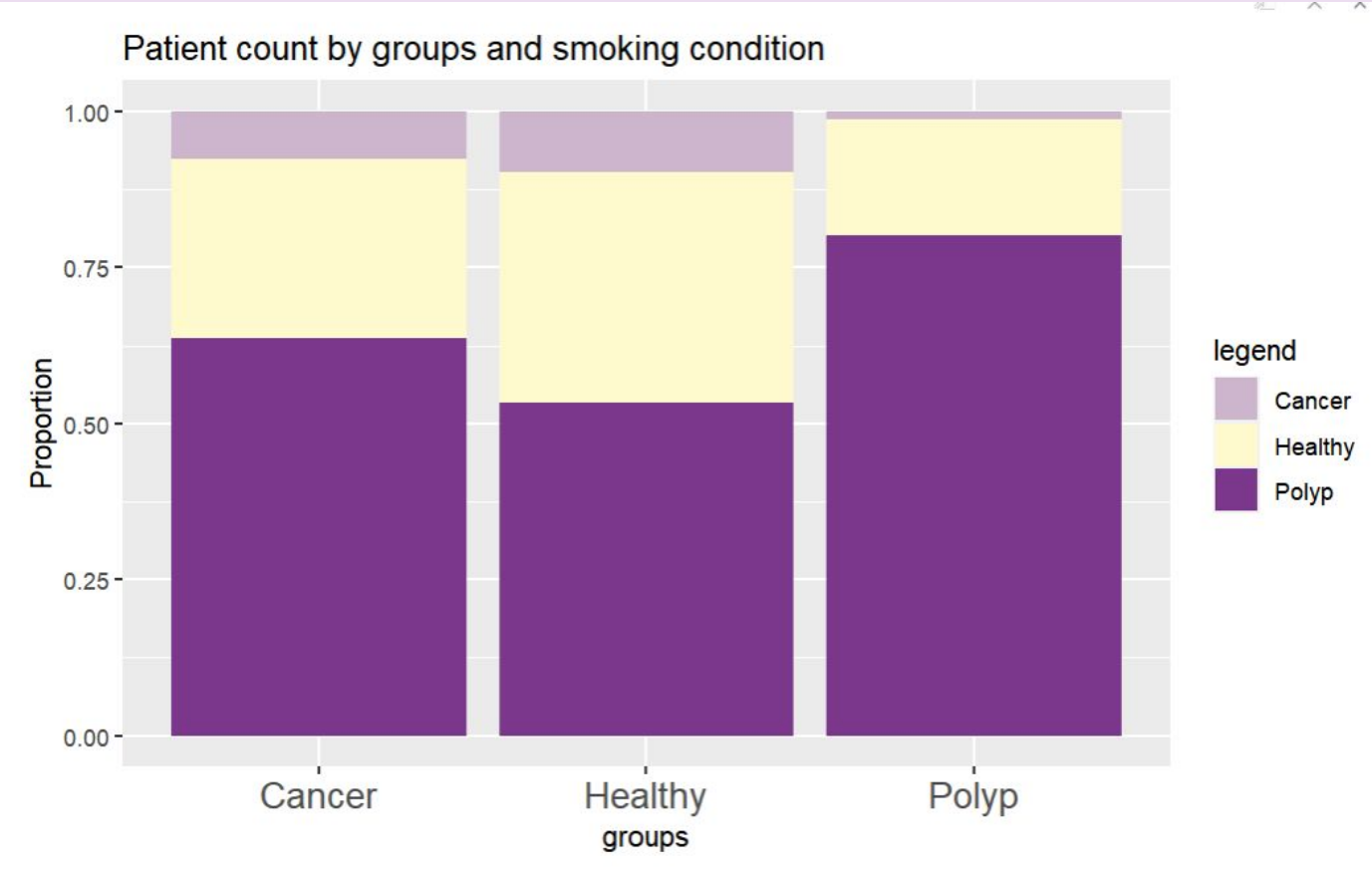
- Participants from Purdue University undergoing treatment or colonoscopy
  - Cancer, Healthy, Polyp
- Age and gender matched
- Measured metabolites in their blood



