

# Cancer Gene Expression Classification

*<https://github.com/tamarahorne/Springboard/tree/main/Capstone%20Project%203>*

Tamara Horne | May 2023| Springboard—DSC Capstone Project 3



# Introduction

## The Value of Predicting Cancer Type from Gene Expression Data

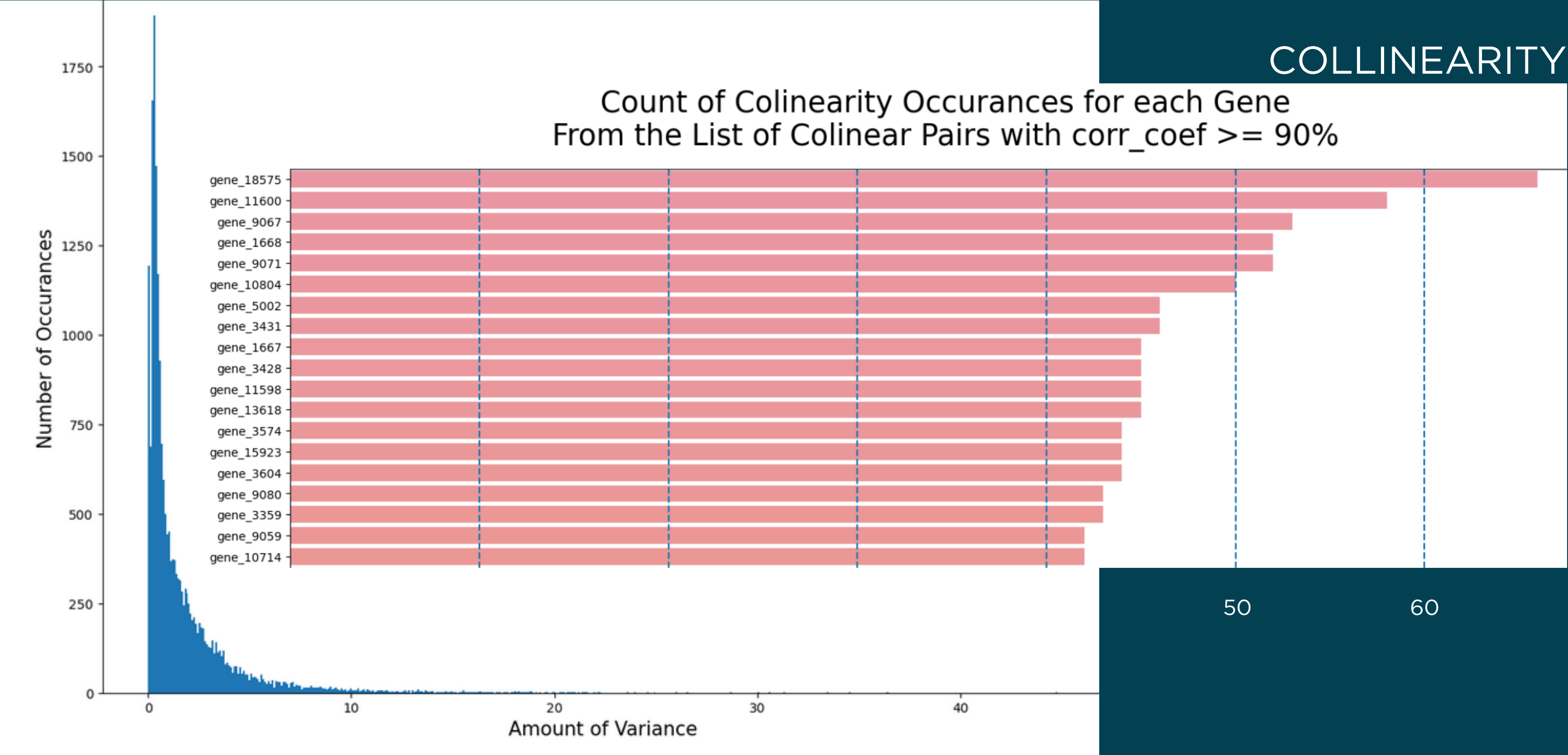
# THE DATA

- subset of the RNA-Seq (HiSeq) PANCAN data set
- samples from five cancer types
  - BRCA (breast)
  - KIRC (kidney)
  - COAD (colon)
  - LUAD (lung)
  - PRAD (prostate)
- 801 rows; 20531 columns
- one row represents one sample
- one column contains a value for the gene expression for one gene
- columns have dummy names (gene\_XX)
- no missing values

# Concerns

## VARIANCE DISTRIBUTION

## COLLINEARITY

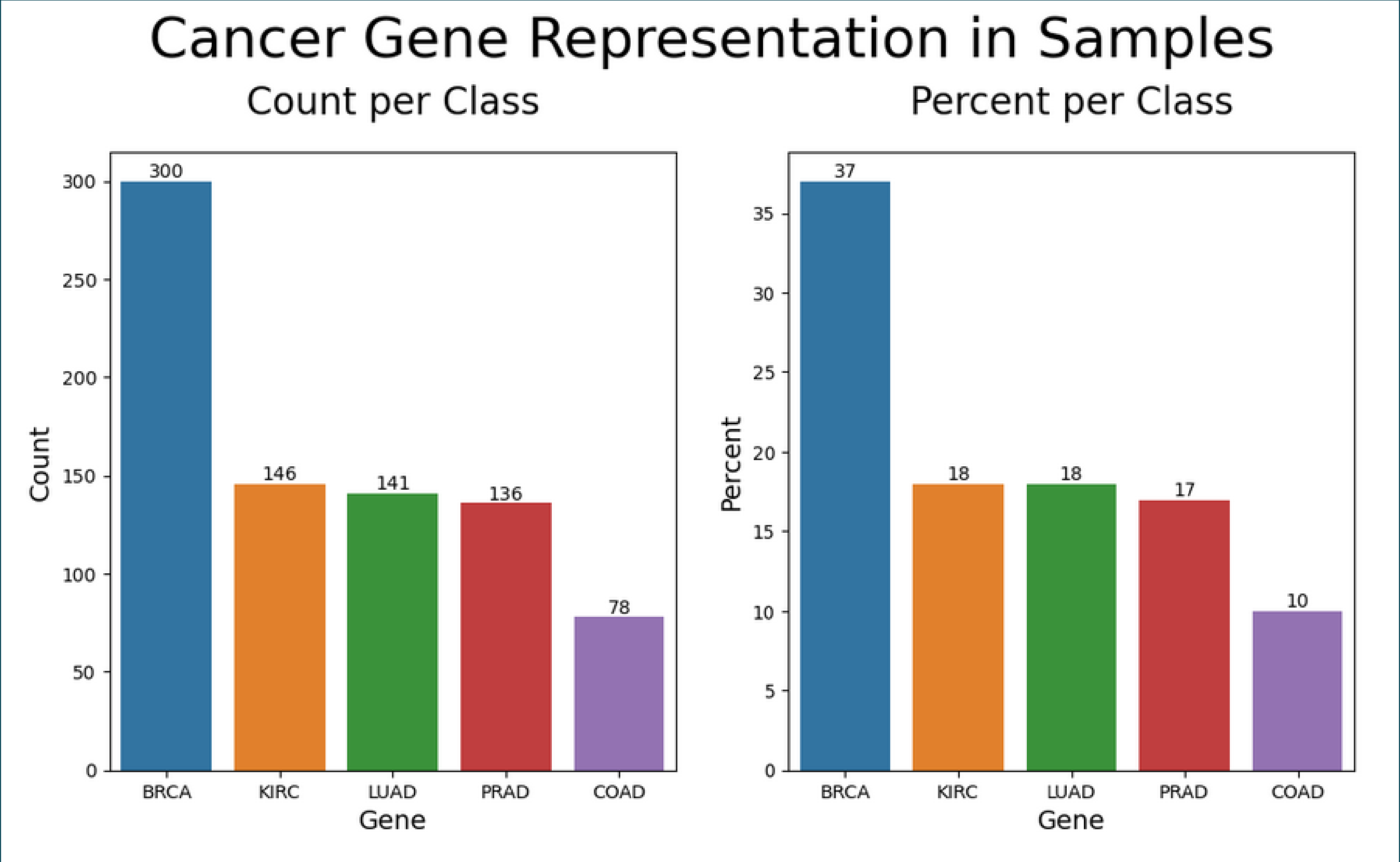


50

60

# Concerns

BALANCE



# Decisions

VARIANCE

**PCA**

COLLINEARITY

**PCA**

BALANCE

**Stratify**



# Baseline Modeling

with  
Logistic Regression

## First Steps:

Fixing the Landing Dimension  
(Rule of Thumb)

