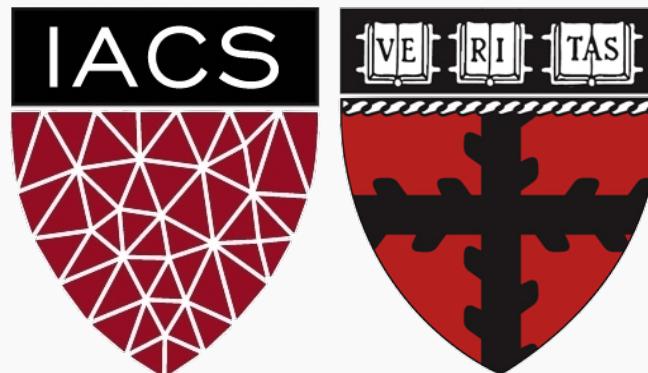


# Classification Metrics

CS109A Introduction to Data Science  
Pavlos Protopapas, Natesh Pillai



# Approach #1: Dry definitions



# Classification Metrics

---

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+|D+)P(D+) + P(T+|D-)P(D-)}$$

- Sensitivity:  $P(T+|D+)$
- Specificity:  $P(T-|D-)$
- Prevalence:  $P(D+)$
- Positive Predictive Value:  $P(D+|T+)$
- Negative Predictive Value:  $P(D-|T-)$

D + - Disease  
D - - Doesn't have disease

		predicted condition		
		prediction positive	prediction negative	Sensitivity
true condition	total population			
	condition positive	True Positive (TP)	False Negative (FN) (Type II error)	<b>Recall =</b> $\frac{\sum \text{TP}}{\sum \text{condition positive}}$
	condition negative	False Positive (FP) (Type I error)	True Negative (TN)	<b>Specificity =</b> $\frac{\sum \text{TN}}{\sum \text{condition negative}}$
	<b>Accuracy =</b> $\frac{\sum \text{TP} + \sum \text{TN}}{\sum \text{total population}}$	<b>Precision=</b> $\frac{\sum \text{TP}}{\sum \text{prediction positive}}$		<b>F1 Score =</b> $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

# THE END



# Approach #2: Case Study



# Covid Case Study

- At the peak of the pandemic, many nations with poor healthcare were running short of hospital beds to admit patients.
- Hospital authorities had to take a call on who to admit and who to send home.
- What if we could build a classifier that suggests whether the patient should be immediately admitted to the hospital or sent home ?

## Pan-India-survey: ‘Only 4% Covid patients who needed ICU bed able to get it through routine process’

LocalCircles, a community social media platform that enables people and small businesses to escalate issues for policy and enforcement interventions, decided to conduct a survey to get the pulse on the issue, and received over 17,000 responses from over 211 districts across India



Written by [Anuradha Mascarenhas](#) | Pune | September 21, 2020 4:34:22 am



### • LIVE BLOG

[IPL 2020 Live Score, MI vs SRH Live Cricket Score Online: Mumbai eye 200](#)  
18 mins ago

[UPSC CSE prelims exam 2020 LIVE UPDATES: Shift 2 concludes; check analysis, candidates' reactions](#)  
19 mins ago

[Bigg Boss Tamil Season 4 grand launch LIVE UPDATES: Kamal Haasan all set for 'new life, new normal and new beginning'](#)  
37 mins ago

[Punjab, Haryana Farmers Protest Live Updates: If farmers are happy, why are they protesting, says Rahul Gandhi](#)  
1 hour ago

# Who should get medical attention first ?



# Covid case study

## ISSUES?

This analysis is for **educational** purpose only

- The data is sourced by online forms and thus is of questionable source.
- A lot of missing values in the original dataset are simply ignored for simpler analysis.
- The entire premise of predicting urgency of admission is false because some people had to wait longer to be admitted because of lack of hospital beds & resources.



# Covid case-study

## Primary predictors

- **age** (if an age range was provided in the source data, only the first number is used)
- **sex**
- **cough, fever, chills, sore throat, headache, fatigue**

## Outcomes

Classification: `urgency_of_admission`

- 0-1 days from onset of symptoms to admission -> **High**
- 2+ days from onset of symptoms to admission or no admission -> **Low**

Karandeep Singh @kdpsinghlab · Mar 16

I generated a COVID-19 machine learning dataset for my #LHS610 course. It's intended for educational use only.

The purpose is to predict urgency of admission (based on age, sex, and timing/type of symptoms). Take a look and feel free use for teaching! 🙌

[github.com/ml4lhs/covid19...](https://github.com/ml4lhs/covid19...)

COVID-19 Machine Learning Dataset

Intended For Educational Use Only

The dataset is located at [covid\\_ml.csv](https://covid_ml.csv).

"It's hard to over value the importance of really caring about the outcome when learning modeling." - JD Long

The COVID-19 pandemic has affected the lives of many people around the world and is a growing threat to our health. The case volume continues to rise in the United States.

The original data comes from the following source: <http://virological.org/t/epidemiological-data-from-the-ncov-2019-outbreak-early-descriptions-from-publicly-available-data/337>

The original dataset is based on public reports of COVID-19 cases reported internationally. There is a source column that provides a link to the website (or news source) where the case was found.

At the original source, there is a Google Sheet that contains live updating data. The Google Sheet receives very high traffic (blocking access to users), so the data was first exported as an Excel file on March 14 at 5:30 pm. This dataset will be out-of-date by the time you read this as the number of cases is growing exponentially.