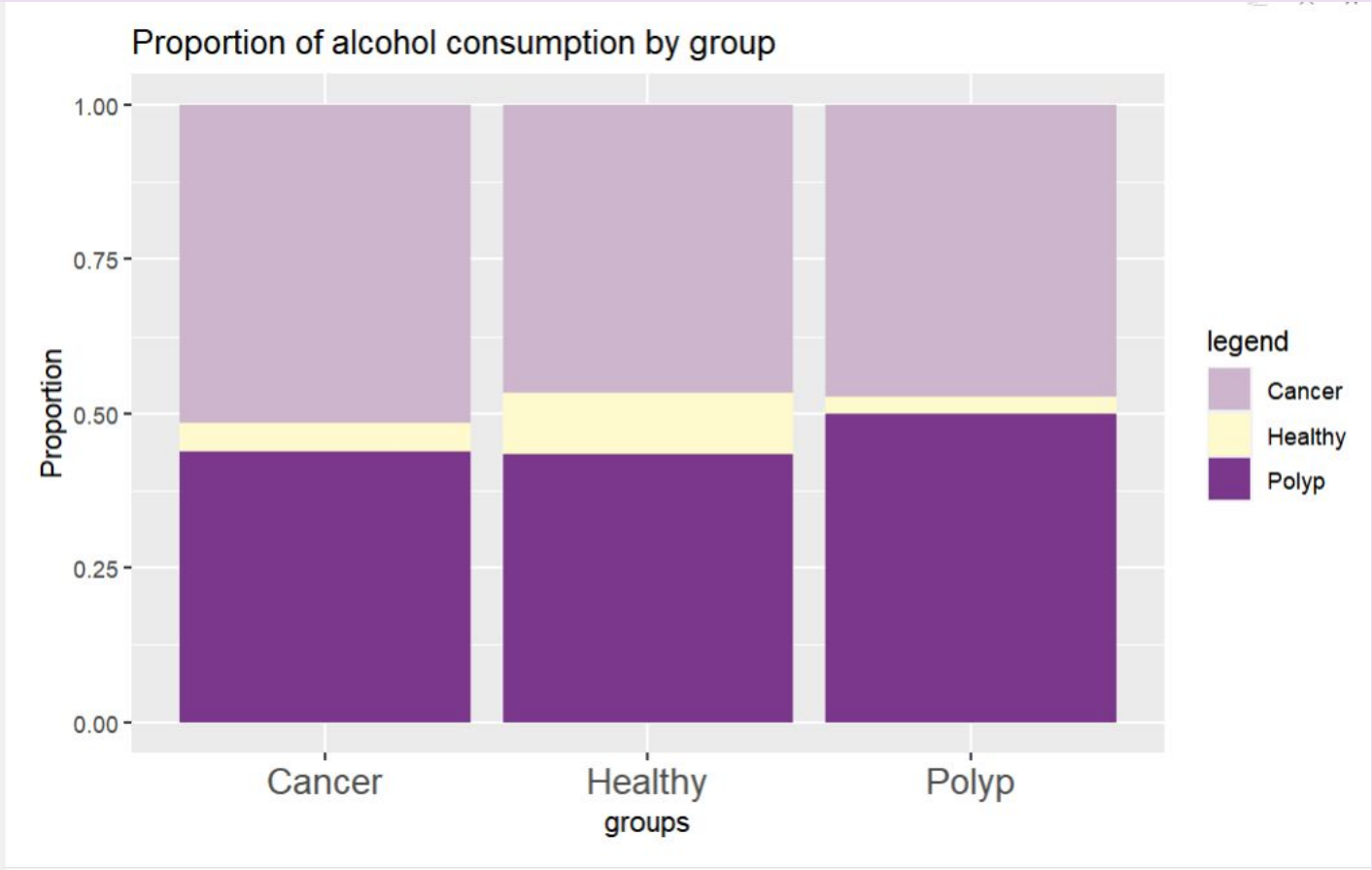
Summary Statistics: BMI

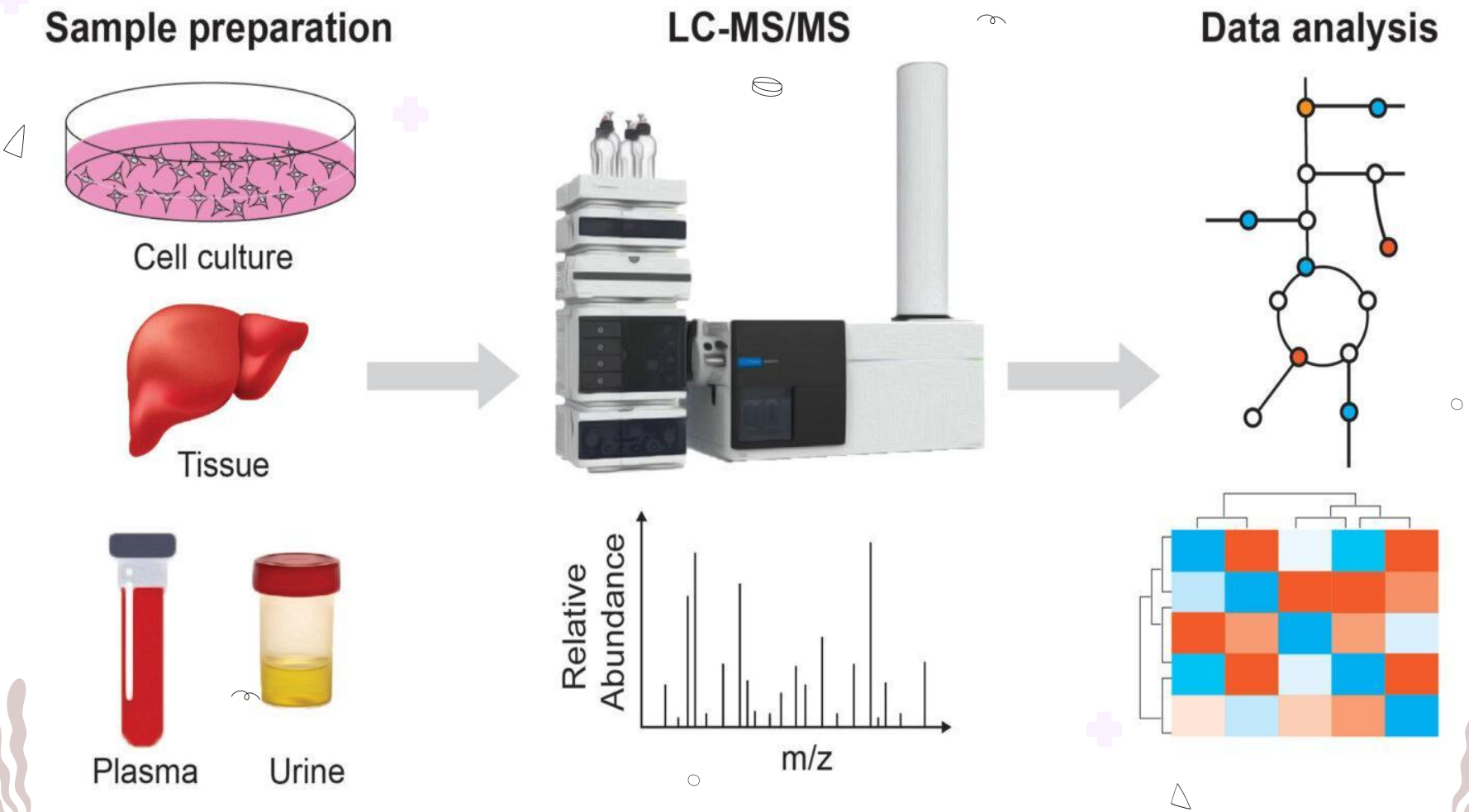


Summary Statistics: Smoking



Summary Statistics: Alcohol




Sample preparation
LC-MS/MS
Data analysis
Cell culture
Tissue
Plasma
Urine
Relative Abundance
m/z

MOTIVATION

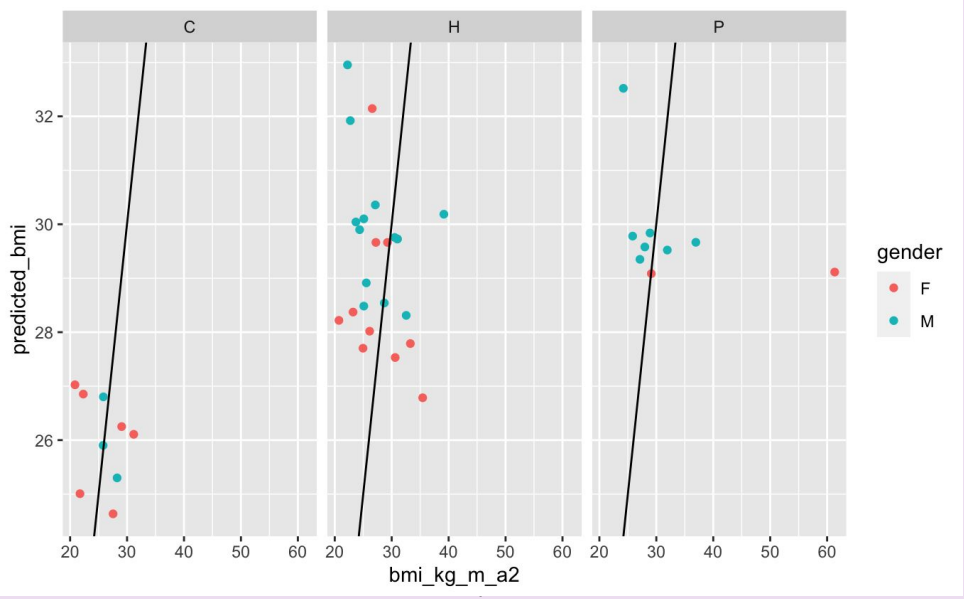
Data Wrangling: Missing Data

	MISSING	TOTAL
CANCER	29	64
POLYP	37	76
HEALTHY	1	84

Data Wrangling: Missing Data

X-Axis
Actual BMI

Y-Axis
Predicted BMI

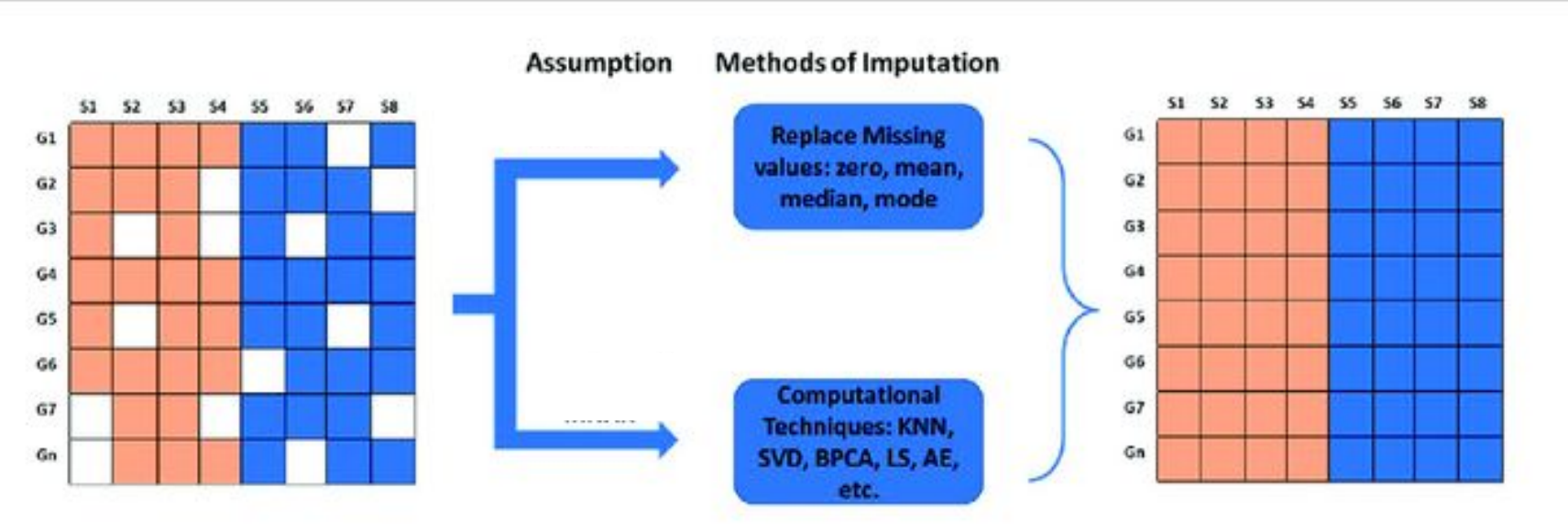


Graphs of the test set's actual versus predicted bmi's split by group.
Colored by gender.