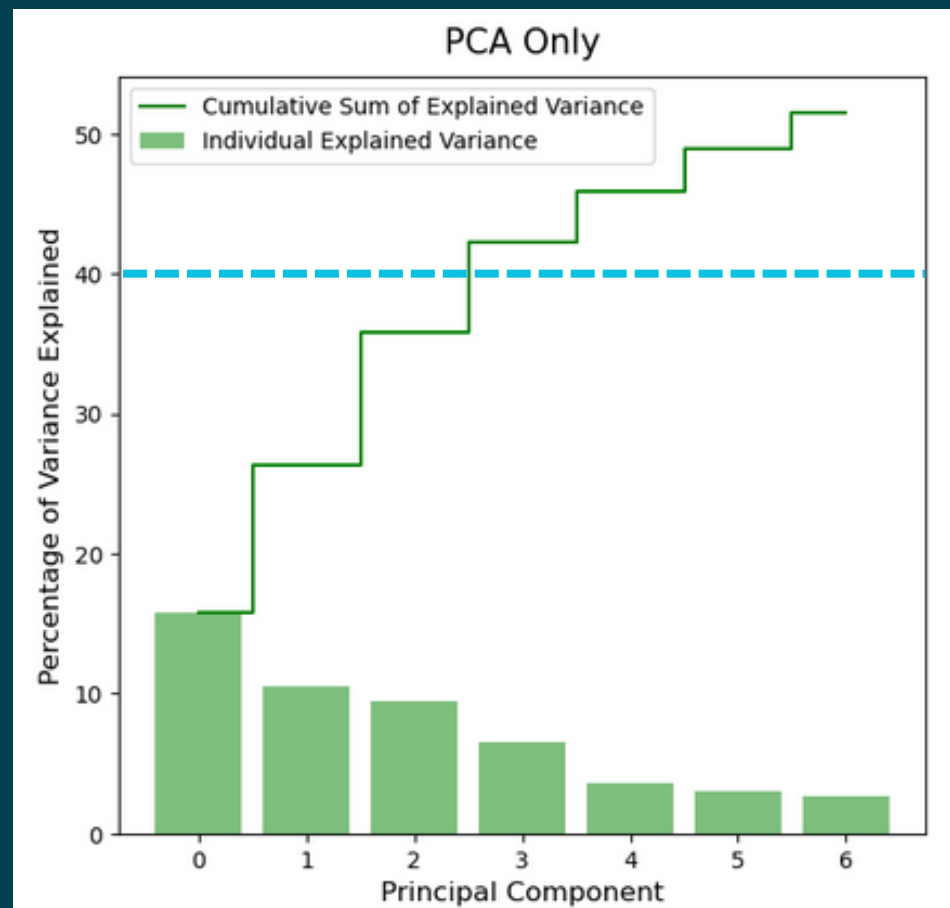


Determine if Using a Scaler in  
the Pipeline with PCA is Wise

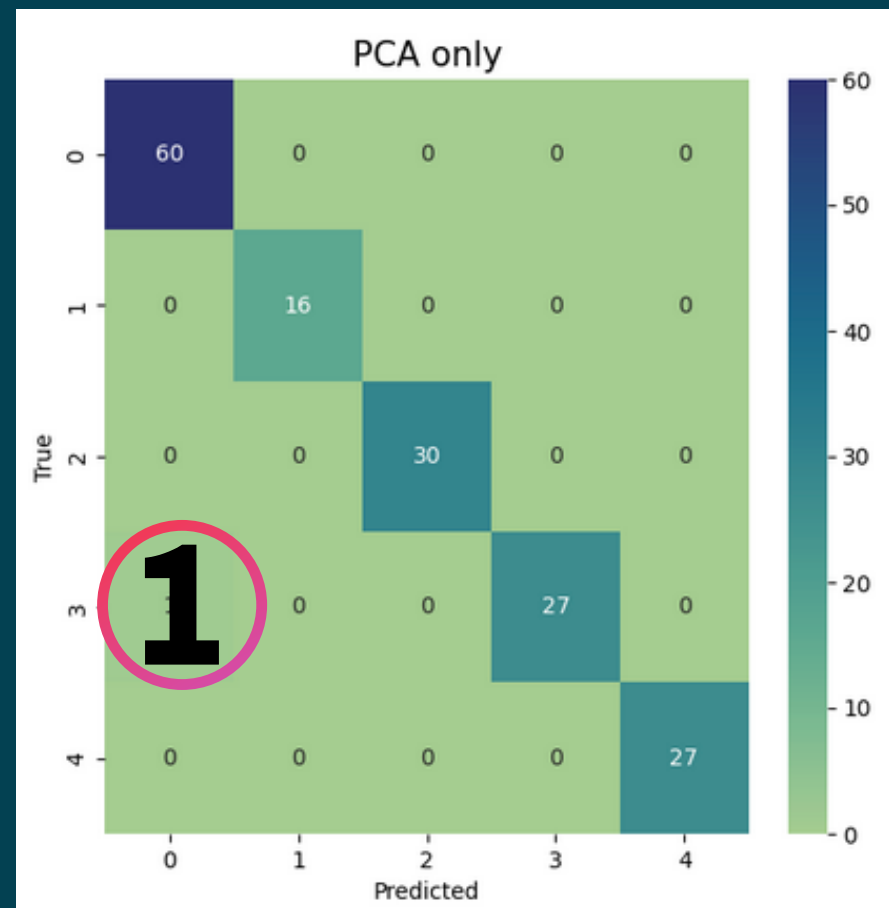


Decide if Balancing of Classes  
is Needed

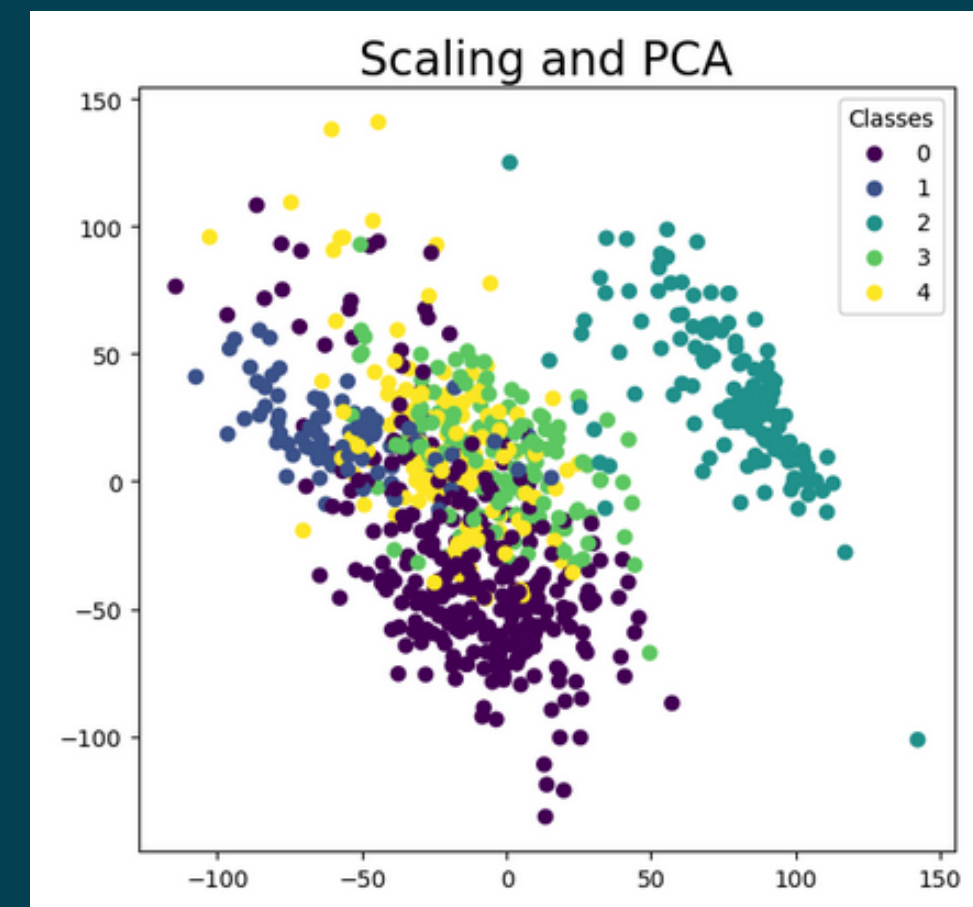
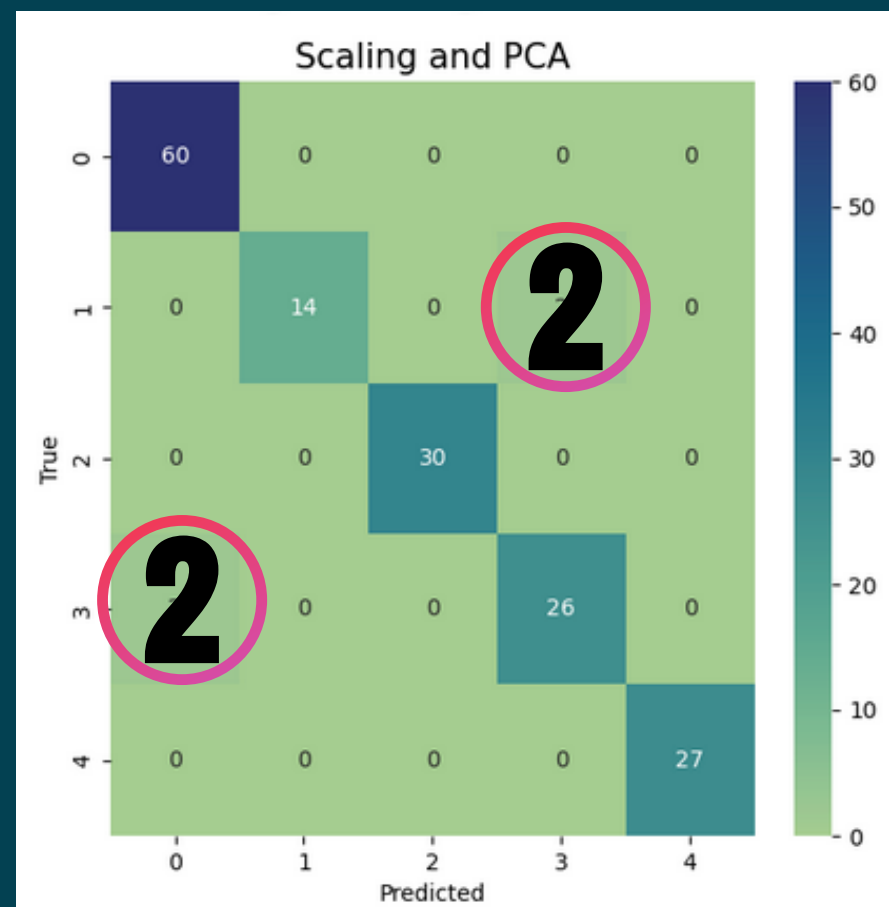
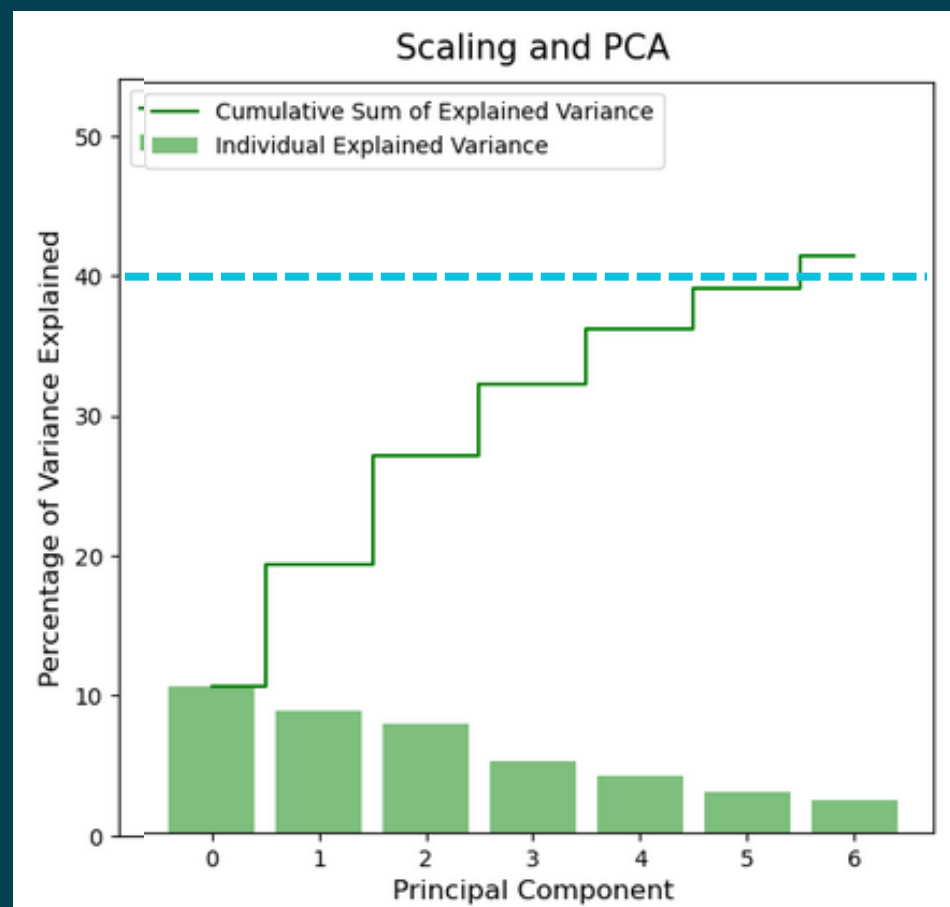
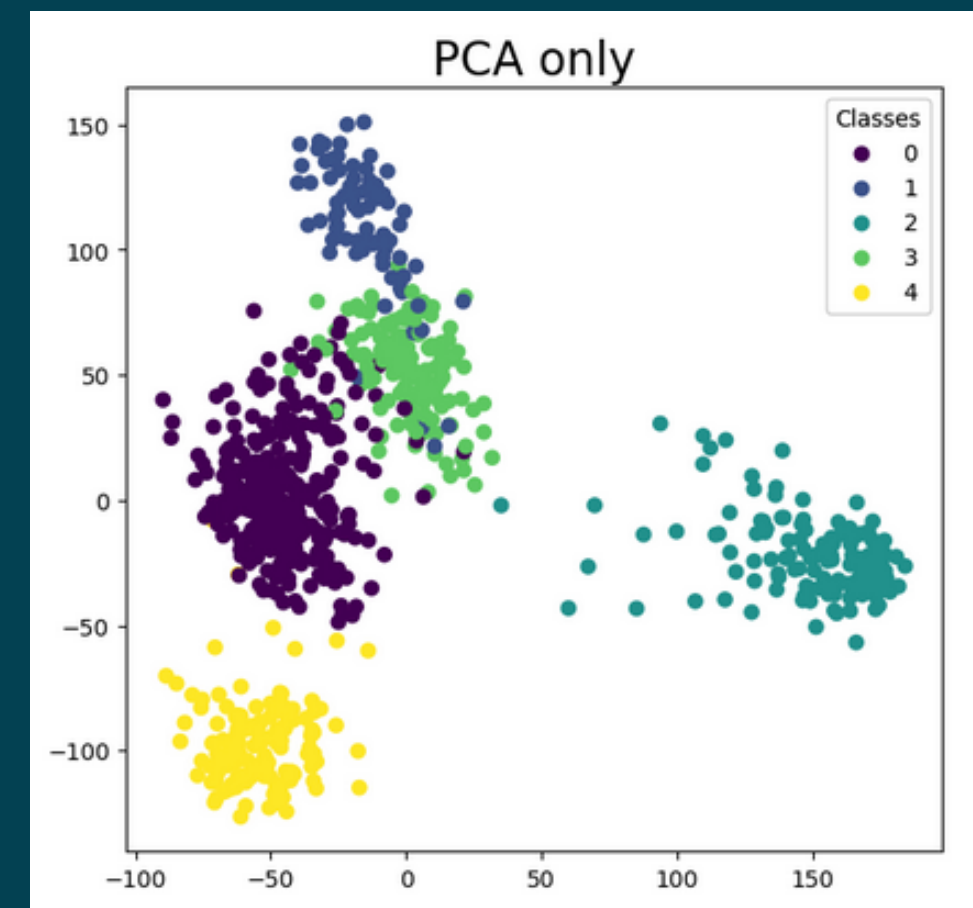
## Percentage of Explained Variance



## Confusion Matrix



## First Two Principal Components





# Baseline Modeling

with  
Logistic Regression

## Decisions:

Using the Rule of Thumb for  
Fixing the Landing Dimensions is  
Sufficient  
(Seven Principal Components)



A Scaler Should Not be Used



Balancing of Classes is Not  
Needed


# Baseline Modeling

with

## Logistic Regression

Scores for each fold using training data  
(PCA, Logistic Regression, StratifiedKFold)


```
Fold number: 0, F1_macro_avg: 0.991
Fold number: 1, F1_macro_avg: 1.000
Fold number: 2, F1_macro_avg: 1.000
Fold number: 3, F1_macro_avg: 0.993
Fold number: 4, F1_macro_avg: 0.991
```



```
Train set classification report:
              precision    recall  f1-score   support

    0           1.00         1.00         1.00        240
    1           1.00         1.00         1.00         62
    2           1.00         1.00         1.00        116
    3           1.00         1.00         1.00        113
    4           1.00         1.00         1.00        109

 accuracy          1.00         1.00         1.00        640
 macro avg          1.00         1.00         1.00        640
 weighted avg       1.00         1.00         1.00        640
```



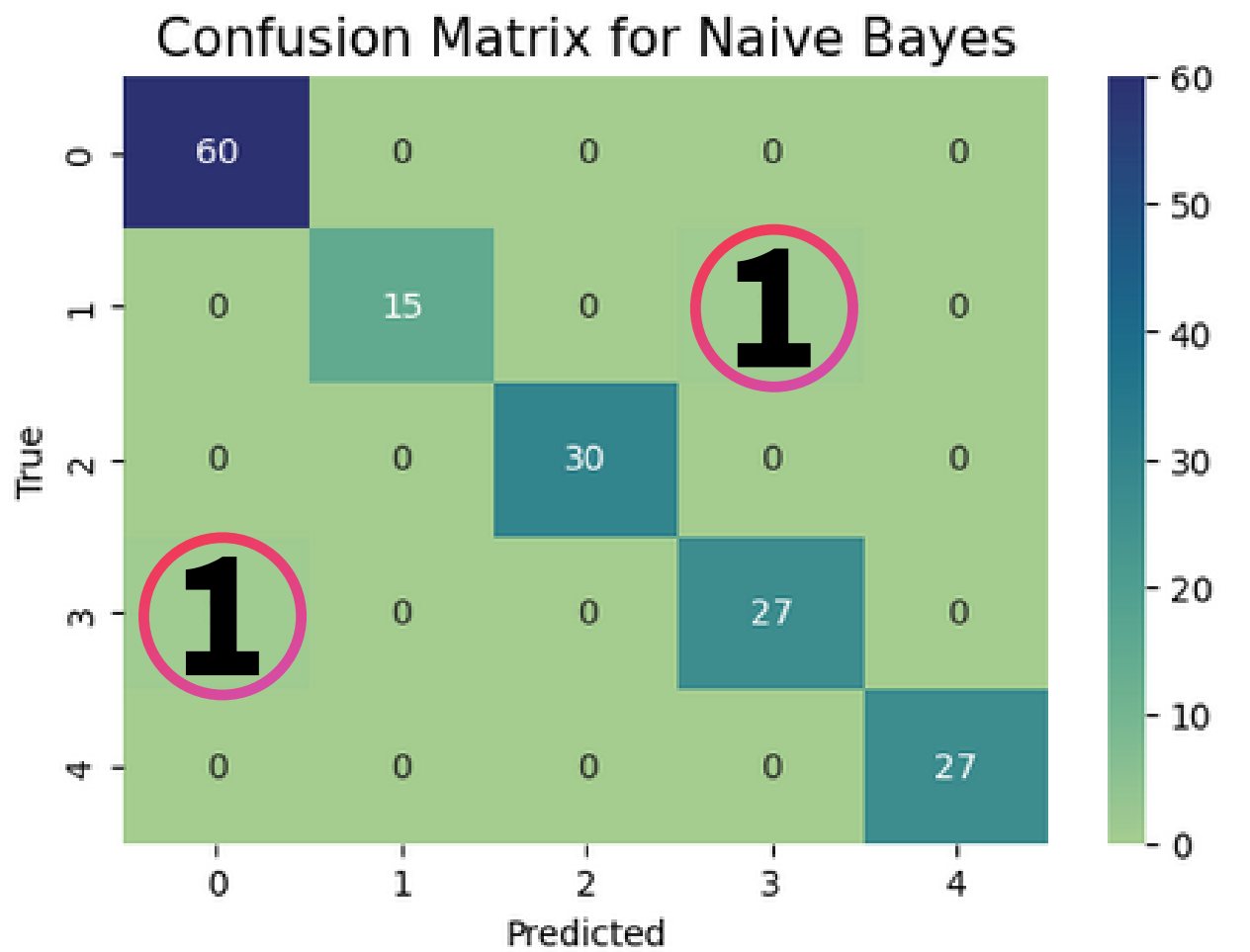
```
Test set classification report:
              precision    recall  f1-score   support

    0           0.98         1.00         0.99         60
    1           1.00         1.00         1.00         16
    2           1.00         1.00         1.00         30
    3           1.00         0.96         0.98         28
    4           1.00         1.00         1.00         27

 accuracy          0.99         0.99         0.99        161
 macro avg          1.00         0.99         0.99        161
 weighted avg       0.99         0.99         0.99        161
```