From the original dataset, the covid19\_ml.csv dataset contains those cases for which:

JD Long

2

12

48



# Scenario #1 - Brazil

## BRAZIL

- The new covid variant is contagious and infecting many Brazilians.
- Brazilian officials however dictate that hospitals do not classify many people at '**high**' risk to avoid bad press and subsequent political global backlash.
- In numbers we need the best classifier with the following restriction.

$$TPR + FPR \leq 0.5$$

Brazil accused of hiding data on coronavirus crisis

Bolsonaro government stops counting total cases and deaths as country becomes global pandemic hotspot



People in Brasília hold flares during a demonstration against president Jair Bolsonaro and racism, and in support of democracy, on Sunday © REUTERS

# Scenario #2 - Germany

## GERMANY

- German officials want the fatality ratio to be as less as possible.
- Thus, it is imperative to find cases in need of urgent attention and give them the best chance of survival.
- In numbers we need the best classifier with the following restriction.

$$TPR \geq 0.85$$



*With Broad, Random Tests for Antibodies, Germany Seeks Path Out of Lockdown*

It was the first large Western democracy to contain the spread of the coronavirus and is now the first to methodically go about reopening its economy. Others are watching.

Taking a blood sample as part of random sampling for antibodies to the coronavirus.

# Scenario #3 - India

## INDIA

- India has only 1 million beds left, and there are already 2 million people suspected of having the disease
- The officials need to work out a strategy to find the people at most need of urgent
- In numbers we need the best classifier with the following restriction



$$TPR + FPR \leq 1$$

# Two models

---

## Logistic Regression



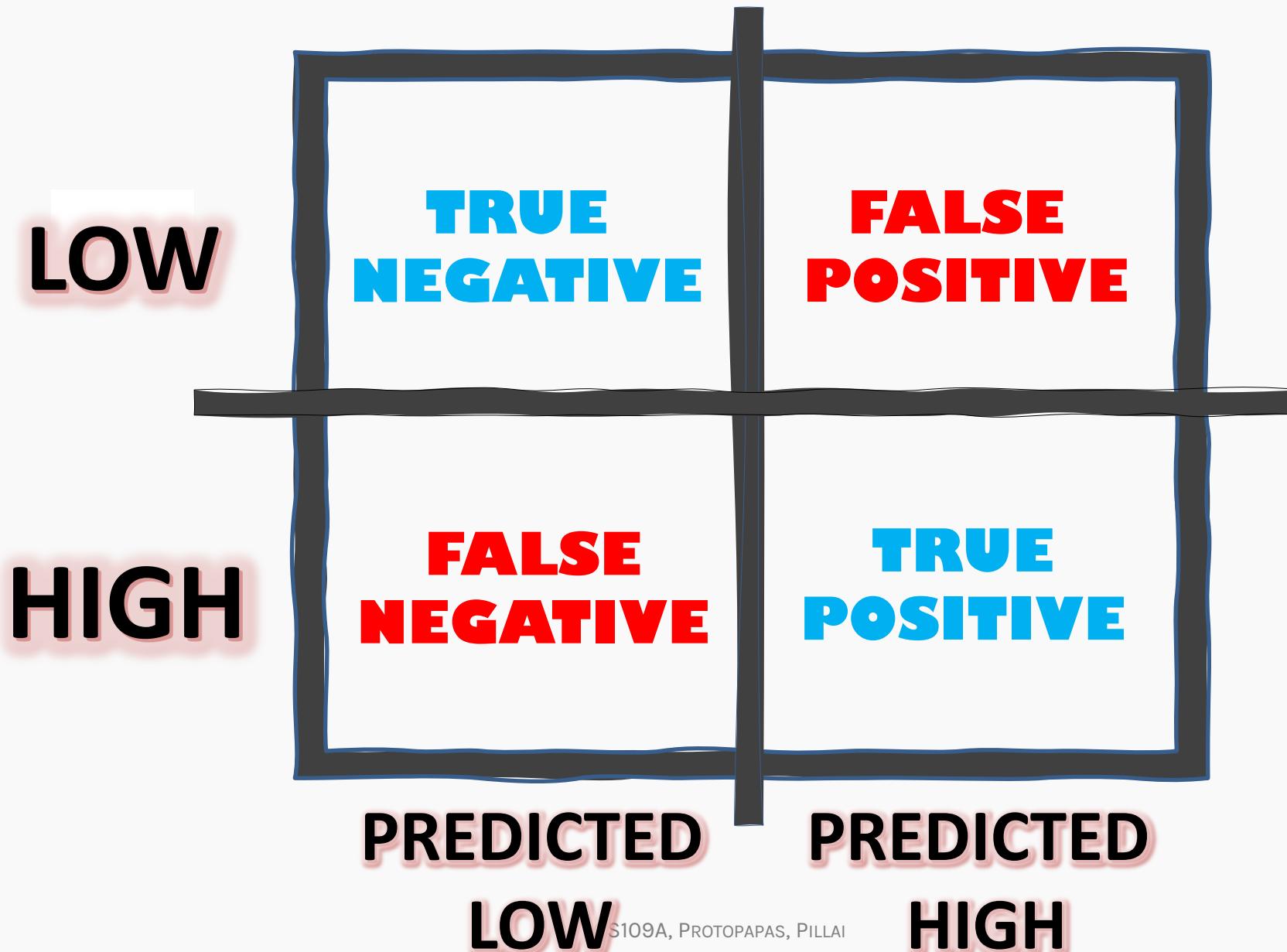
## kNN Classification



# Model Comparison – Logistic vs kNN

Classification Metric	Formula	Logistic Regression	kNN Classification
Accuracy			
Sensitivity (Recall)			
Specificity			
Precision			
F1 score			