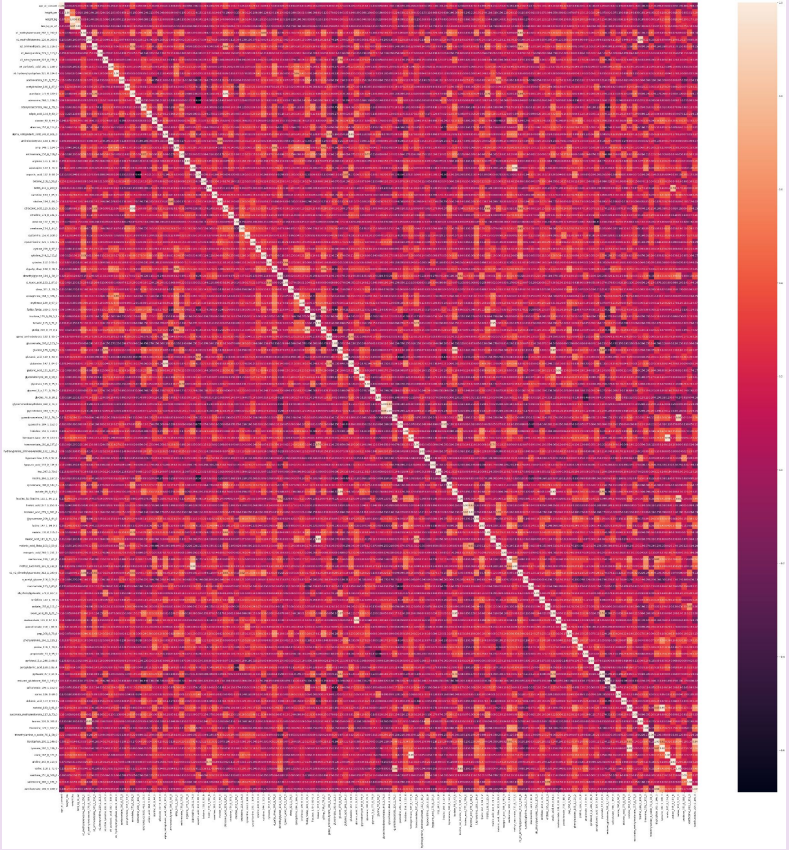
Black Line
Y = x to represent correct prediction

Three Graphs
Three groups: Cancer, Healthy, Polyp

Missing Data: Outliers



Correlation Matrix



HIGHLY CORRELATED (>.85)

glucose & lactate & oxalic acid

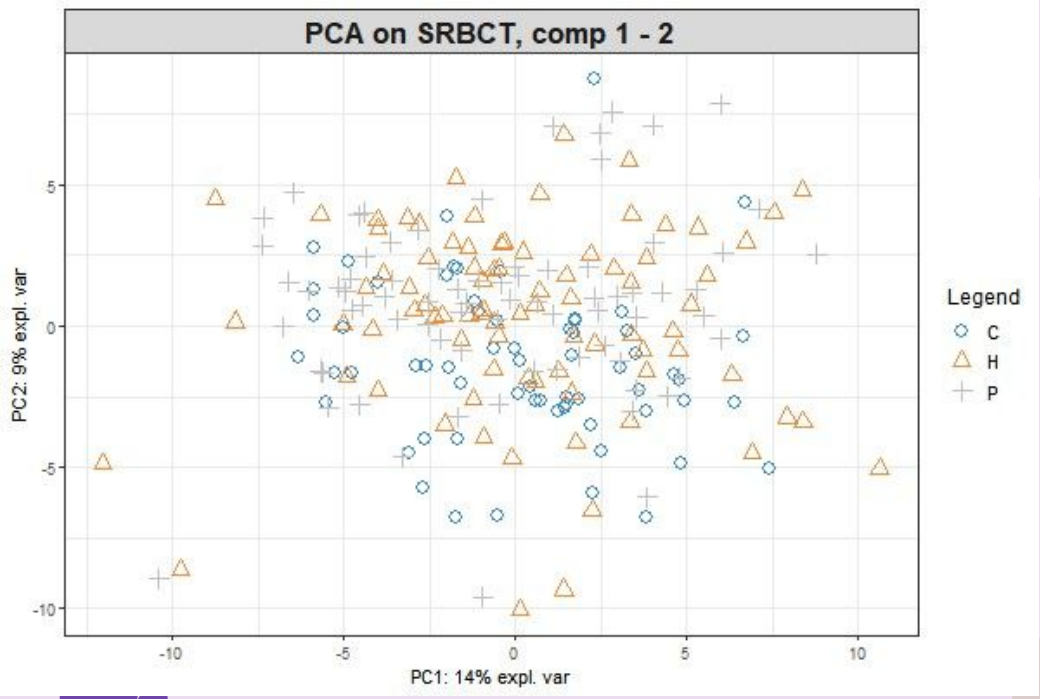
- weight kg & bmi
- methylhistamine & taurine
- aconitate & citraconic acid
- alpha ketoglutaric acid & glutamine
- asparagine & methyl succinate
- fumaric & maleic acid
- glutaric acid & oxaloacetate
- guanidinoacetate & valine
- guanosine & inosine
- homogentisate & urate
- leucine isoleucine & valine
- sorbitol & tyrosine

PCA

- Dimensionality reduction
- No meaningful clusters

- Recommends standardization

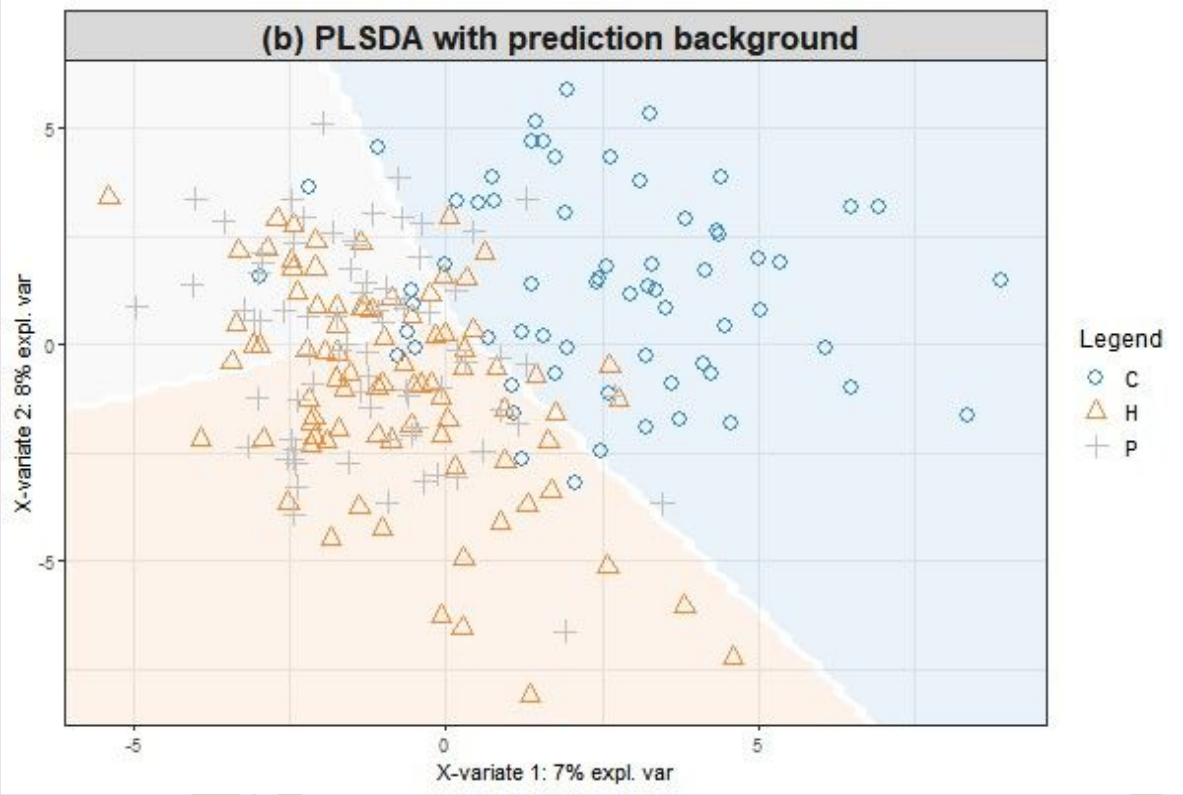
- 56 Principal Components explain 95.48% of the variance