# Extended Modeling: *Linear*

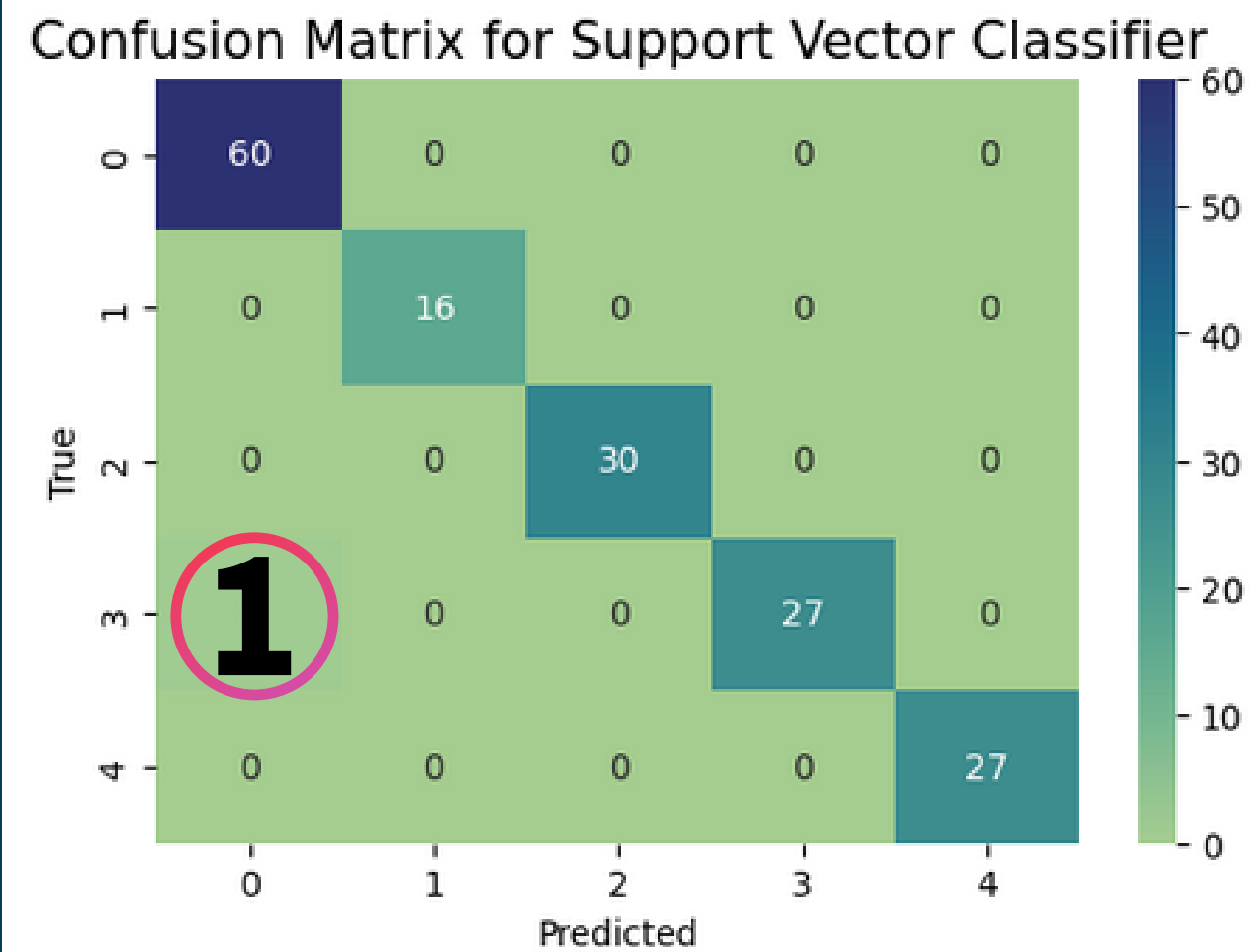
## Naive Bayes



Naive Bayes classification report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	60
1	1.00	0.94	0.97	16
2	1.00	1.00	1.00	30
3	0.96	0.96	0.96	28
4	1.00	1.00	1.00	27
accuracy			0.99	161
macro avg	0.99	0.98	0.98	161
weighted avg	0.99	0.99	0.99	161

## Support Vector Classifier

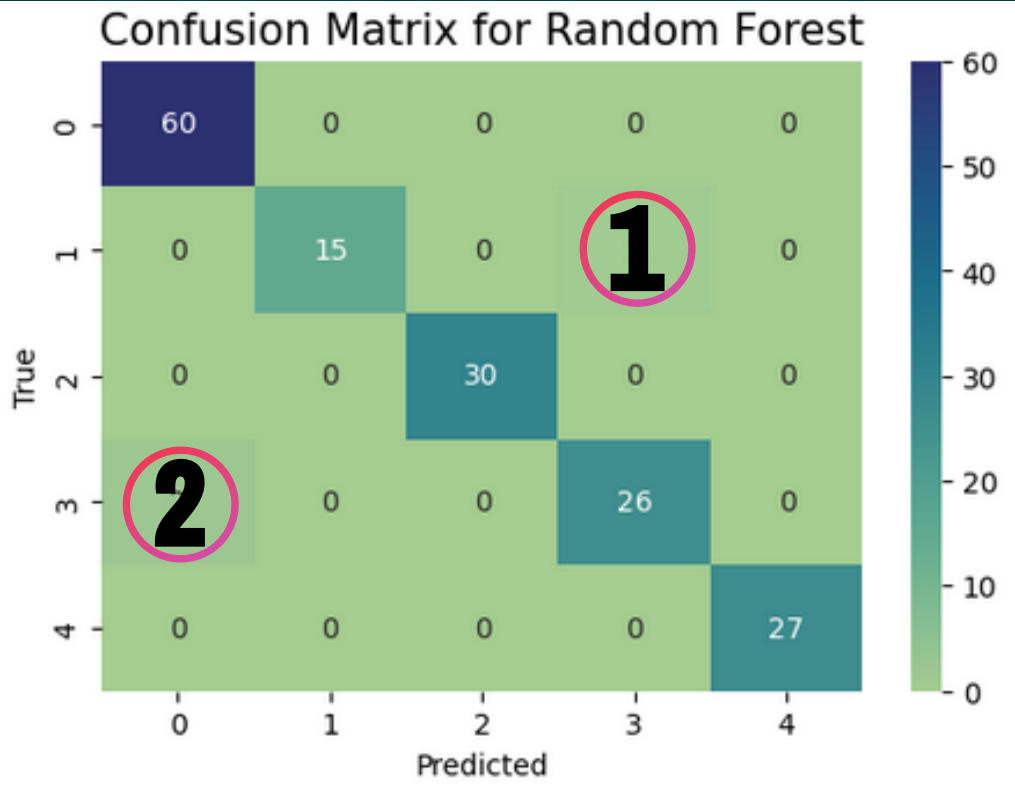


Support Vector Classifier classification report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	60
1	1.00	1.00	1.00	16
2	1.00	1.00	1.00	30
3	1.00	0.96	0.98	28
4	1.00	1.00	1.00	27
accuracy			0.99	161
macro avg	1.00	0.99	0.99	161
weighted avg	0.99	0.99	0.99	161

# Extended Modeling: *Non-linear* with hyper-parameter tuning

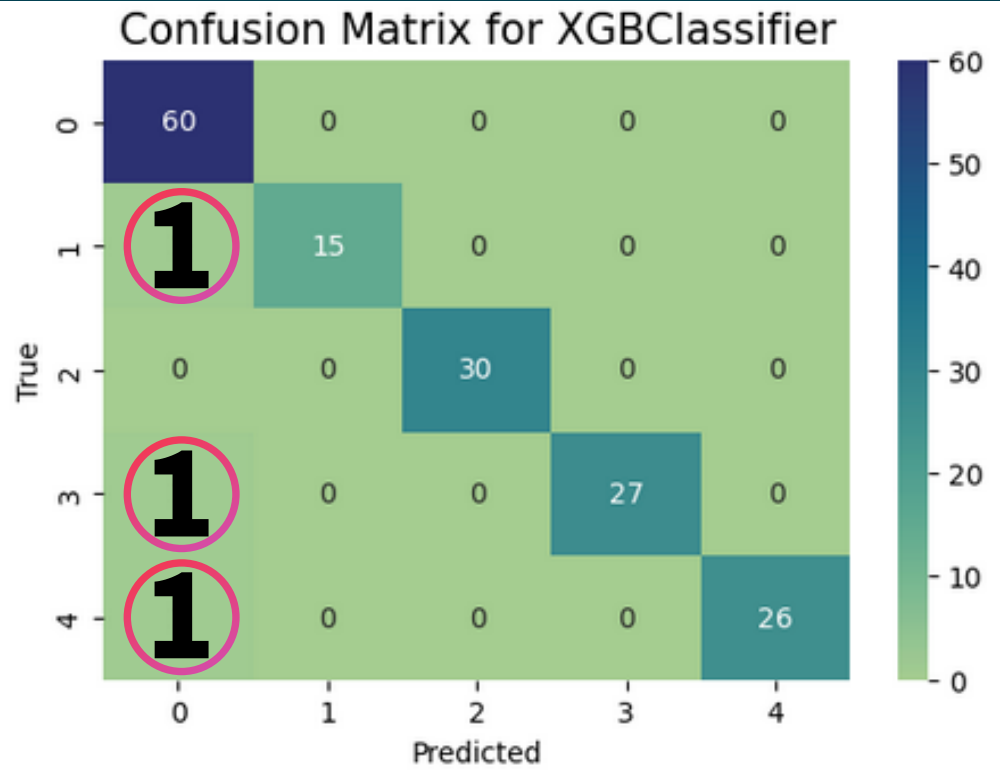
## Random Forest Classifier



Random Forest classification report:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	60
1	1.00	0.94	0.97	16
2	1.00	1.00	1.00	30
3	0.96	0.93	0.95	28
4	1.00	1.00	1.00	27
accuracy			0.98	161
macro avg	0.99	0.97	0.98	161
weighted avg	0.98	0.98	0.98	161

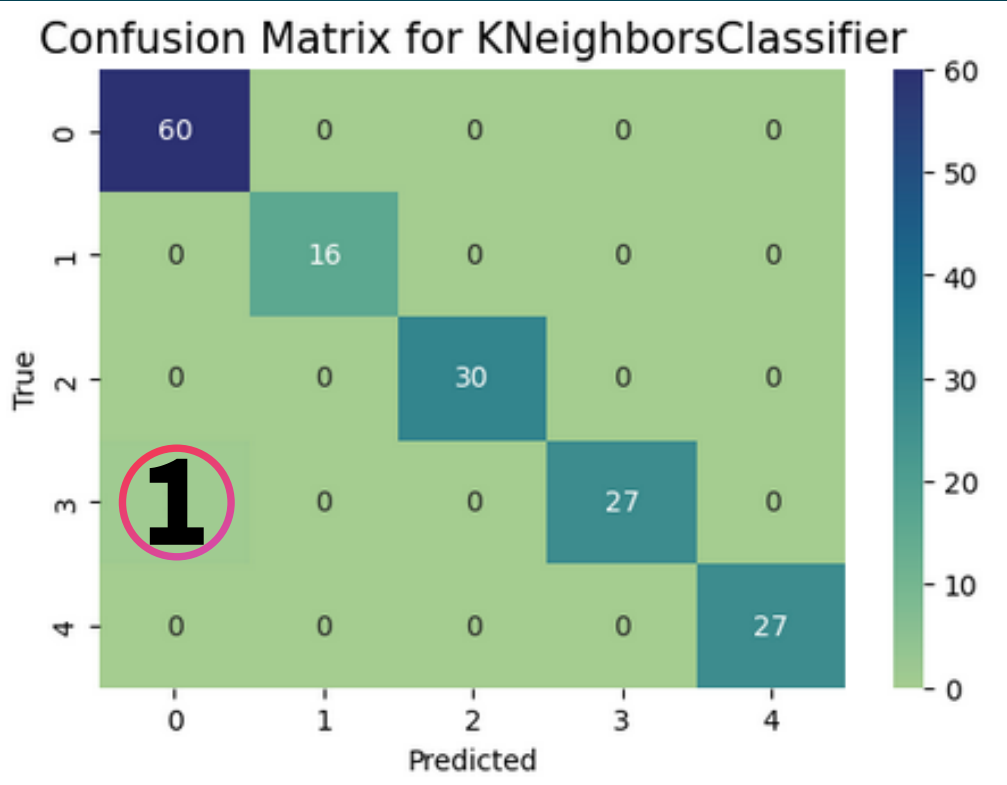
## XGBoost Classifier



XGBClassifier classification report:

	precision	recall	f1-score	support
0	0.95	1.00	0.98	60
1	1.00	0.94	0.97	16
2	1.00	1.00	1.00	30
3	1.00	0.96	0.98	28
4	1.00	0.96	0.98	27
accuracy			0.98	161
macro avg	0.99	0.97	0.98	161
weighted avg	0.98	0.98	0.98	161