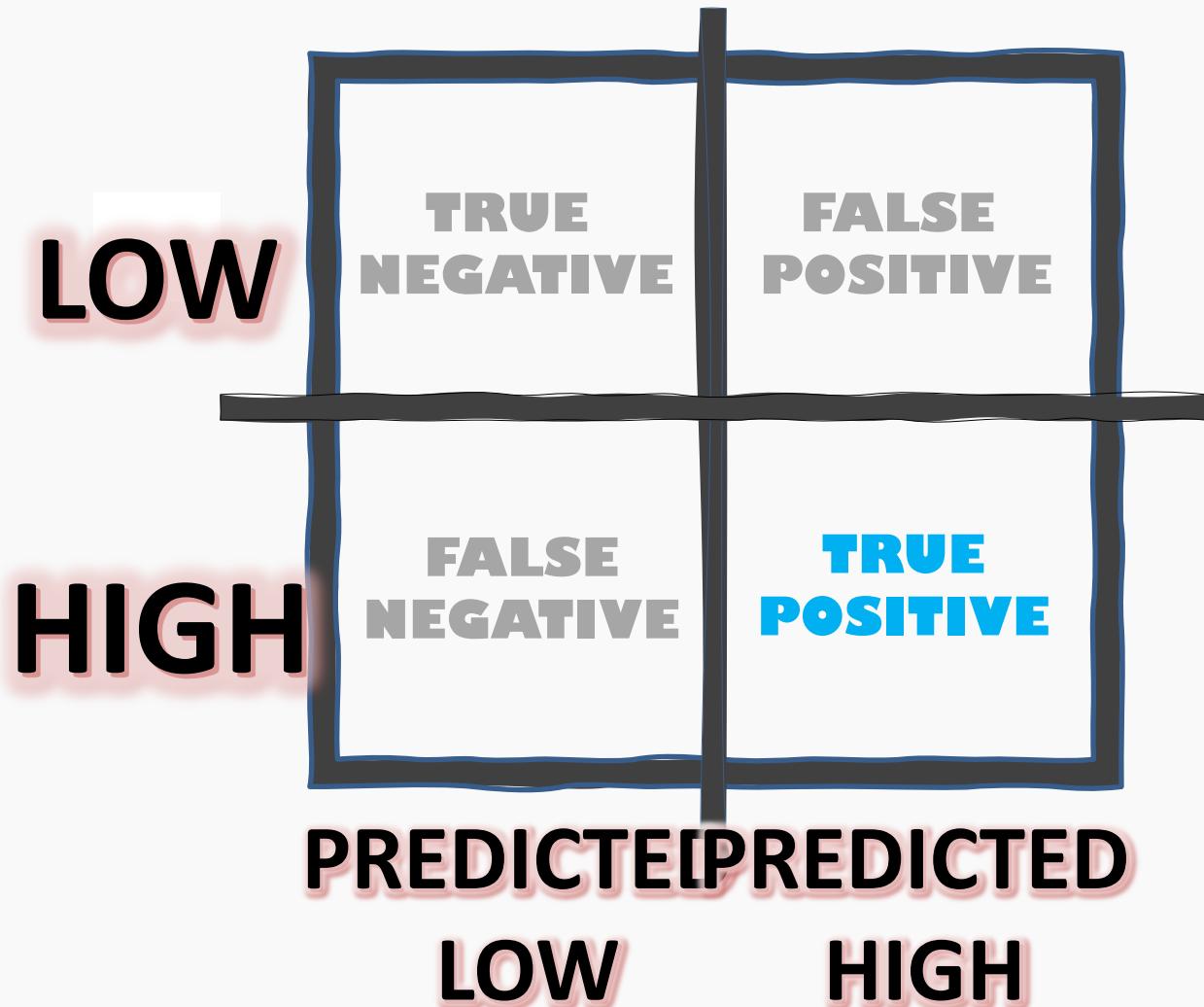
# The 'Confusion' Matrix



# The 'Confusion' Matrix

## TRUE POSITIVE (TP)

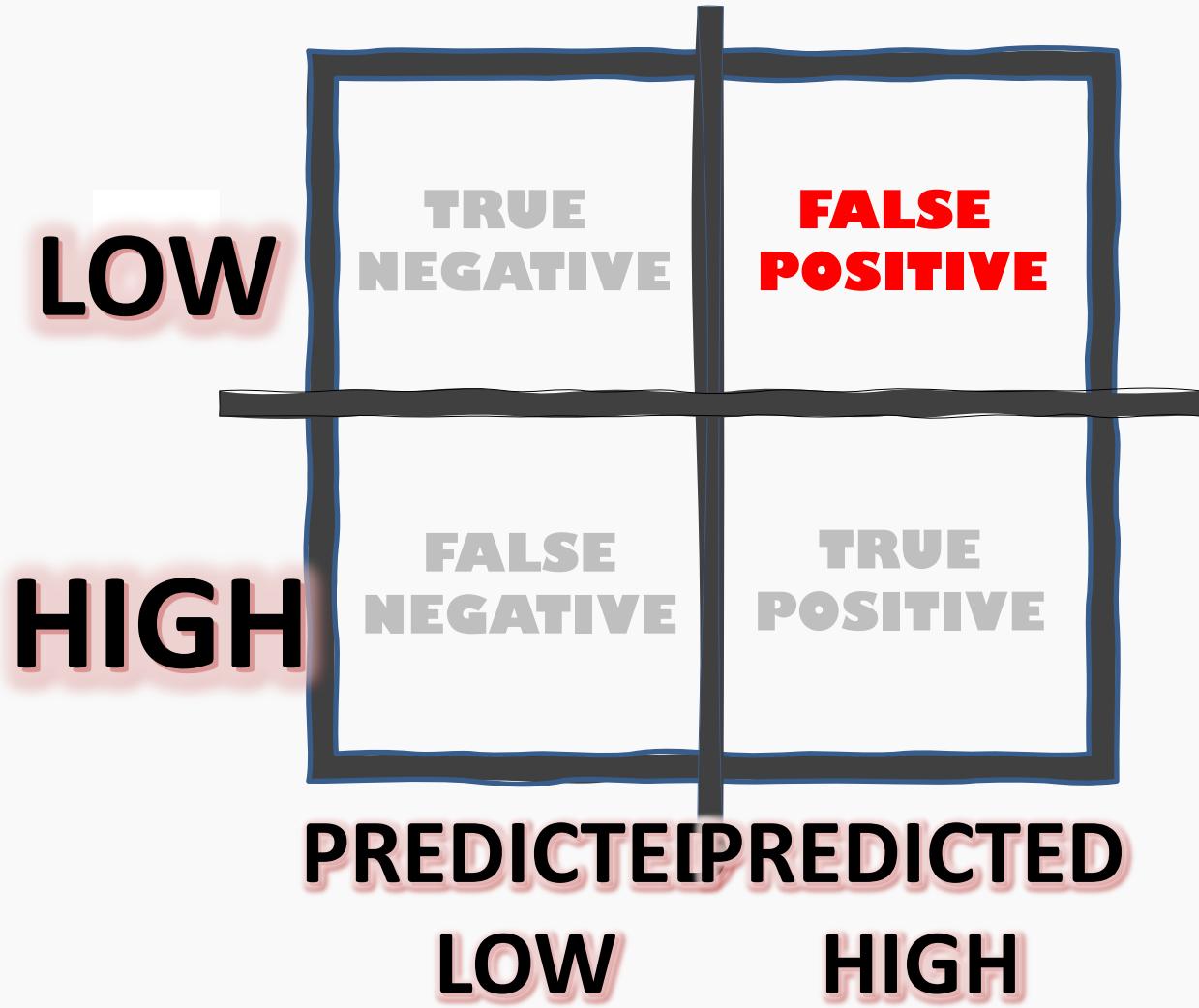
- Samples that are positive and the classifier predicts them as positive are called True Positives.
- For eg. a positive Covid test result would be a TRUE POSITIVE if you actually have Covid.