PLS-DA

- Partial Least Squares Discriminant Analysis

- Finds principle components that best explains covariance between groups

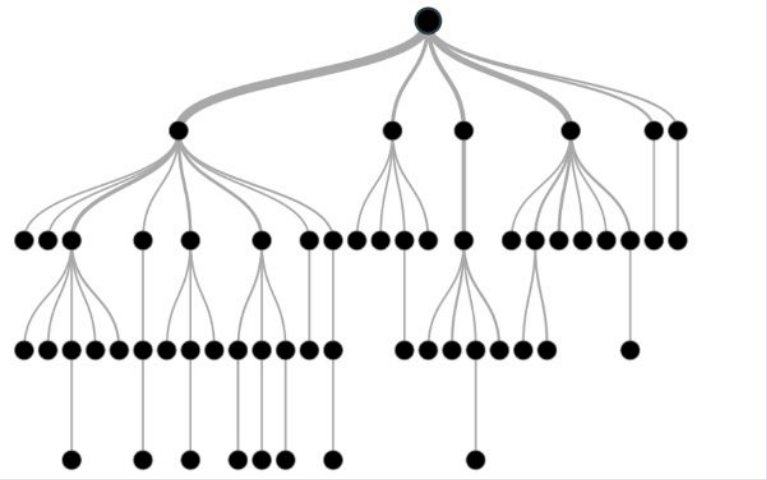


Models

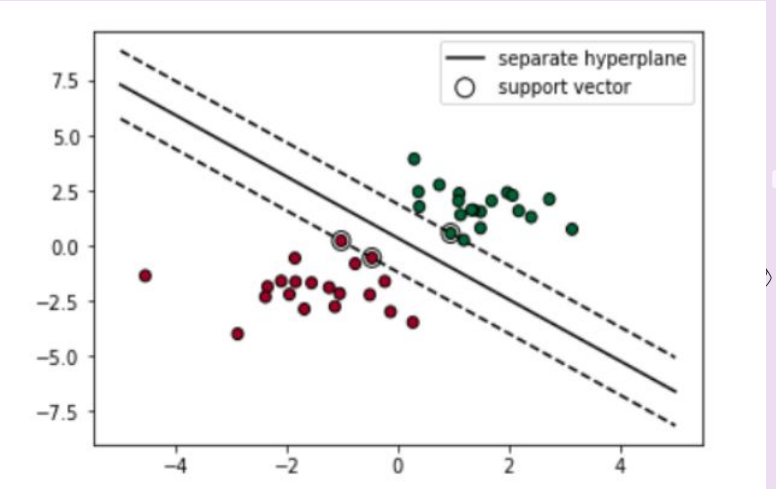
What Models We Used and Results

Classification Models

Trees:



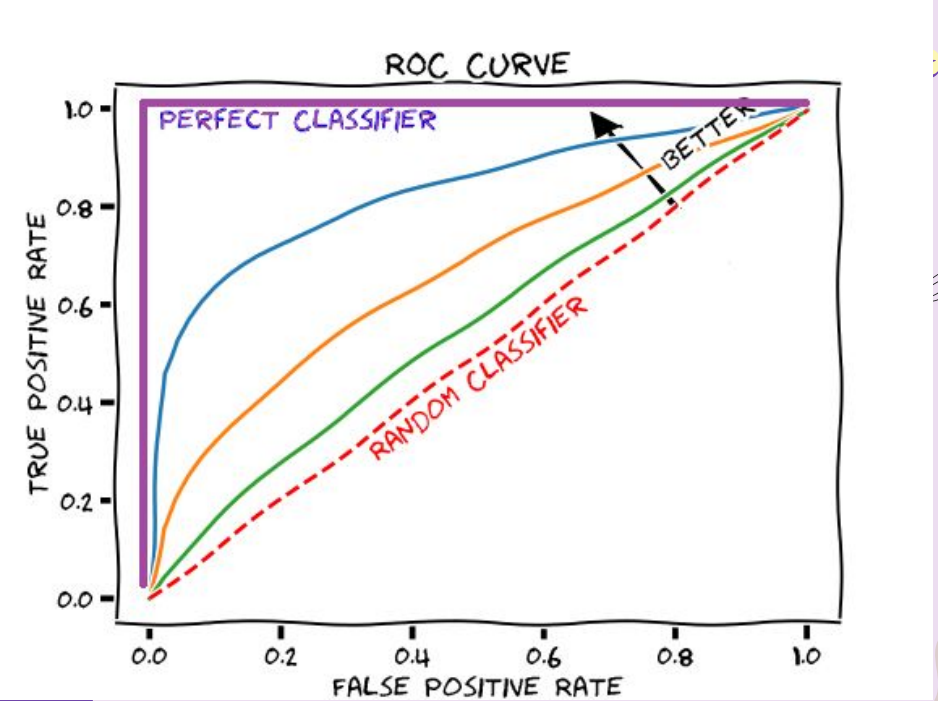
Support Vector Machines (SVM):



- Kth Nearest Neighbor (KNN): good at making predictions
- Naive Bayes (NB): quick works well with small datasets
- Linear Discriminant Analysis (LDA): can handle multicollinearity

Area Under the Curve (AUC)

- Our measurement for “best model”
- Most popular metric in diagnostic testing
- Considers false positives and false negatives
- Better than accuracy



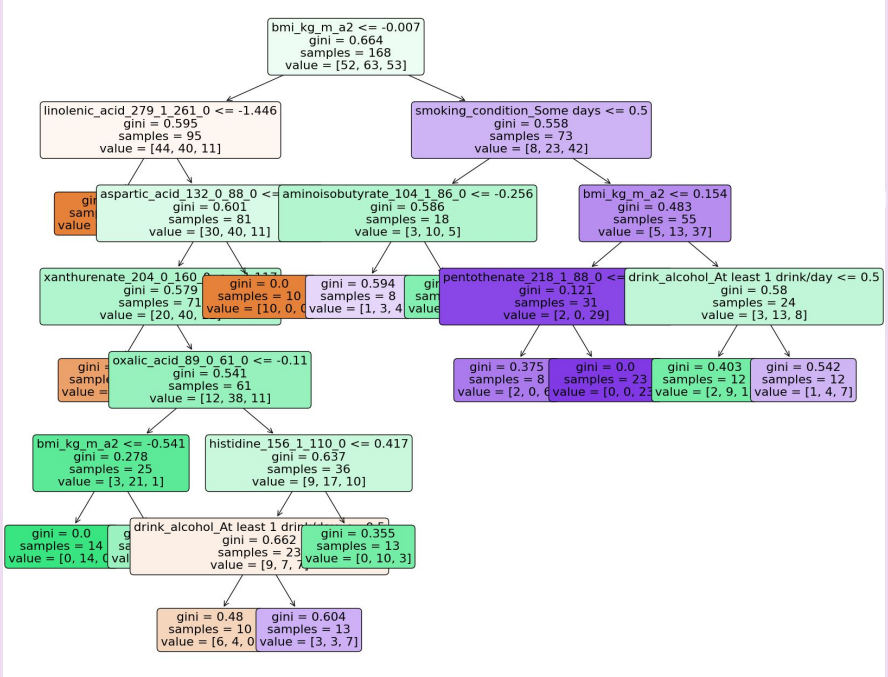
Decision Trees

- Randomly dropped a variable with |correlation| > .85

- only principal components ≥ .01 (22)