## N Neighbors Classifier



KNeighborsClassifier classification report:

	precision	recall	f1-score	support
0	0.98	1.00	0.99	60
1	1.00	1.00	1.00	16
2	1.00	1.00	1.00	30
3	1.00	0.96	0.98	28
4	1.00	1.00	1.00	27
accuracy			0.99	161
macro avg	1.00	0.99	0.99	161
weighted avg	0.99	0.99	0.99	161

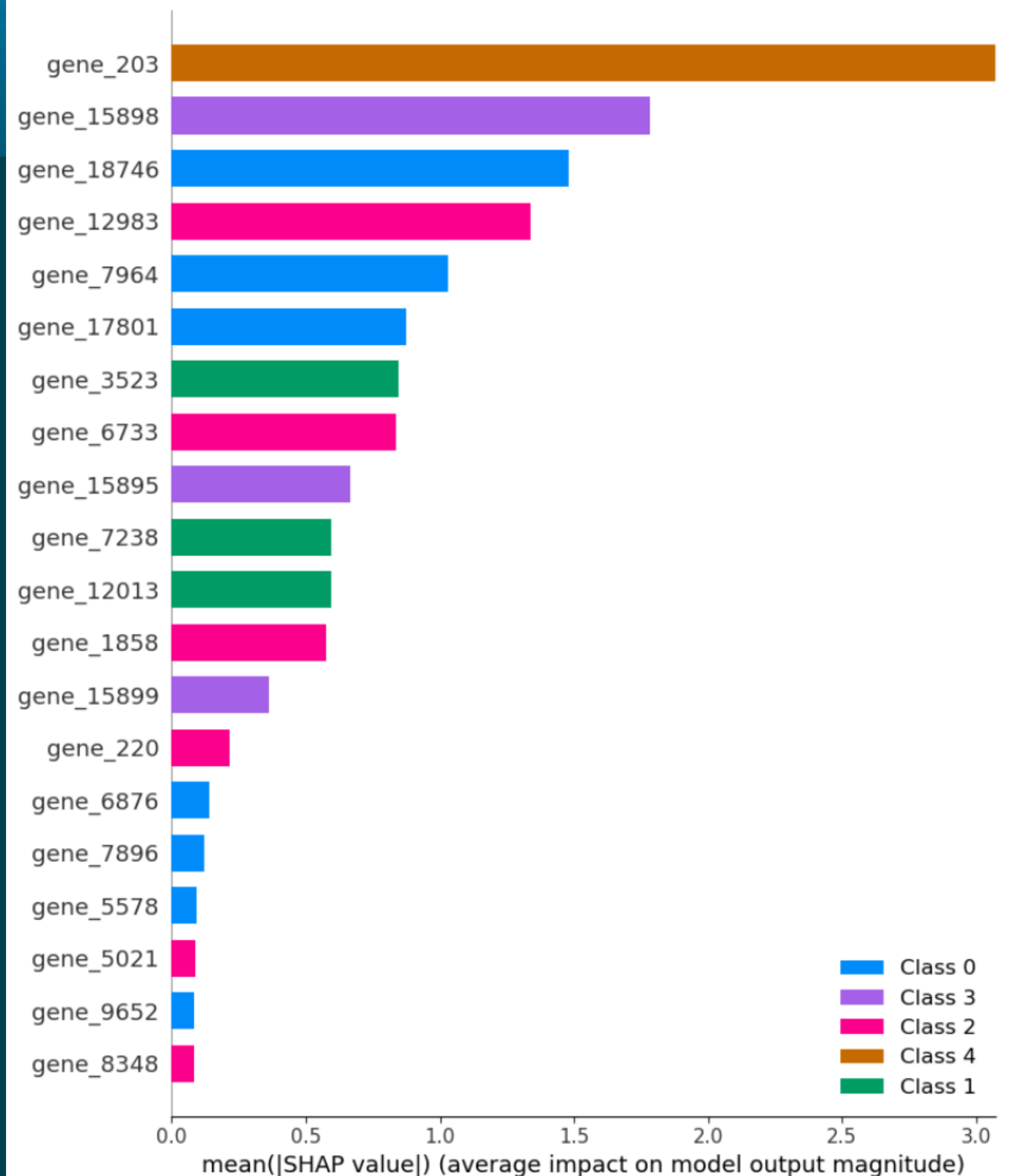
# Model Comparison: *f1 macro scores*

	model names	f1 scores
3	KNeighborsClassifier	0.994711
0	Logistic Regression	0.994711
1	Support Vector Classifier	0.994711
2	Naive Bayes	0.984753
5	XGBoostClassifier	0.981260
4	Random Forest	0.979361

# Interpretion:

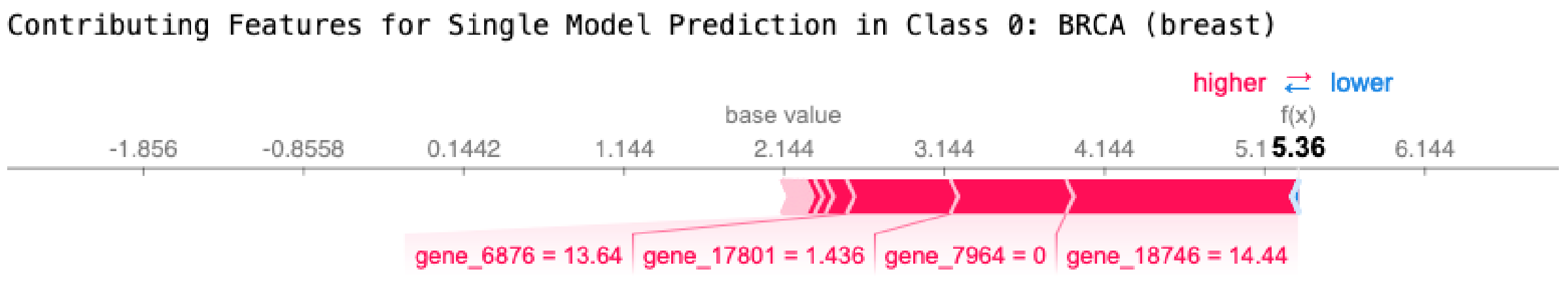
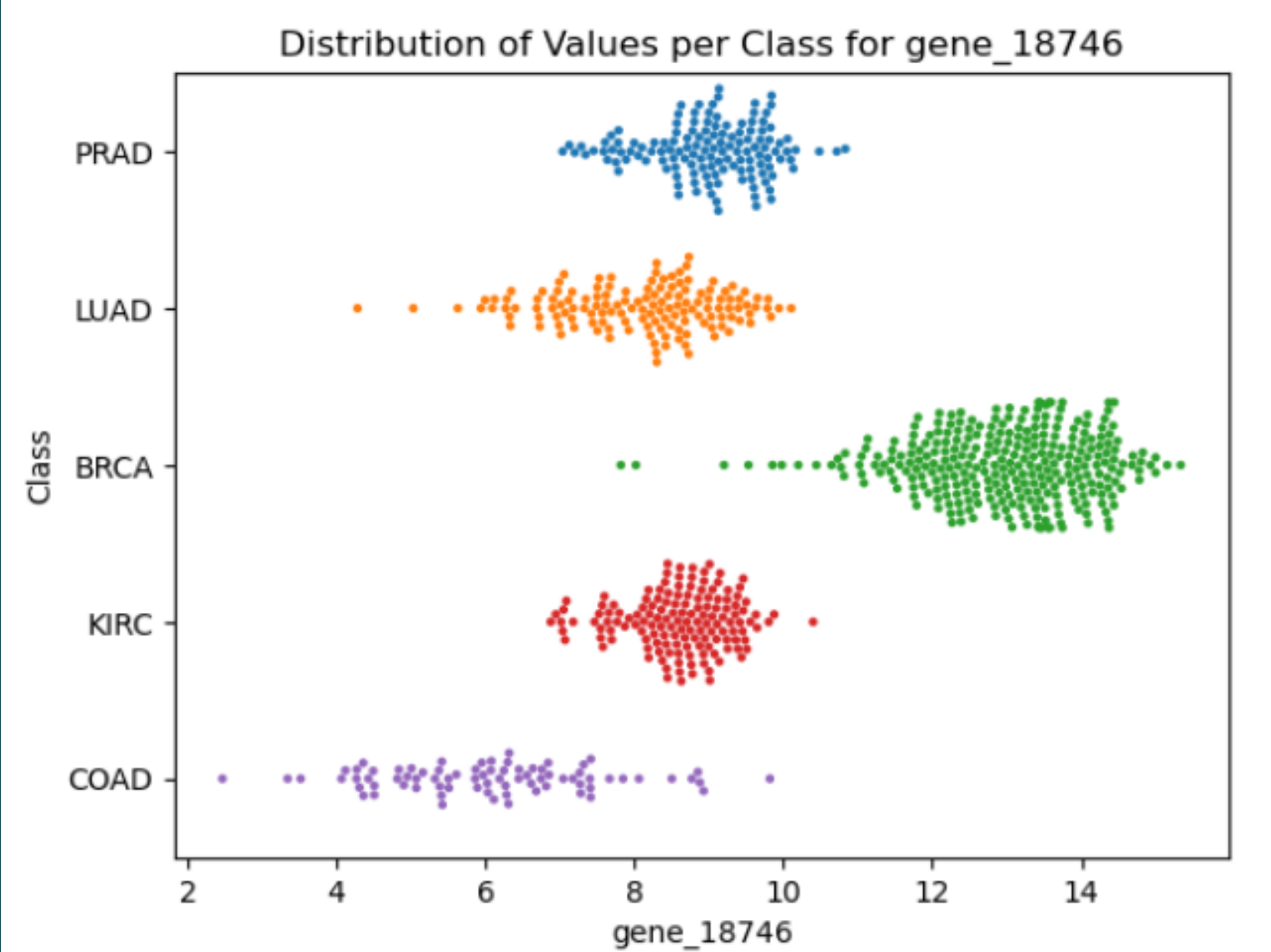
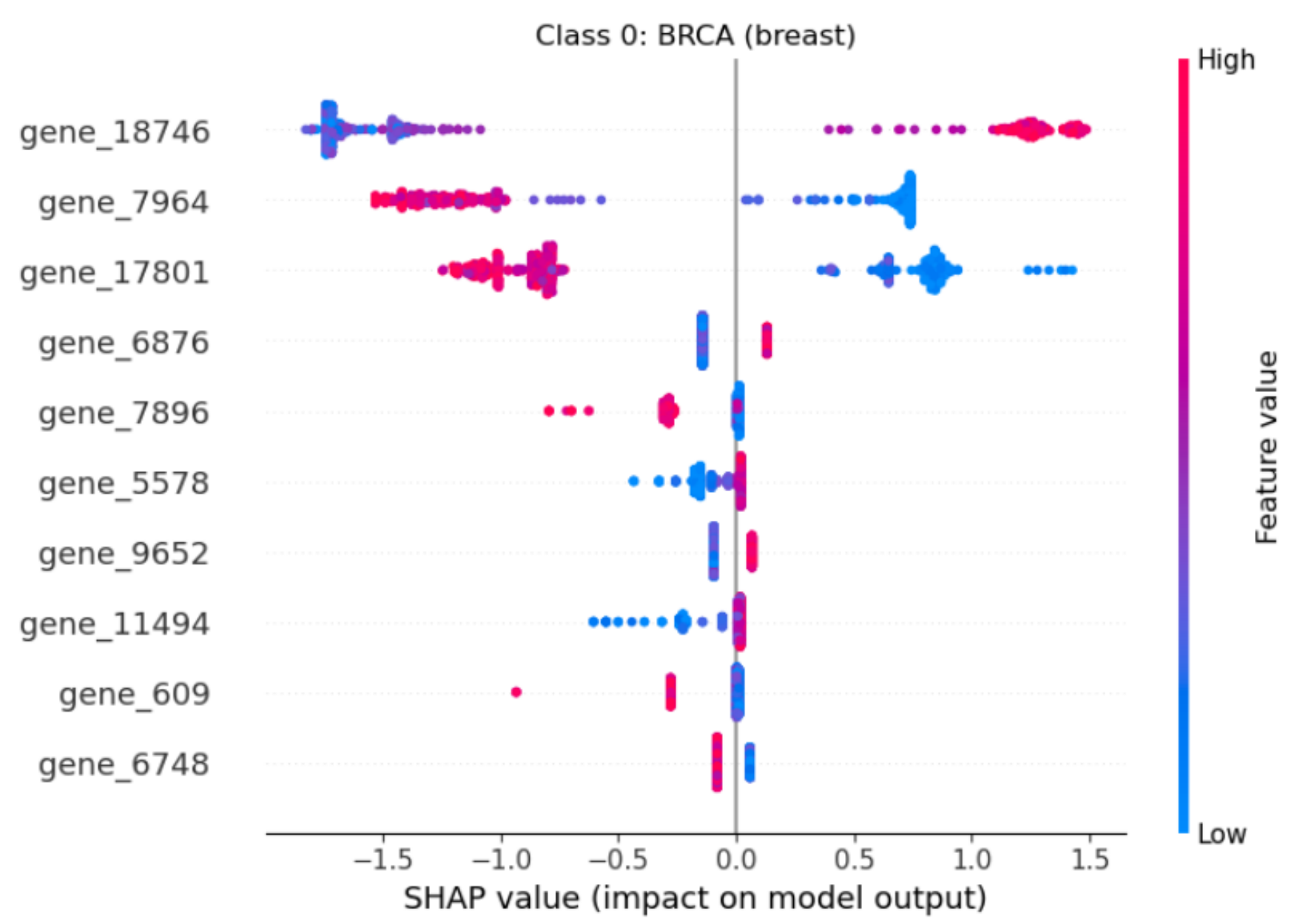
*XGBoost Classifier and SHAP*