# The 'Confusion' Matrix

## FALSE POSITIVE (FP)

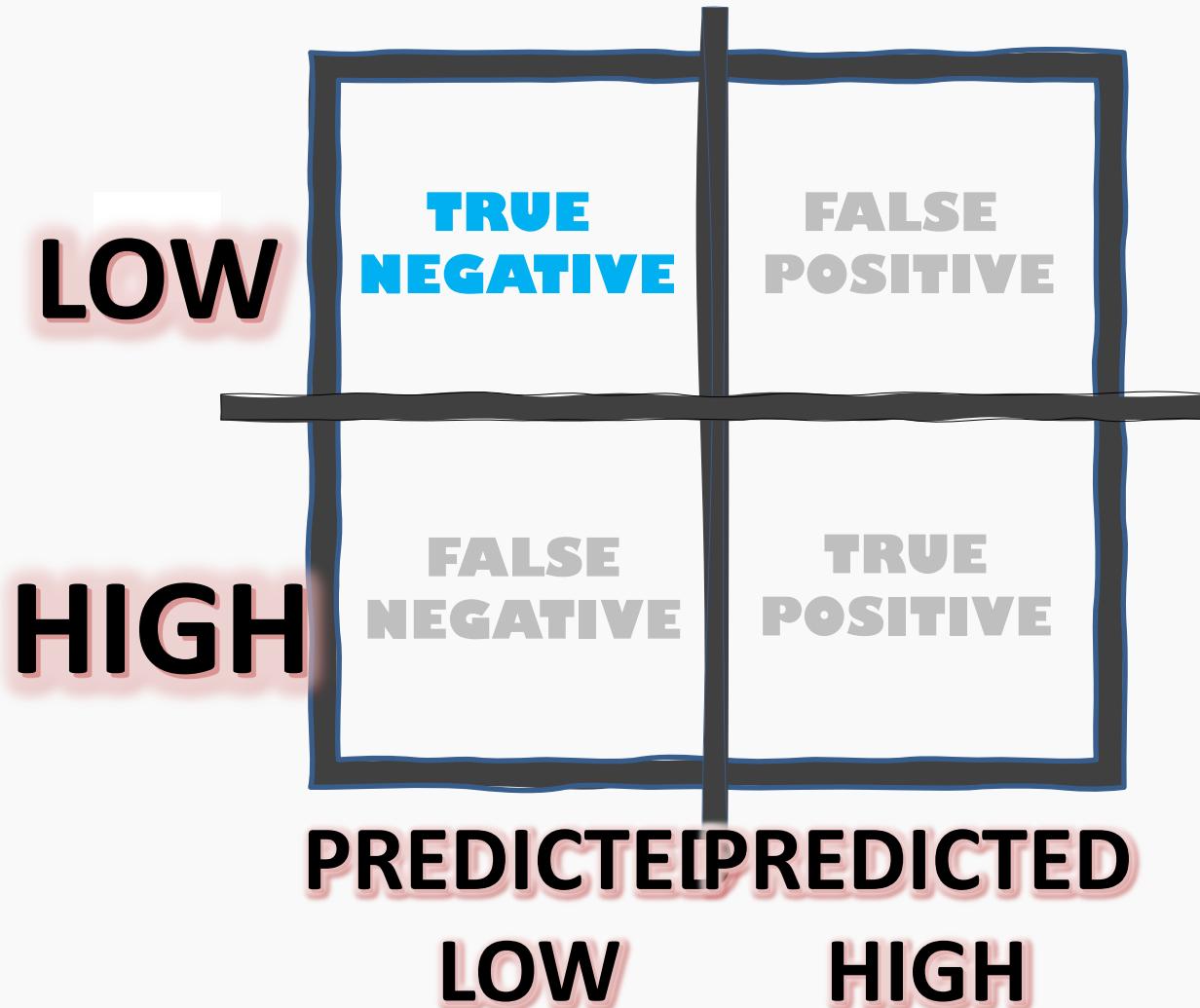
- Samples that are negative and the classifier predicts them as positive are called False Positives.
- For eg. a positive Covid test result would be a FALSE POSITIVE if you actually don't have Covid.