- Histidine
- Oxalic_acid
- Aspartic_acid
- Linolenic_acid
- Pantothenate
- Xanthurenate
- Aminoisobutyrate
- smoking_condition_Some days
- drink_alcohol_At least 1 drink/day
- BMI

- accuracy was about 66.07%



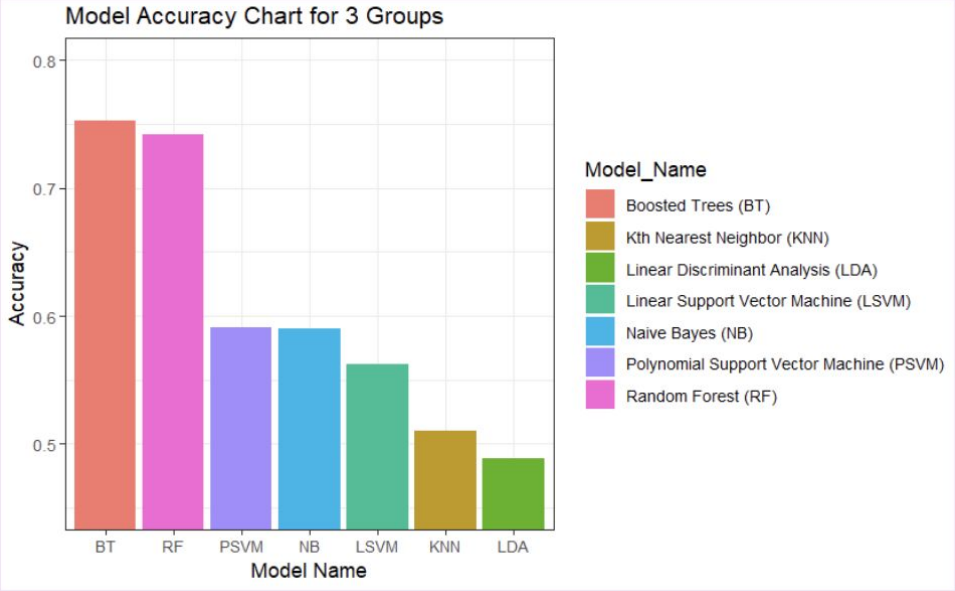
3 Class Model Results (ACC)

Boosted Trees

.753088

Random Forest

.7424281



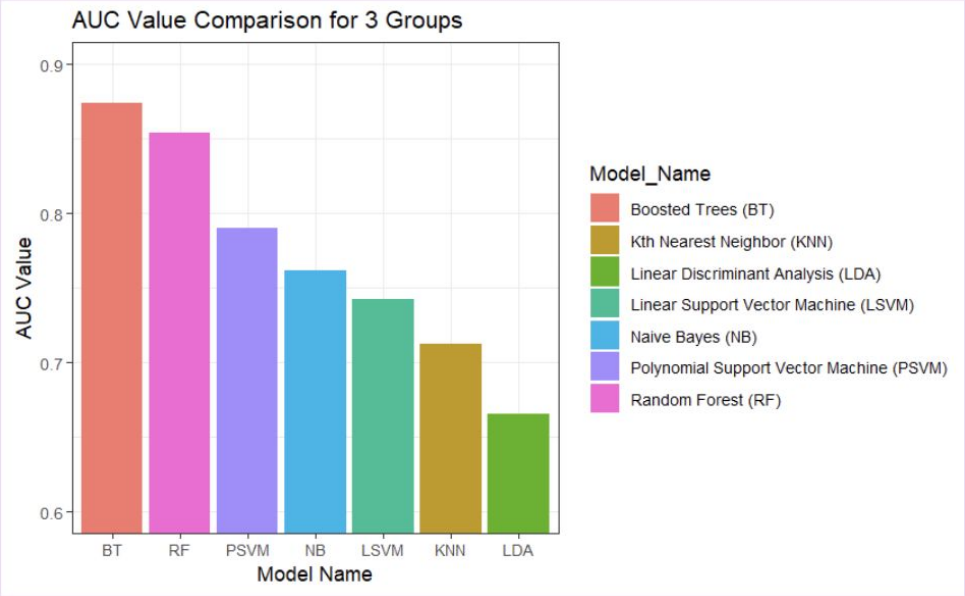
Accuracy

of true positives and true negatives

Not good

Doesn't show enough about the data

3 Class Model Results (AUC)



Area Under the Curve

Under the ROC curve

Why

Considers true positive (sensitivity) and false positive (1 - specificity) rates

Boosted Trees

.8741

Random Forest

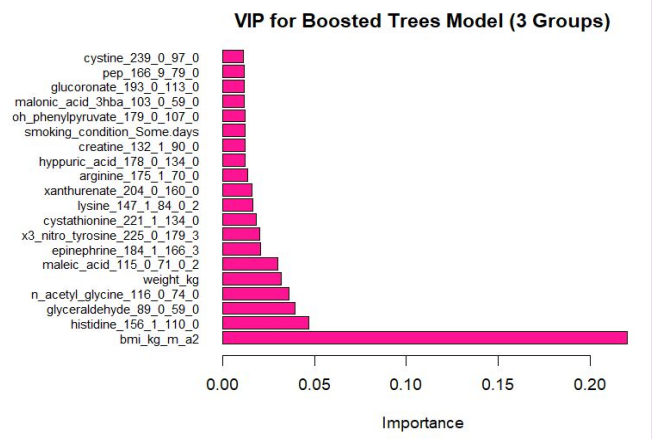
.8541

Important Variables in 3 Class Models

Boosted Trees

Random Forest

Graph

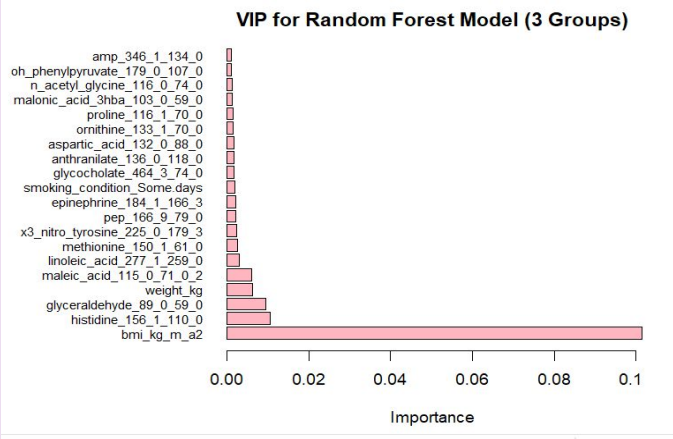


OVERLAP

Different for Both

- Smoking condition
- Maleic Acid
- NitroTyrosine
- Epinephrine
- PEP
- Malonic Acid
- N Acetyl Glycine