## Feature Importance per class

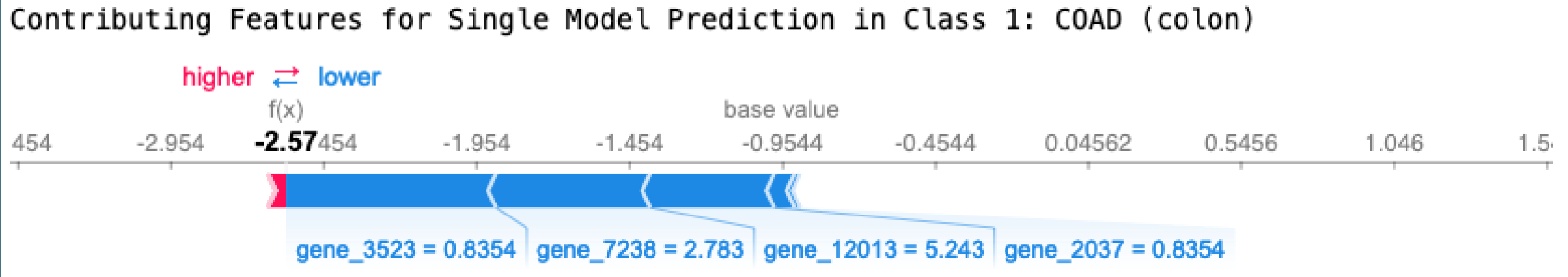
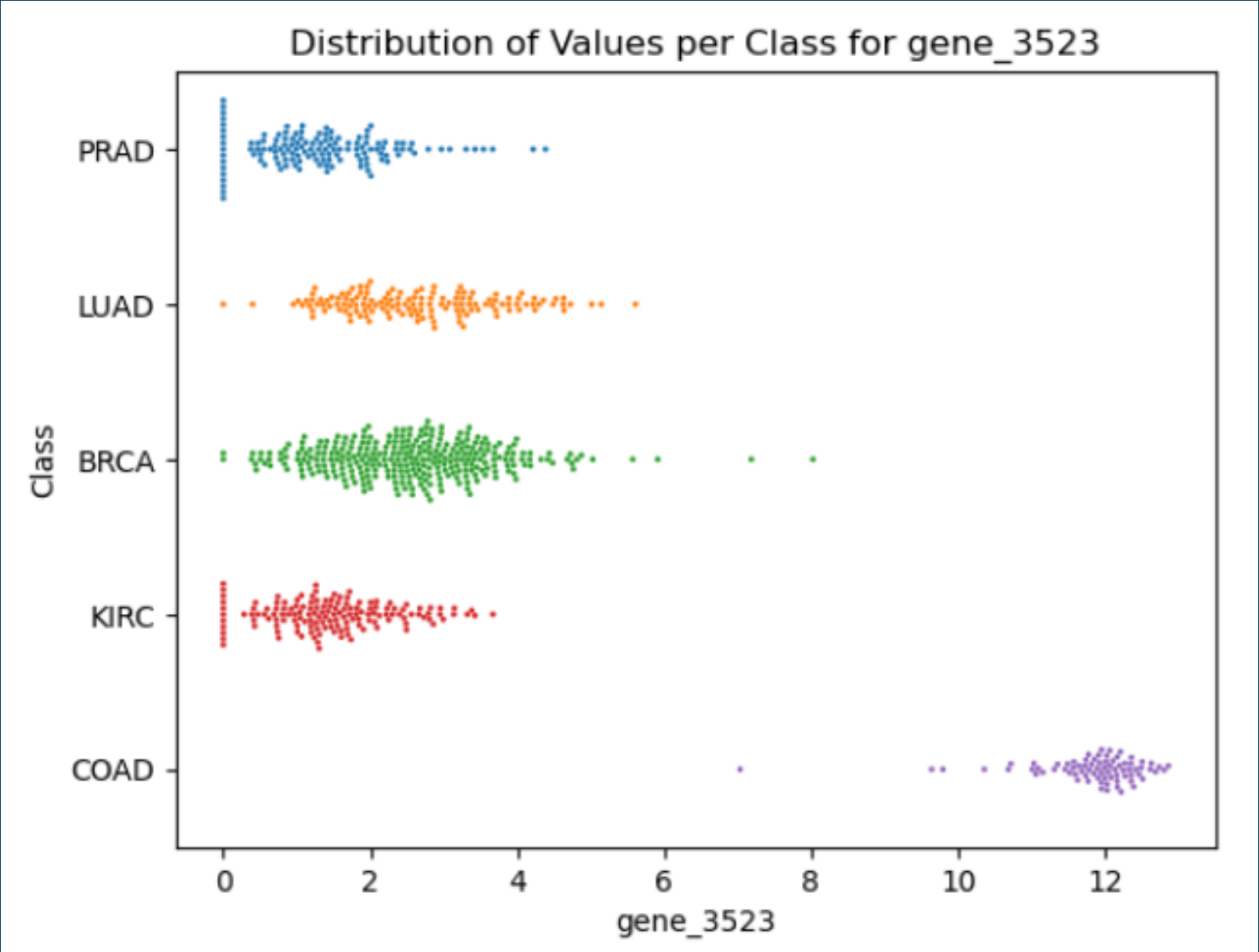
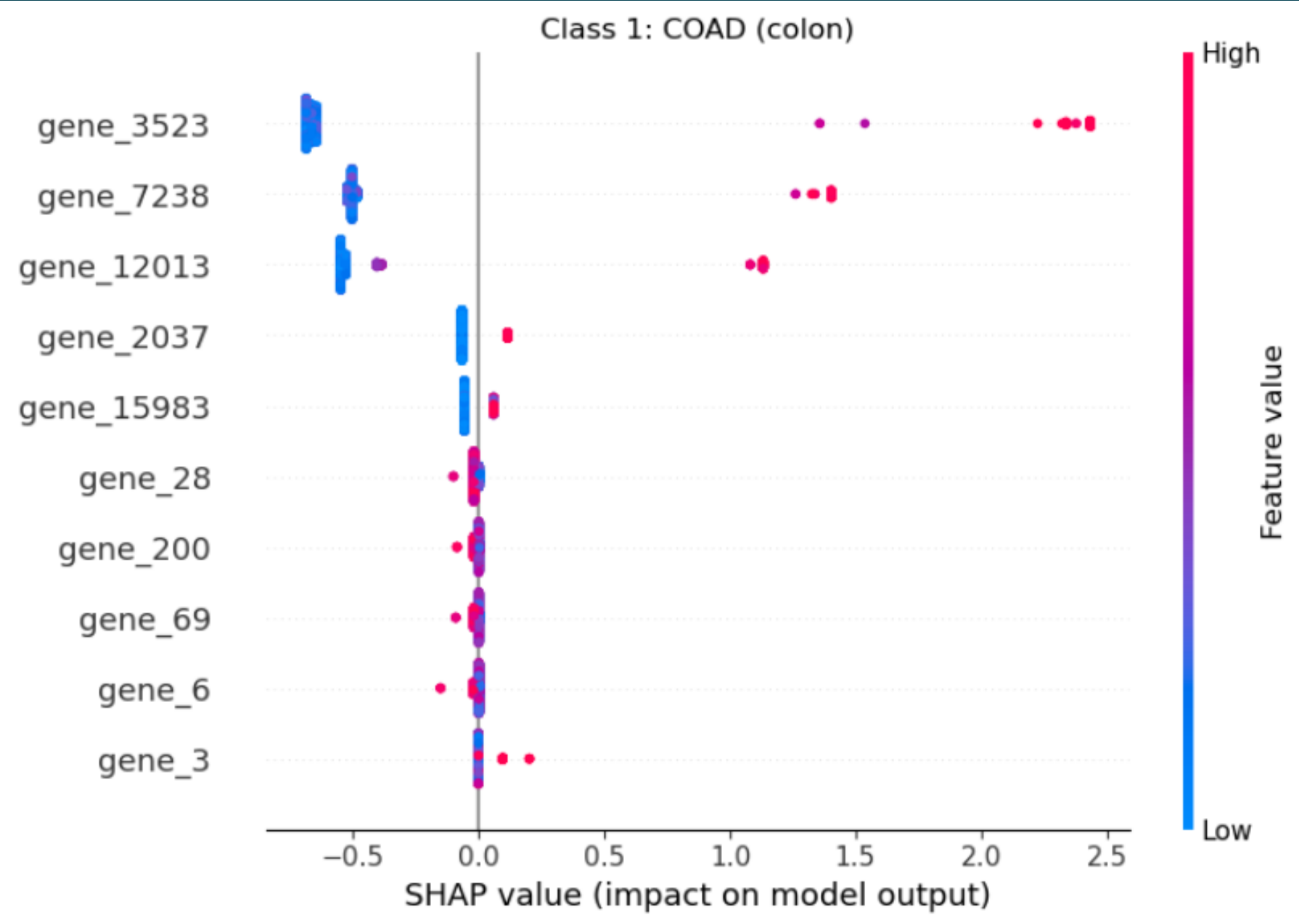




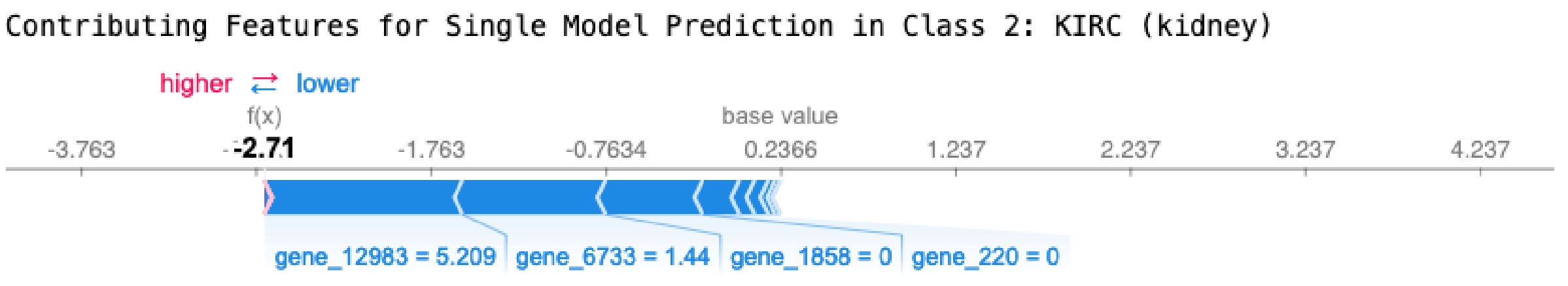
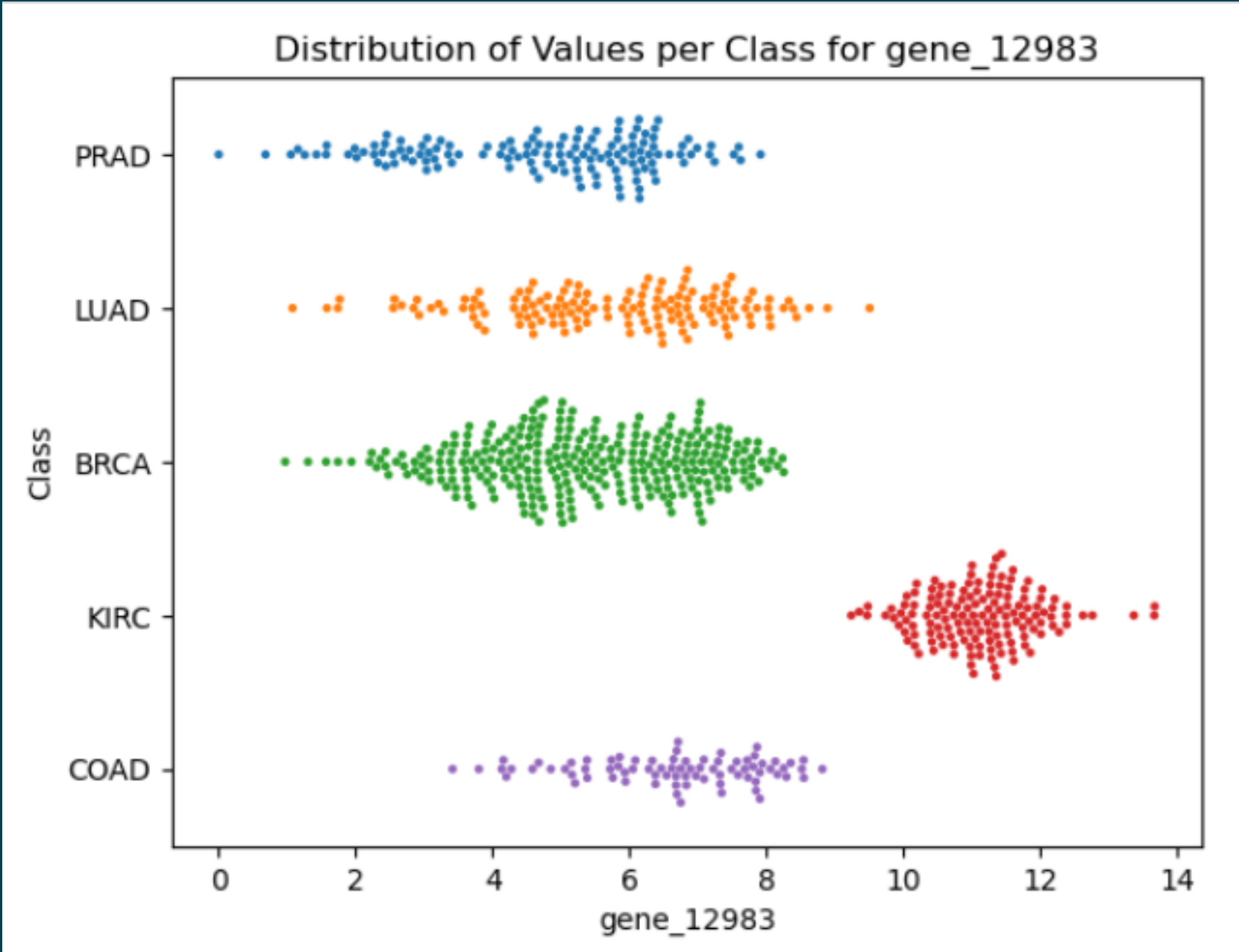
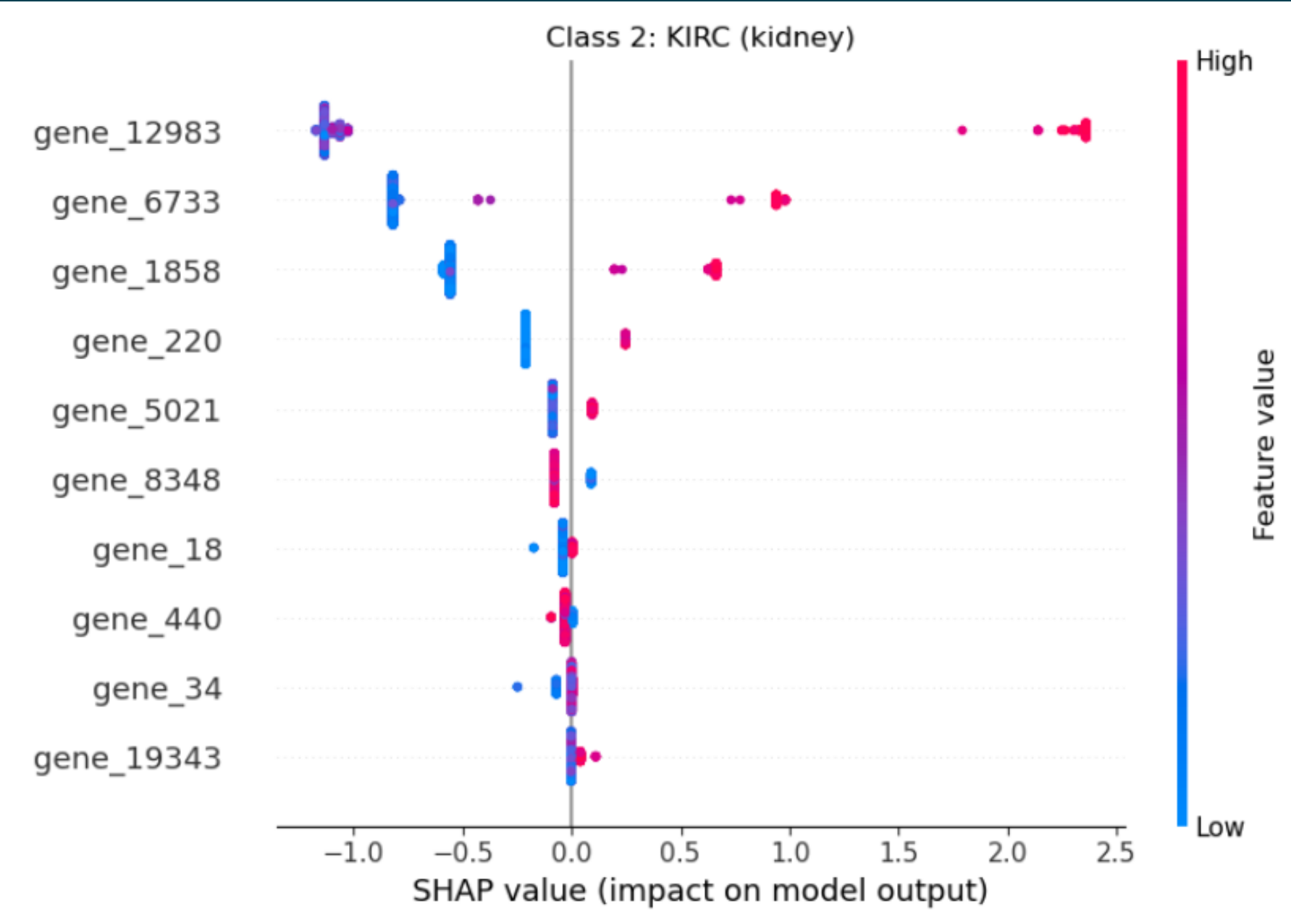
# BRCA (breast): *Class 0*