# The 'Confusion' Matrix

## TRUE NEGATIVE (TN)

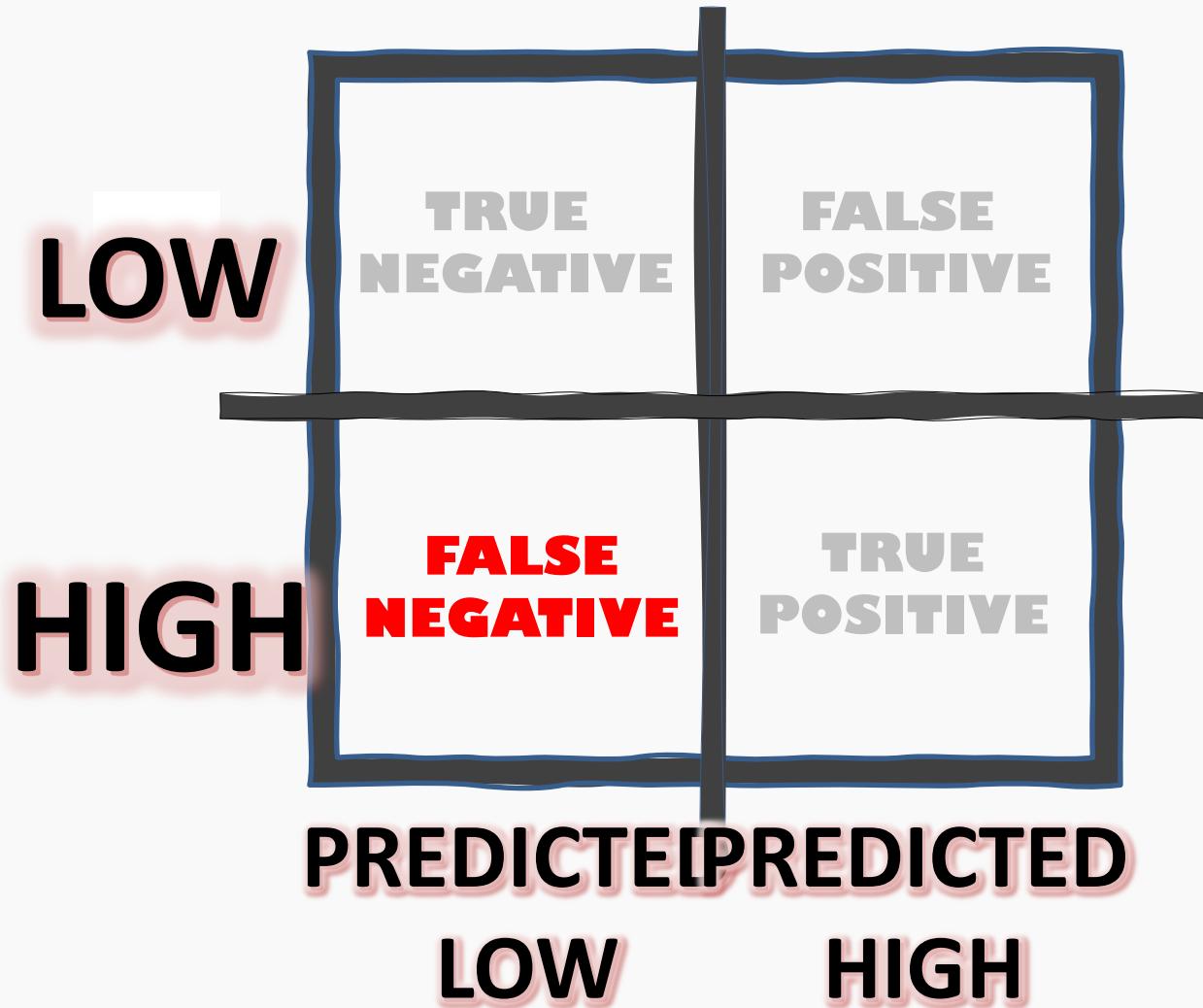
- Samples that are negative and the classifier predicts them as negative are called True Negatives.
- For eg. a negative Covid test result would be a TRUE NEGATIVE if you actually don't have Covid.