Graph



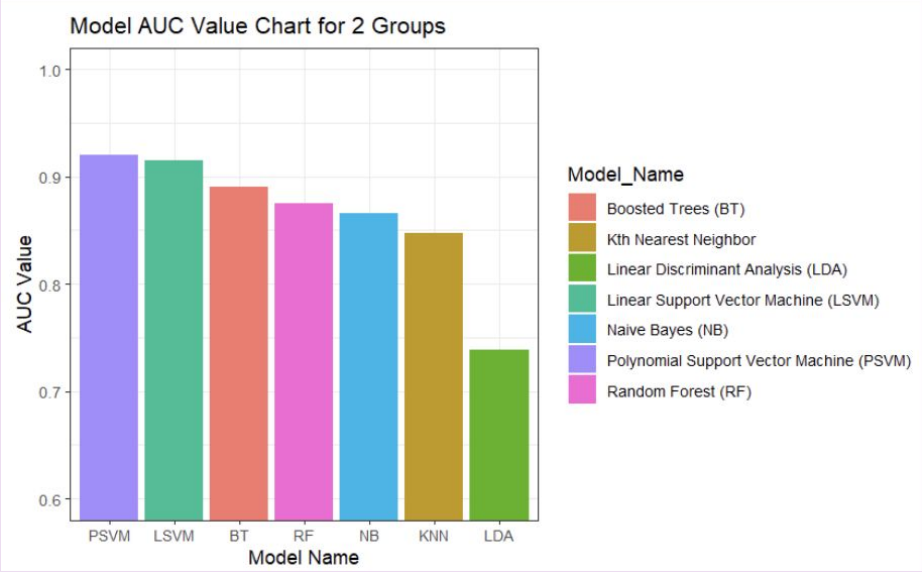
2 Class Model

- PLS-DA → 2 Class Model
- See how model determines a polyp patient
- Train and test set created on Cancer & Healthy subset

Cancer **Healthy**

Polyp

2 Class Model Results (AUC)



Polynomial Support Vector Machine
.9204496

Linear Support Vector Machine
.9151449

Boosted Trees
.89

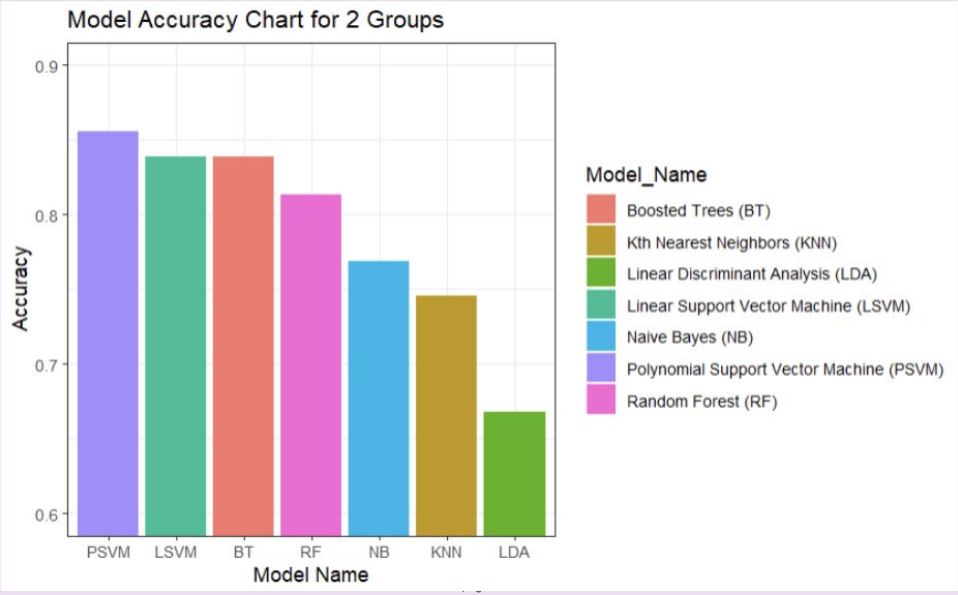
Random Forest
.8754

Higher AUC in 2 class model!

2 Class Model Results (Accuracy)

Polynomial Support Vector Machine
.8553768

Linear Support Vector Machine
.8383478



Boosted Trees
.8383188

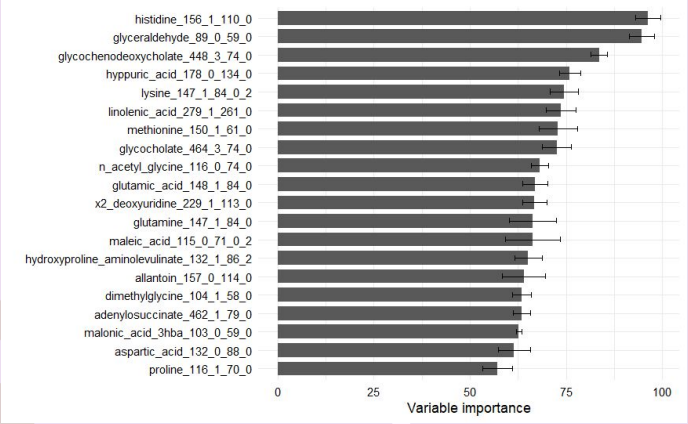
Random Forest
.8132899

Higher accuracy in 2 class model!

Important Variables in 2 Class Models

Polynomial SVM

Graph

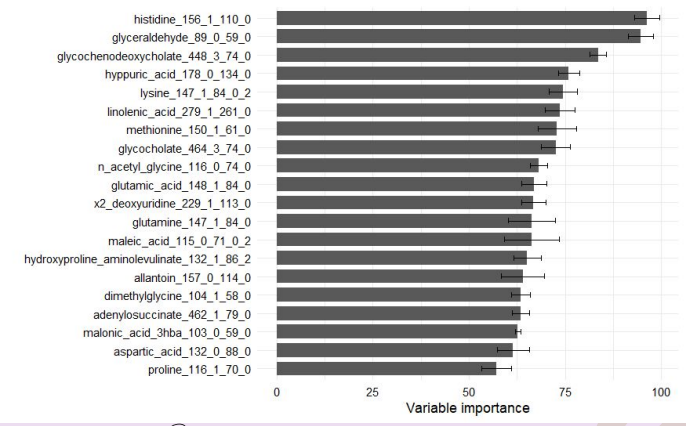


Most Important

1. Histidine
2. Glyceraldehyde

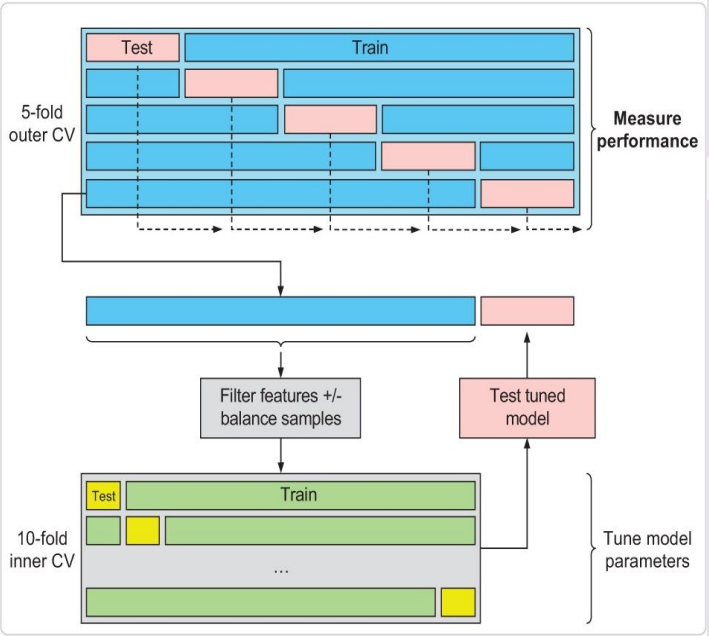
Linear SVM

Graph



Nested CV

- Cross-Validation (CV) is a technique used to evaluate the performance of a machine learning model.
- However, it has been demonstrated that filtering on the whole dataset creates a bias when determining accuracy of models
- Nested CV is CV nested within a CV
- Train Set of the outer fold becomes the data for the inner fold
- Outer Fold is for model evaluation
- Inner Fold is for hyperparameter tuning
- Reduces Bias and Overfitting
- Computationally Taxing



Nested CV Models (3 Group)