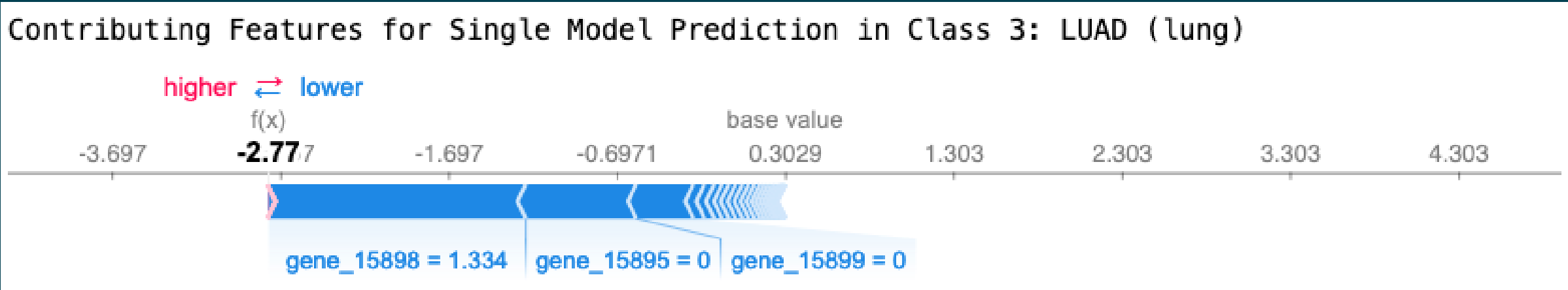
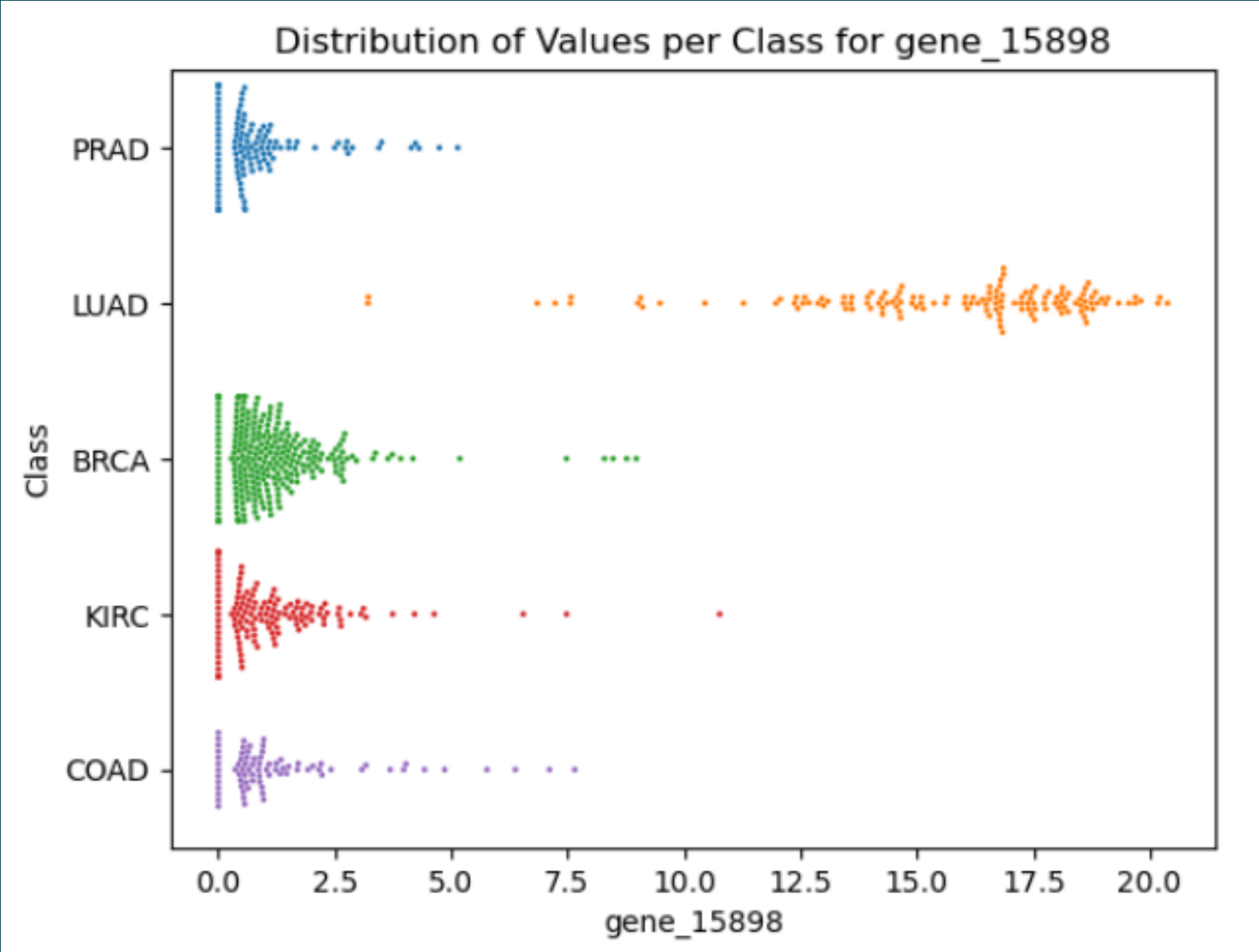
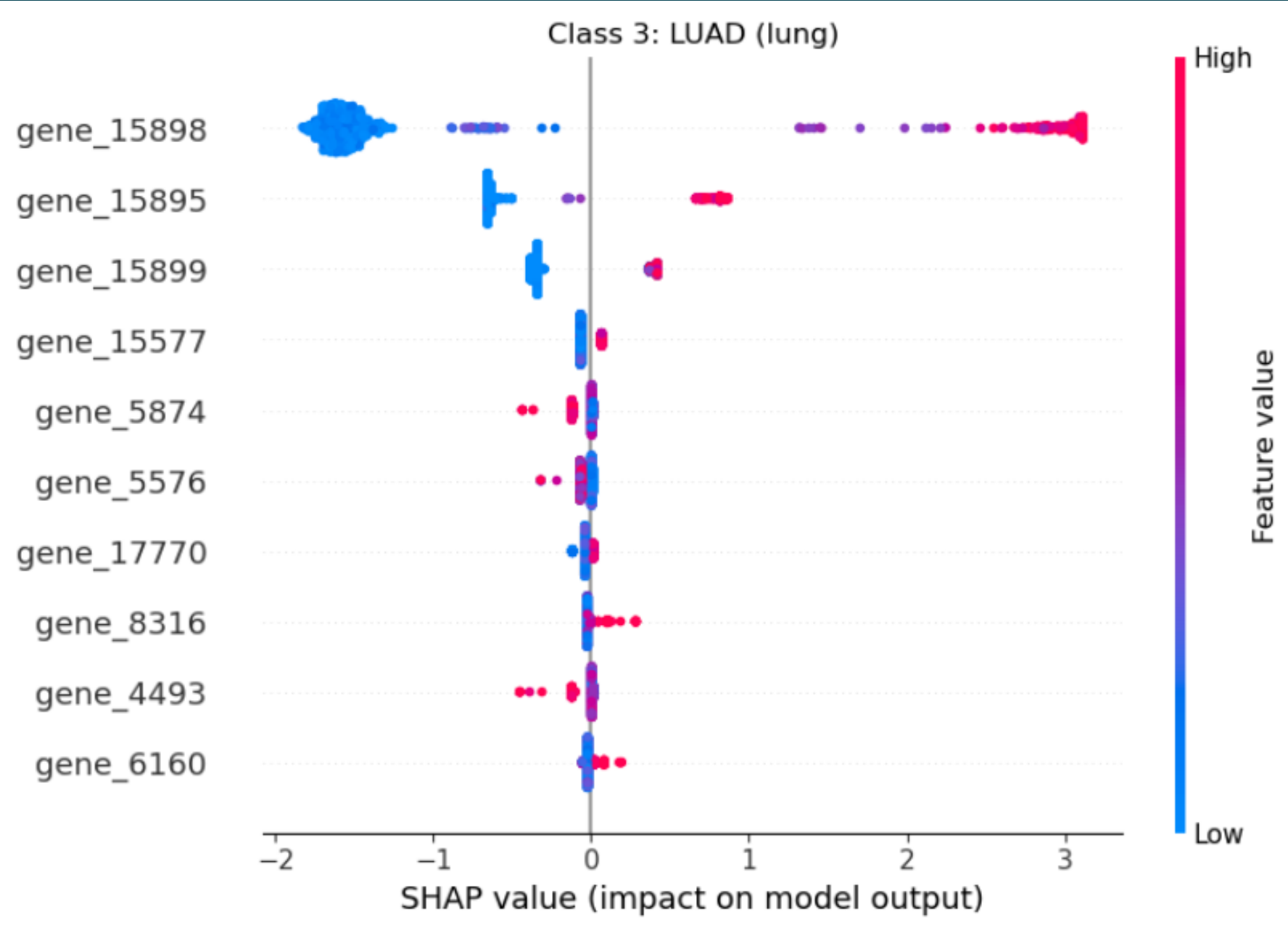
# COAD (colon): *Class 1*