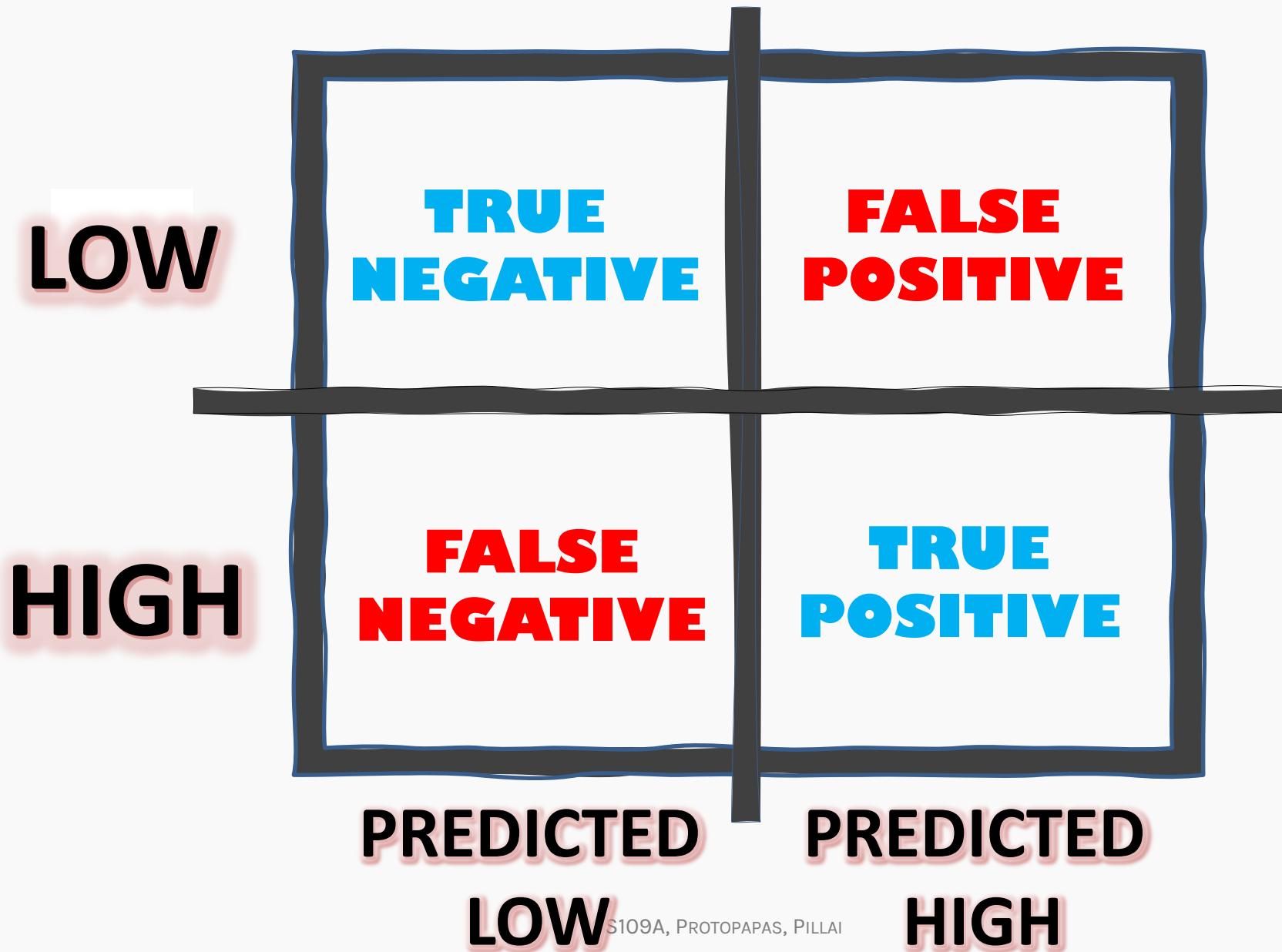
# The 'Confusion' Matrix

## FALSE NEGATIVE (FN)

- Samples that are negative and the classifier predicts them as positive are called False Negatives.
- For eg. a negative Covid test result would be a FALSE NEGATIVE if you actually have Covid.