Models <chr>	AUC <dbl>	Accuracy <dbl>
Boosted Trees (BT)	0.8344	0.6786
Random Forest (RF)	0.8245	0.6607
Polynomial Support Vector Machine (PSVM)	0.7780	0.5804
Naive Bayes	0.7539	0.5402
Linear Support Vector Machine (LVSM)	0.7584	0.6071
Kth Nearest Neighbor(KNN)	0.7204	0.5179
Linear Discriminant Analysis (LDA)	0.7113	0.5089

*All models ran with 10 outer and 5 inner folds

Nested CV: Cancer vs. Healthy

Models <chr>	AUC <dbl>	Accuracy <dbl>
Boosted Trees (BT)	0.8875	0.8243
Random Forest (RF)	0.8277	0.7432
Polynomial Support Vector Machine (PSVM)	0.8558	0.7770
Naive Bayes	0.7874	0.7230
Linear Support Vector Machine (LVSM)	0.8186	0.7095
Kth Nearest Neighbor(KNN)	0.6816	0.6284
Linear Discriminant Analysis (LDA)	0.6151	0.6284

*All models ran with 10 outer and 5 inner folds

Polyp Prediction

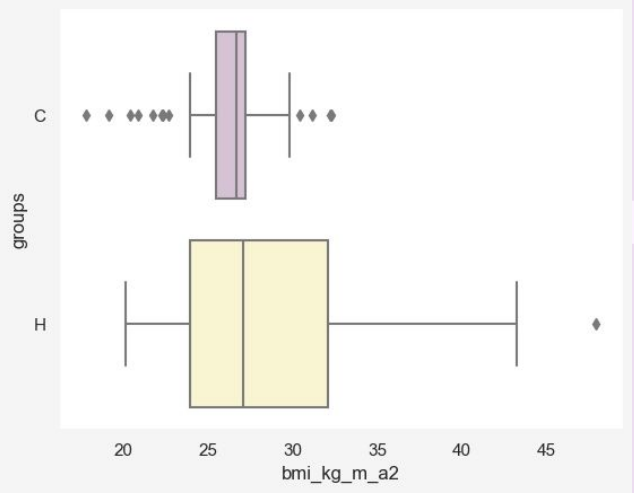
Models <chr>	C <dbl>	H <dbl>	CI_95 <chr>
Boosted Trees (BT)	20	56	(0.6232, 0.8313)
Random Forest (RF)	27	49	(0.5266, 0.7512)
Polynomial Support Vector Machine (PSVM)	12	64	(0.7404, 0.9157)
Naive Bayes	20	56	(0.6232, 0.8313)
Linear Support Vector Machine (LVSM)	14	62	(0.7103, 0.8955)
Kth Nearest Neighbor(KNN)	36	40	(0.4084, 0.6421)
Linear Discriminant Analysis (LDA)	14	62	(0.7103, 0.8955)

BMI: Healthy vs Cancer

- We found that an increase in BMI actually decreases the odds ratio of CRC. Why?
- High BMI is correlated with CRC
- Recent weight loss is also strongly correlated with CRC

Odds Ratio

- Estimate:
  - 44% decrease
- Confidence Interval:
  - [-65%, -18%]
- p-Value
  - 0.005
- Marginal



	Log Mean	Standard Deviation
Cancer	26.21	2.66
Healthy	28.75	6.39

Smoking & Alcohol

Smoking

- Smoking is correlated with colorectal cancer (CRC), via metabolic changes

Some smoking:
- Estimate: 40% decrease
- 95% CI: [-71%, 21%]
- p-value: 0.17

Smoking everyday
- Estimate: 10% increase
- 95% CI: [-74%, 236%]
- p-value: 0.87

Alcohol

- Alcohol consumption is correlated with CRC by causing changes in genetic abnormalities, epigenetic dysregulation, cell signaling, and changing the tumor microenvironment

Some drinking:
- Estimate: 127% increase
- 95% CI: [-38%, 994%]
- p-value: 0.25

One drink a day
- Estimate: 196% increase
- 95% CI: [-29%, 1100%]
- p-value: 0.2

Warburg Effect

What is it?

Glycolysis bypasses Krebs cycle in favor of lactic acid fermentation

Correlation

Glucose and Lactate have a .886 correlation coefficient

Krebs and ETC

producing 32 ATP VS glycolysis' yield of 2 ATP

Less Energy

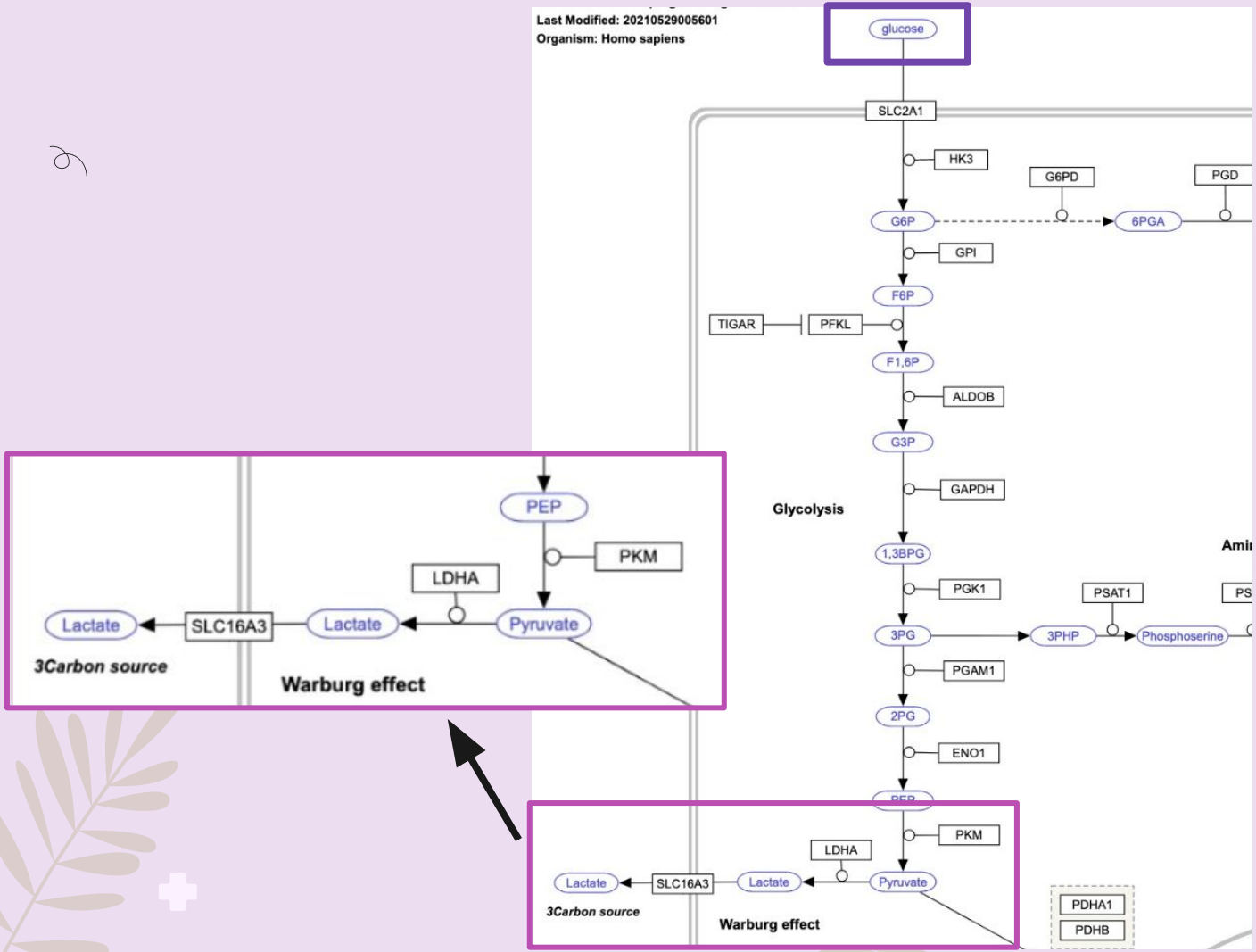
greater rate of glucose metabolism

Why?

competition, cell proliferation, or to continue glycolysis

Risk Factor

Higher levels of lactate could indicate higher risk for CRC.