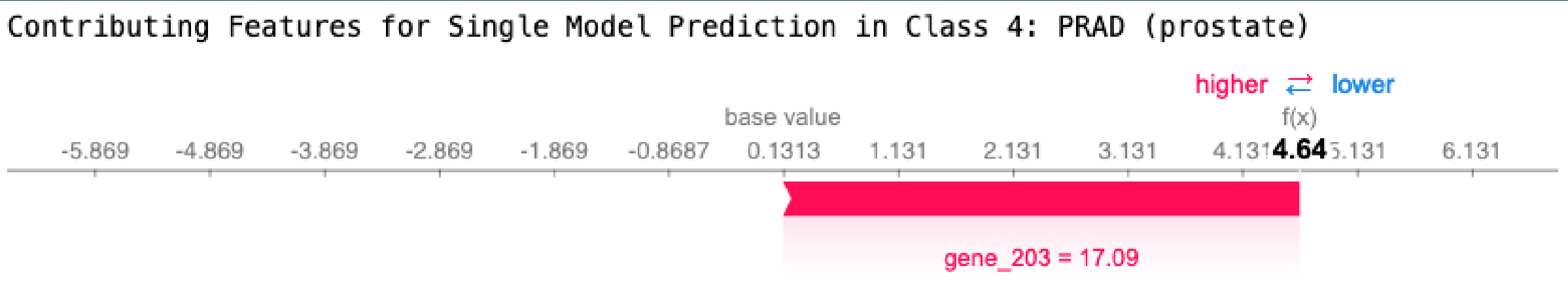
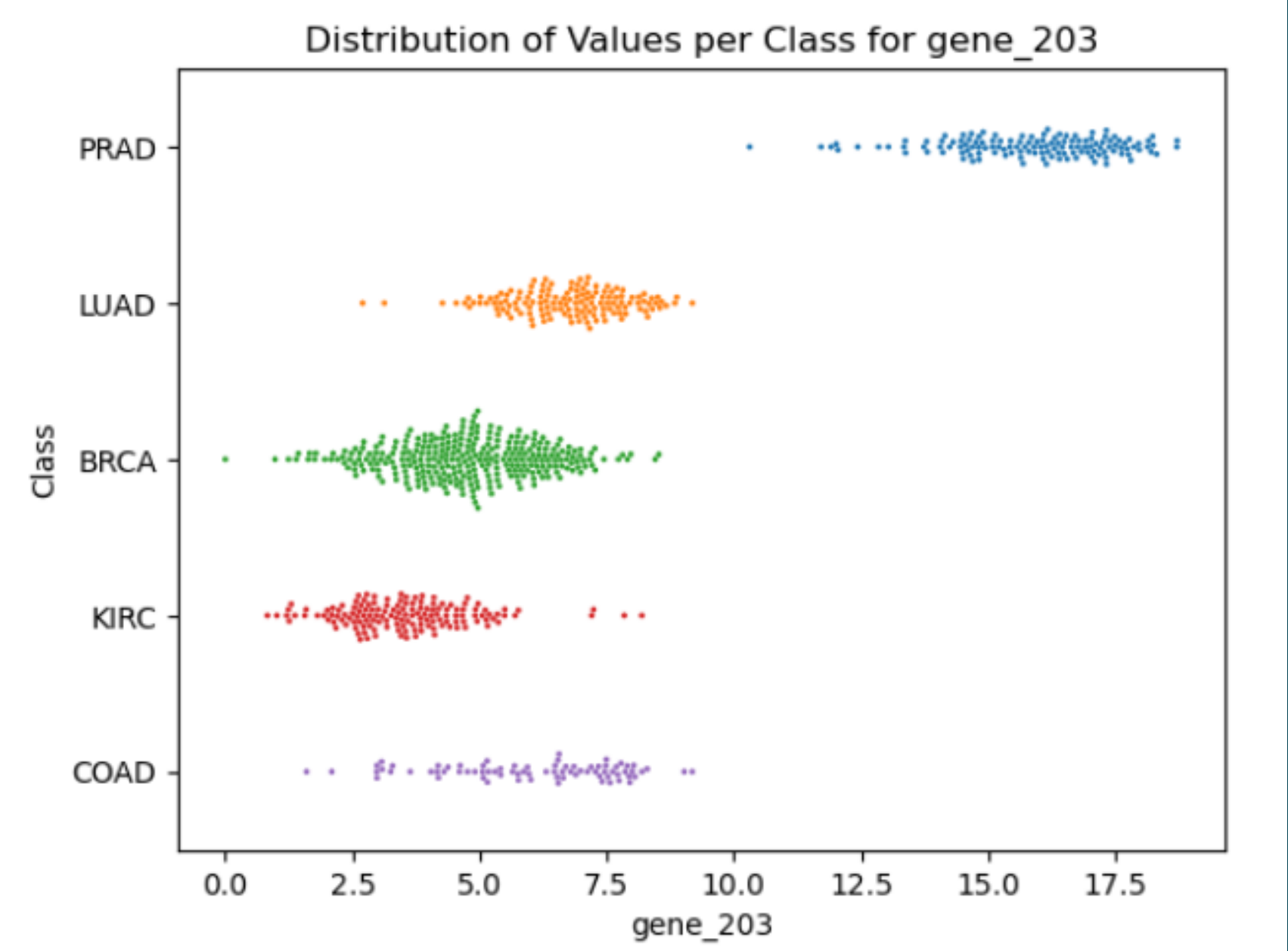
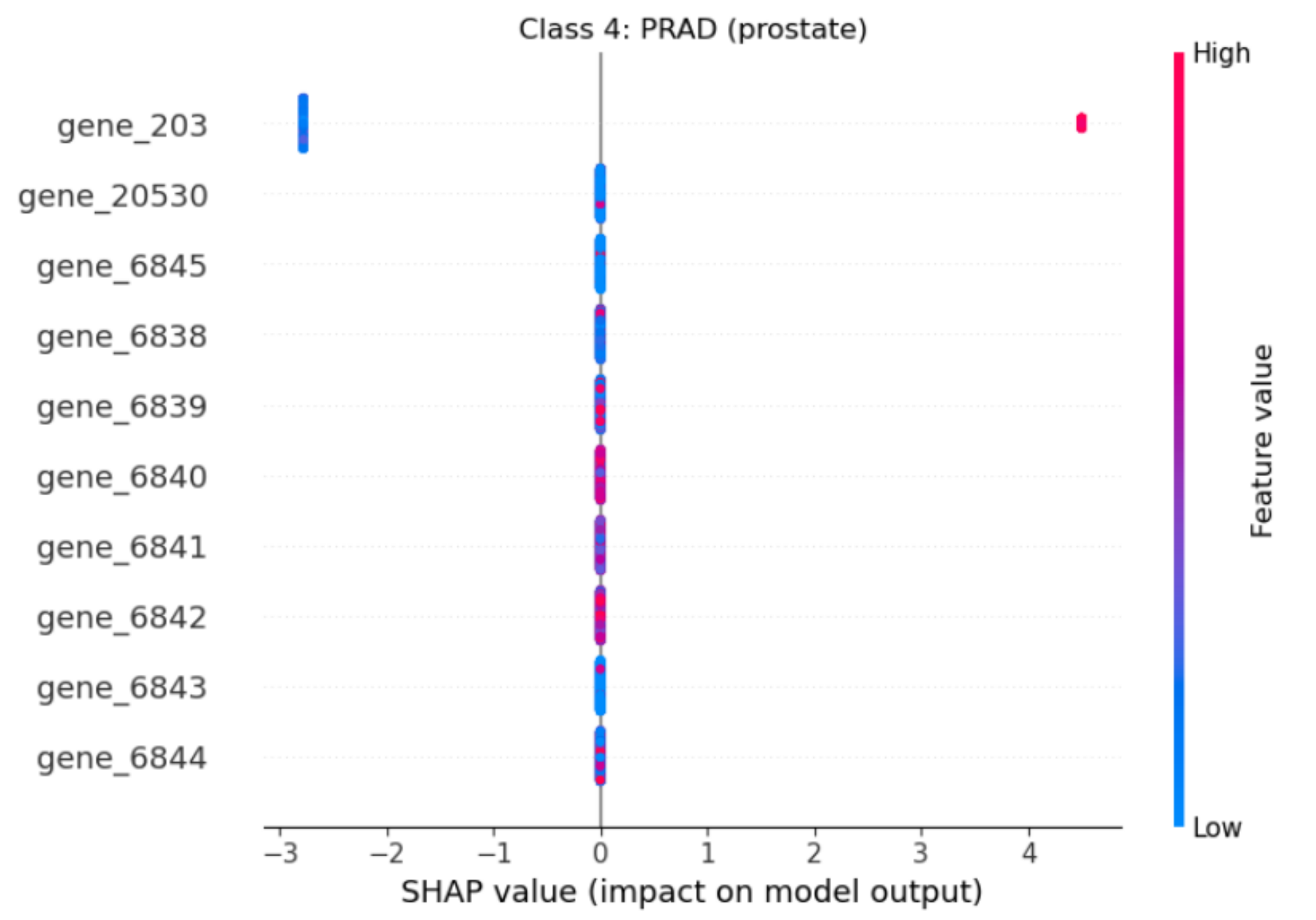
# KIRC (kidney): *Class 2*



# LUAD (lung): *Class 3*



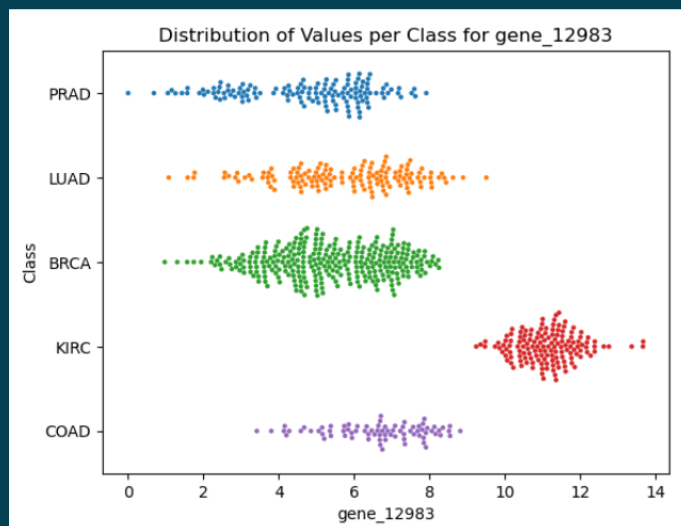
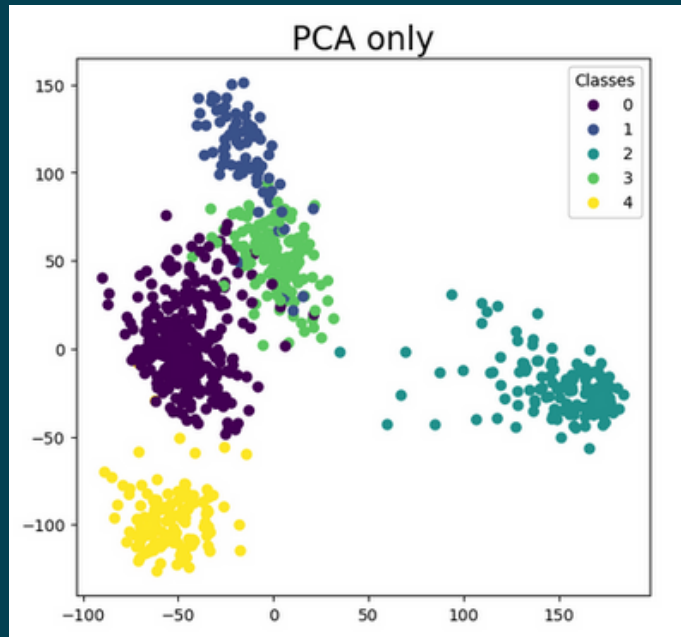
# PRAD (prostate): *Class 4*





# Data Predictability *Skepticism of Model Success*

All models had f1 macro scores of over 97%. Why are the classes so easy to predict?



## What I Know

- Classes were largely separate with just two principal components
- Higher values for top feature seen in membership class
- Data comes from a reputable source

## What I Don't Know

- In the real world, is there a direct relationship between high gene expression values and cancer class membership? We see a correlation in the data but that cannot be misconstrued for causation.
- Are all gene expression datasets this predictable or is there something abnormal, wrong, or exceptional about this particular dataset?



# Conclusions and Client Recommendations

Best Models per Class based on Rate of False Positives/Negatives

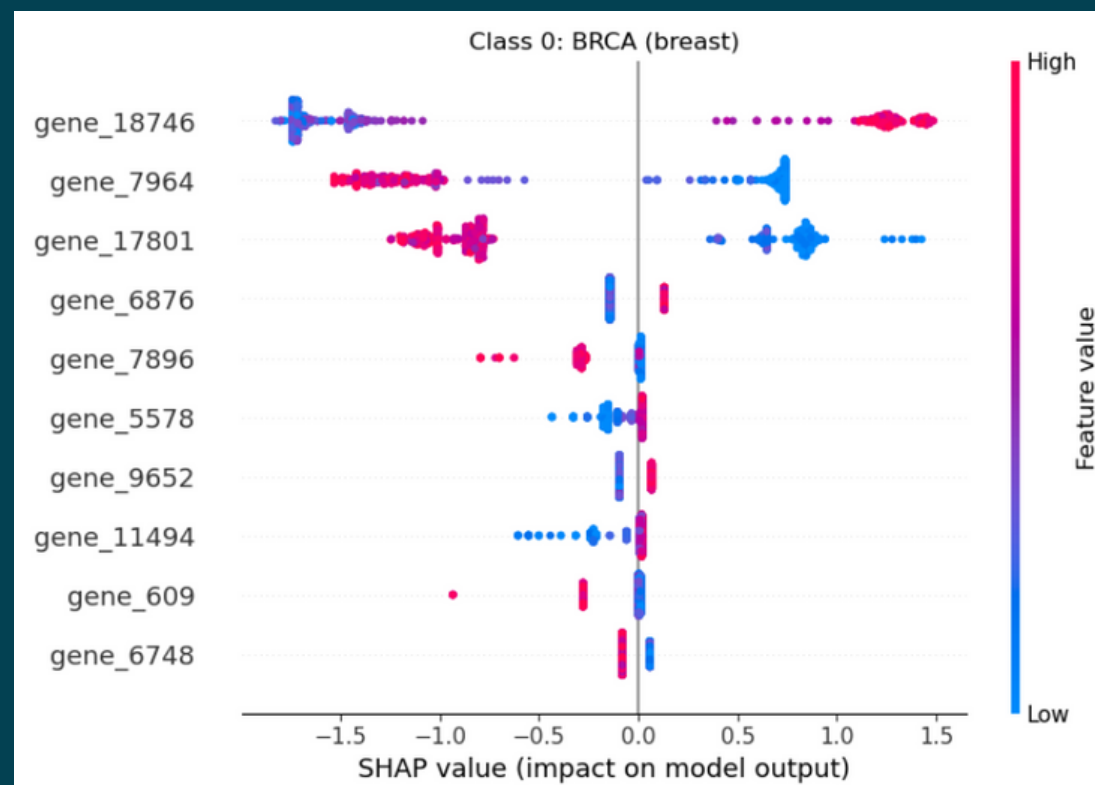
	Score	Class 0 BRCA (breast)		Class 1 COAD (colon)		Class 2 KIRC (kidney)		Class 3 LUAD (lung)		Class 4 PRAD (prostate)	
		FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
K Neighbors Classifier	0.994711	1	0	0	0	0	0	0	1	0	0
Logistic Regression	0.994711	1	0	0	0	0	0	0	1	0	0
Support Vector Classifier	0.994711	1	0	0	0	0	0	0	1	0	0
Naive Bayes	0.984753	1	0	0	1	0	0	1	1	0	0
XGBoost Classifier	0.981260	3	0	0	1	0	0	0	1	0	1
Random Forest Classifier	0.979361	2	0	0	1	0	0	1	2	0	0