# The 'Confusion' Matrix

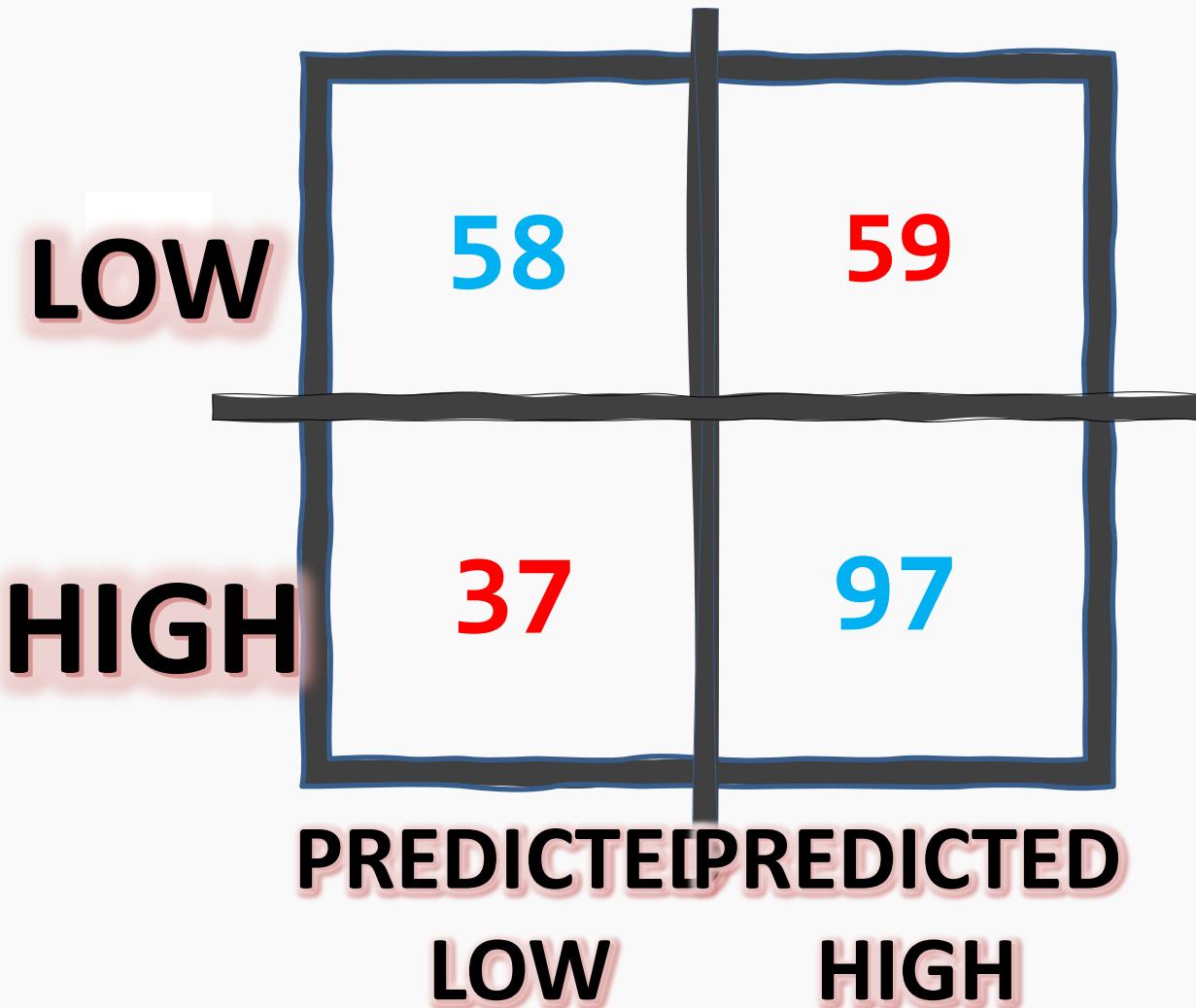


# Let's Begin



# The 'Confusion' Matrix

## Logistic Regression



# The 'Confusion' Matrix

## kNN Classification



		PREDICTED	
		LOW	HIGH
PREDICTED	LOW	55	62
	HIGH	33	101

# The 'Confusion' Matrix

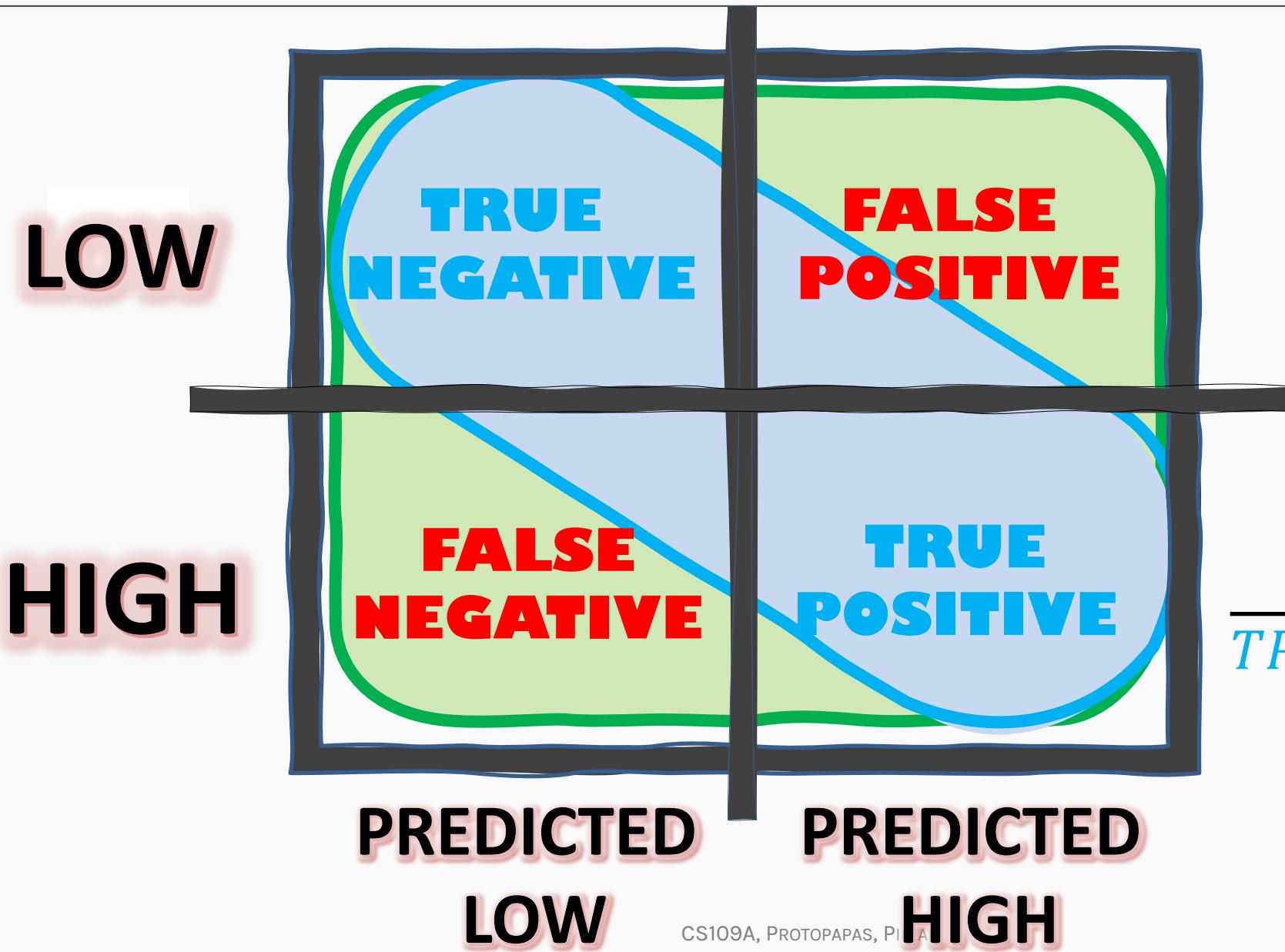
Logistic Regression

		PREDICTED	PREDICTED
		LOW	HIGH
ACTUAL	LOW	58	59
	HIGH	37	97

kNN Classification

		PREDICTED	PREDICTED
		LOW	HIGH
ACTUAL	LOW	55	62
	HIGH	33	101

# Accuracy



$$\frac{TN + TP}{TP + TN + FP + FN}$$

# Accuracy

Logistic Regression