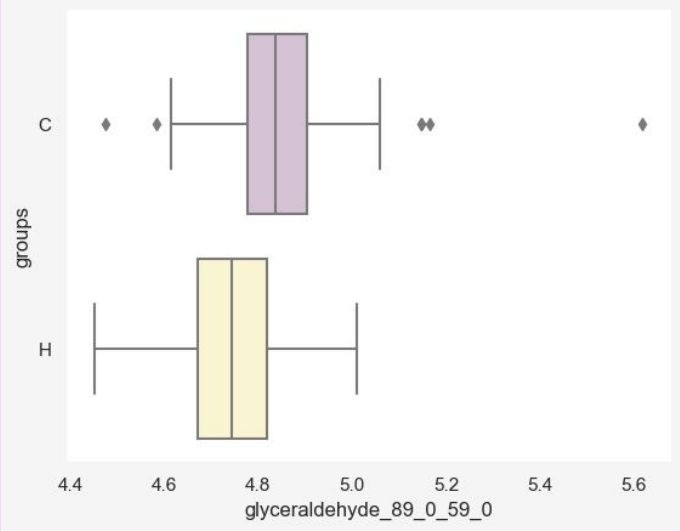

Last Modified: 20210529005601
Organism: Homo sapiens
glucose
SLC2A1
HK3
G6PD
PGD
G6P
6PGA
GPI
F6P
TIGAR
PFKL
F1,6P
ALDOB
G3P
GAPDH
Glycolysis
1,3BPG
PGK1
PSAT1
3PG
3PHP
Phosphoserine
PGAM1
2PG
ENO1
PEP
PKM
LDHA
Lactate
SLC16A3
Lactate
Pyruvate
3Carbon source
Warburg effect
PDHA1
PDHB

Glyceraldehyde

- Glyceraldehyde is needed to produce intermediates in glycolysis

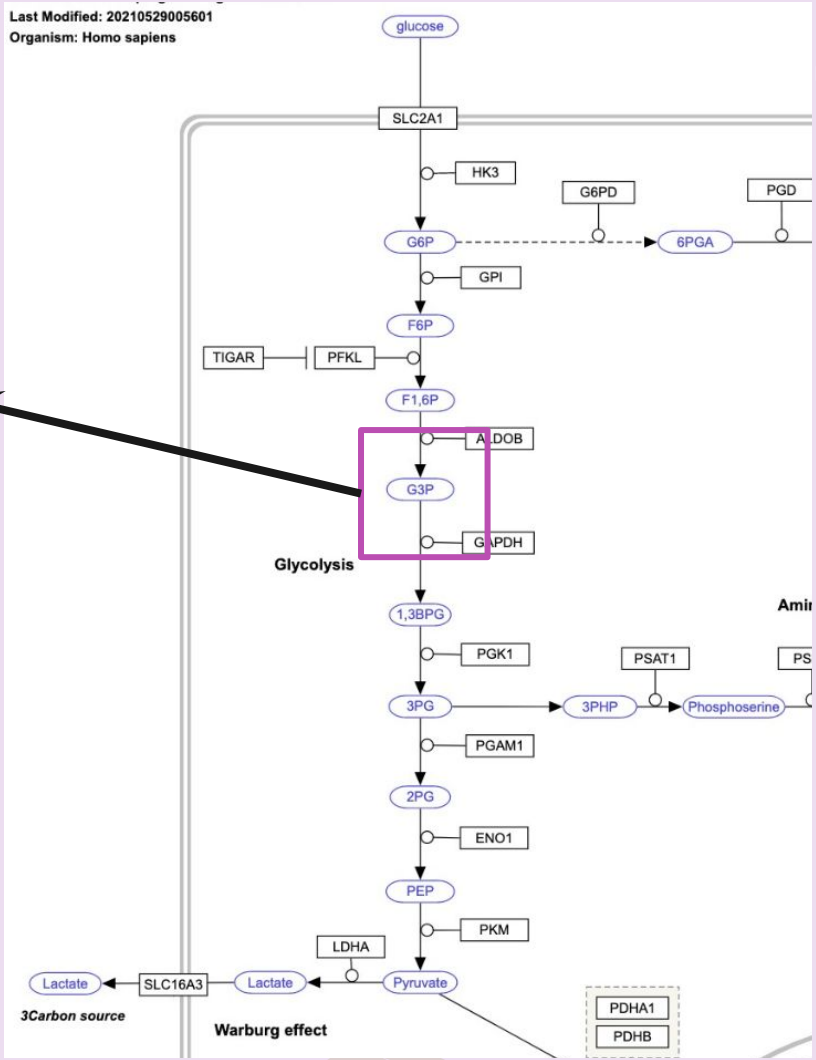
Odds Ratio

- Estimate:
  - 99% increase per unit increase in Glyceraldehyde
- Confidence Interval:
  - [52%, 155%]
- p-Value
  - 0.0001
- Marginal



	Log Mean	Standard Deviation
Cancer	4.85	0.16
Healthy	4.74	0.12

the phosphate of glyceraldehyde: glyceraldehyde 3-phosphate

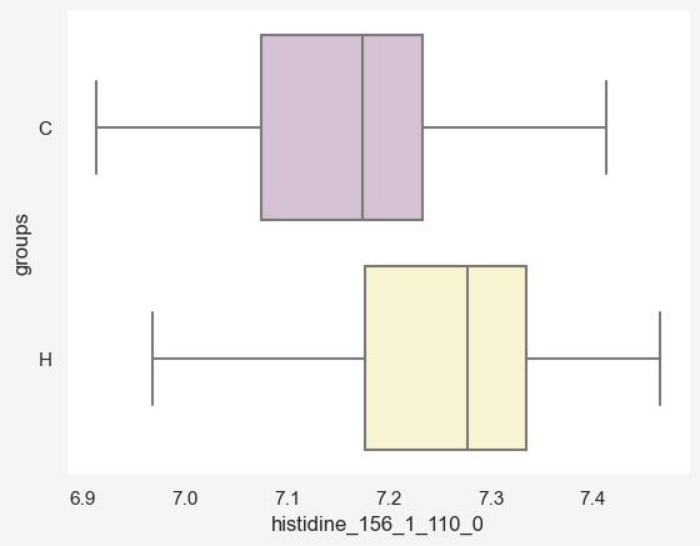


Histidine

- Amino acid levels are generally inversely related with CRC risk

Odds Ratio

- Estimate:
  - 52% decrease per unit increase in Histidine
- Confidence Interval:
  - [-67%, -34%]
- p-Value
  - 0.00003



	Log Mean	Standard Deviation
Cancer	7.16	0.11
Healthy	7.25	0.12

Diagnostic Tests	Polyp Detection Accuracy	Cancer Detection Accuracy	Sensitivity	Specificity
Fecal immunochemical test (FIT)	51%	95%	74%	95%
gFOBT	64%	58%	100%	60%
Fecal DNA testing	46%	42%	92%	88%
CT-Colonography	80%	89%	96%	80%
Colonoscopy	96%	95%	91%	73%
Boosted Trees	74%	85%	77%	88%
Random Forest	70%	85%	77%	88%

Future Studies

Future studies and Conclusion

Future Studies

Glutaminolysis

- As a result of this process, GLS1 and GDH are both upregulated.[

LONGITUDINAL

Lipid Biosynthesis

- upregulate de novo lipid biogenesis and cholesterol synthesis pathways
- which in turn upregulates glycolysis and mitochondrial respiration

One Carbon Metabolism

- Phosphoglycerate dehydrogenase (PHGDH), which is tasked with resupplying 1CM with 3-phosphoglycerate through the serine biosynthetic pathway, is another metabolite which is upregulated due to CRC.[

A Special Thanks To:

Professor Zhaoxia Yu

Professor Min Zhang

Thanasi Bakis

Thank You!