*FP = False Positive (a false positive for class x is one where a true class x sample is predicted to be class y)*

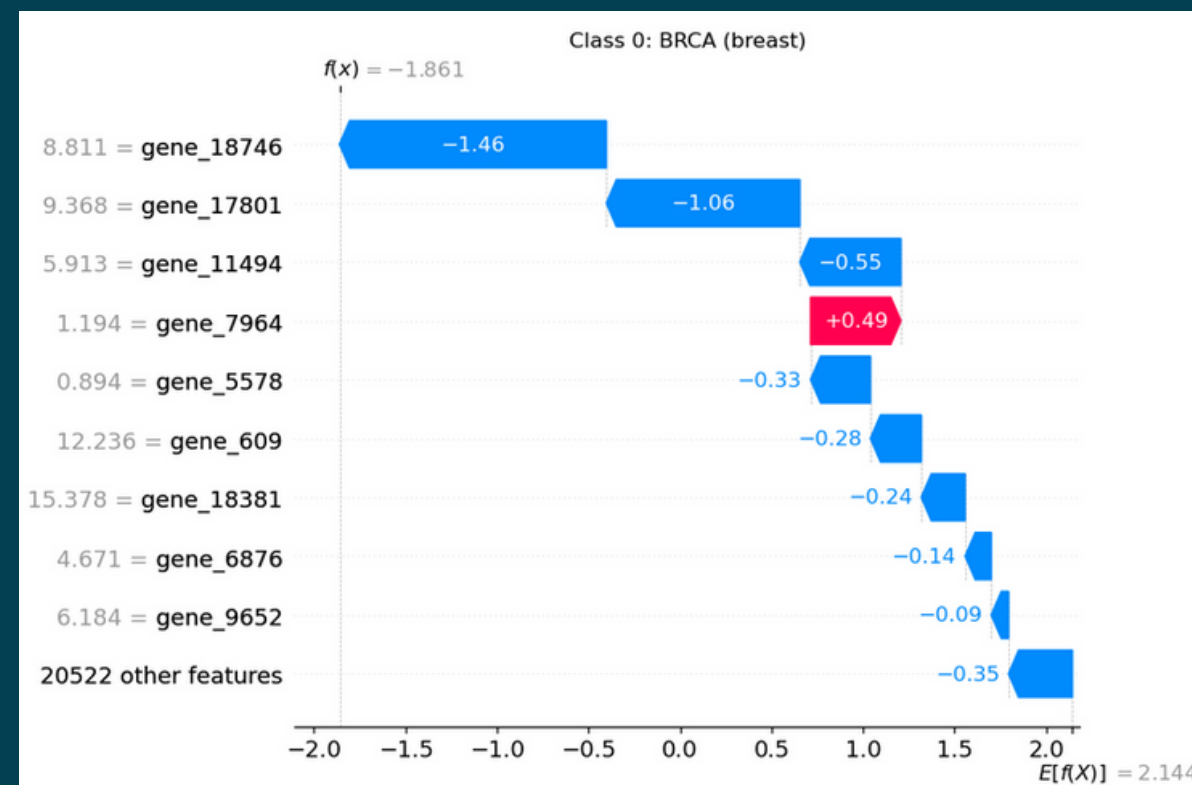
*FN = False Negative (a false negative for class x is one where a true class y sample is predicted to be class x)*

# Conclusions and Client Recommendations

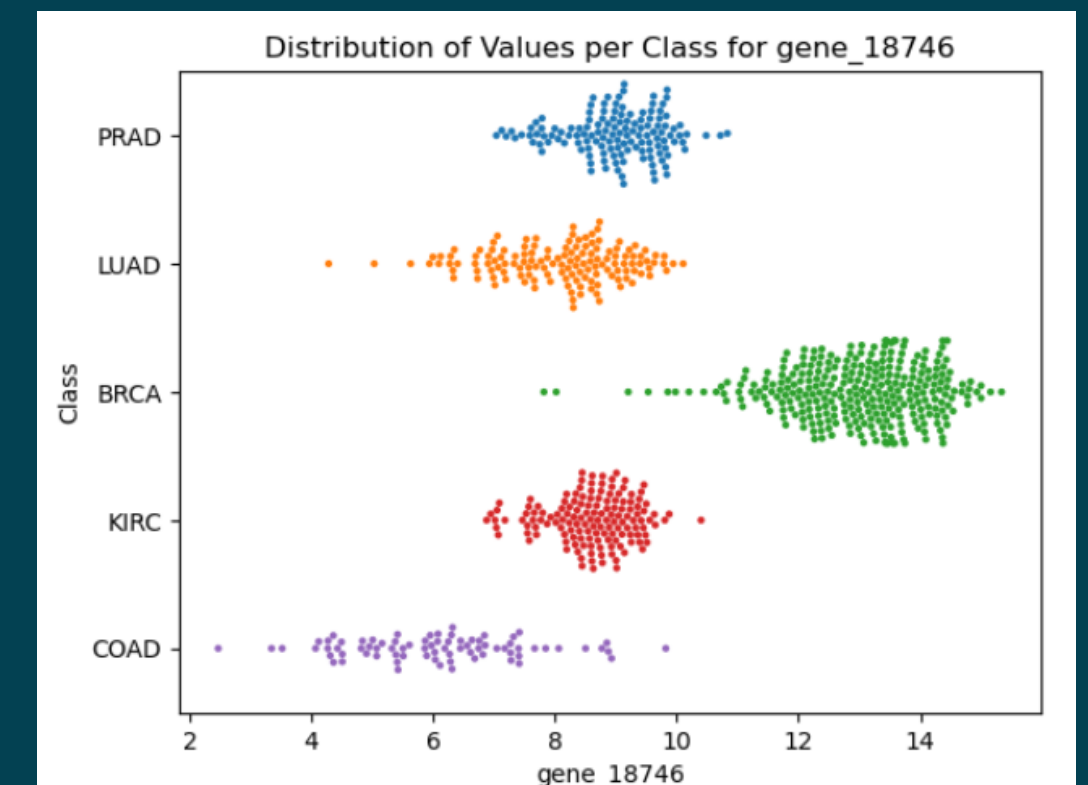
## The Knowledge Gained Through Interpretation



Relationship between SHAP values and Feature Importance



Direction and magnitude of feature importance for single sample

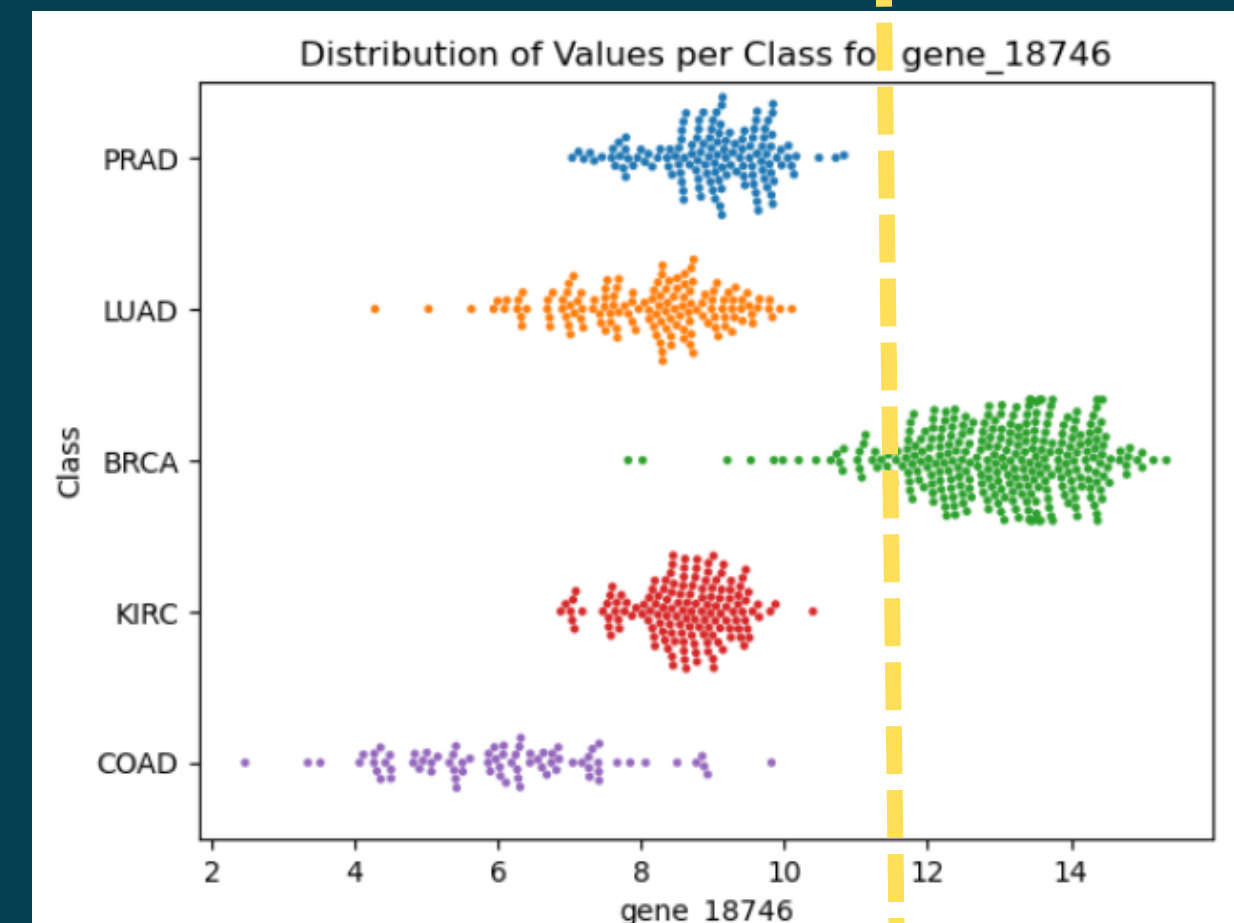
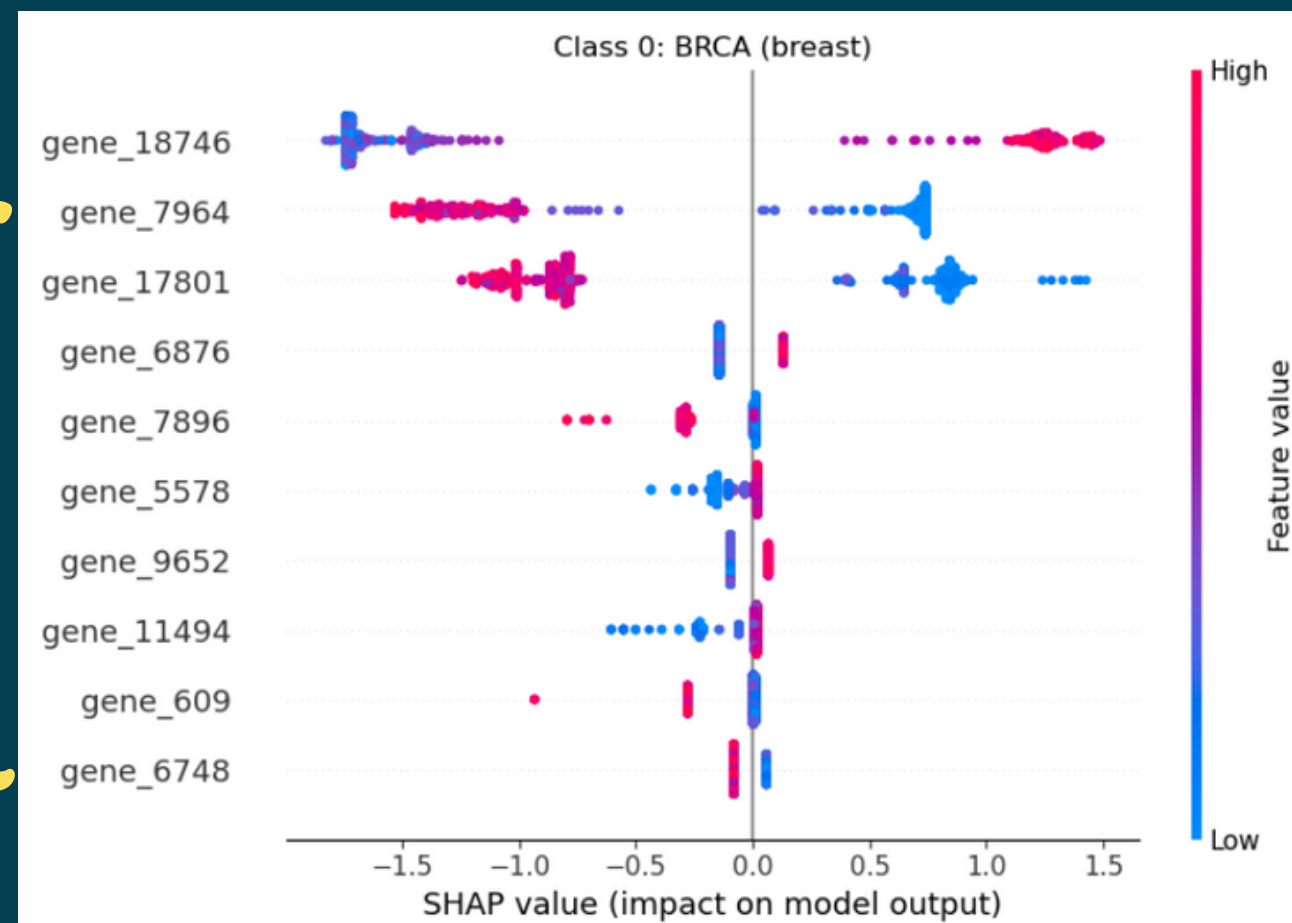


Distribution of values per class for single feature

# Possible Future Project Extension

*Explore each feature in greater depth*

- Plot feature gene expression value distribution for the other important features for each class.
- Find value threshold for important features above which there is a 100% chance of class membership.







# For More Information Please Visit

*<https://github.com/tamarahorne/Springboard/tree/main/Capstone%20Project%203>*