		PREDICTED	PREDICTED
		LOW	HIGH
PREDICT	LOW	58	37
HIGH	HIGH	59	97

$$\text{Accuracy} = \frac{58+97}{58+97+37+59} = 0.62$$

kNN Classification

		PREDICTED	PREDICTED
		LOW	HIGH
PREDICT	LOW	55	33
HIGH	HIGH	62	101

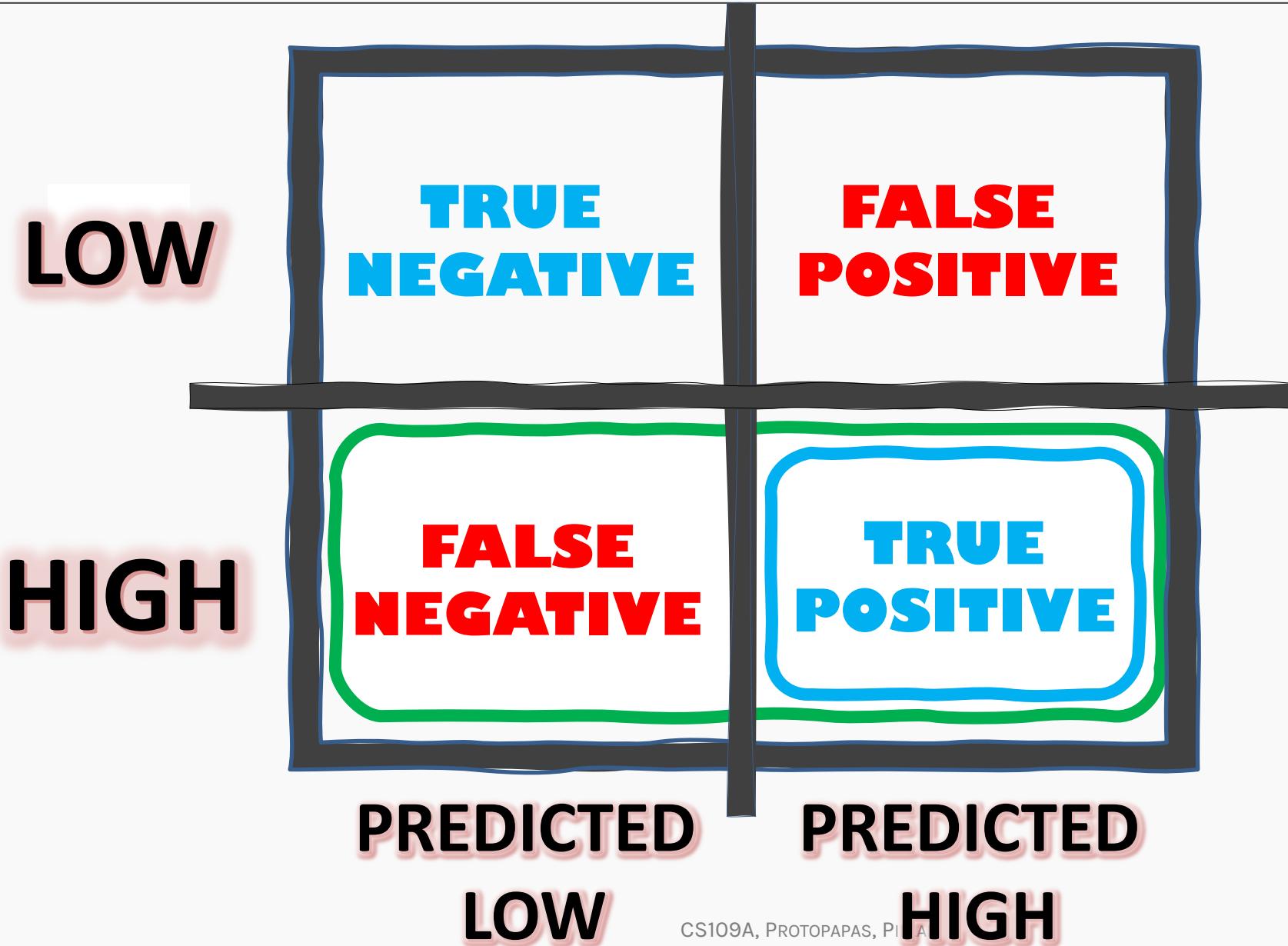
$$\text{Accuracy} = \frac{55+101}{55+101+33+62} = 0.62$$

# Model Comparison – Logistic vs kNN

Classification Metric	Formula	Logistic Regression	kNN Classification
Accuracy	$\frac{TN + TP}{TP + TN + FP + FN}$		
Sensitivity (Recall)			
Specificity			
Precision			
F1 score			



# Sensitivity/True Positive Rate/Recall

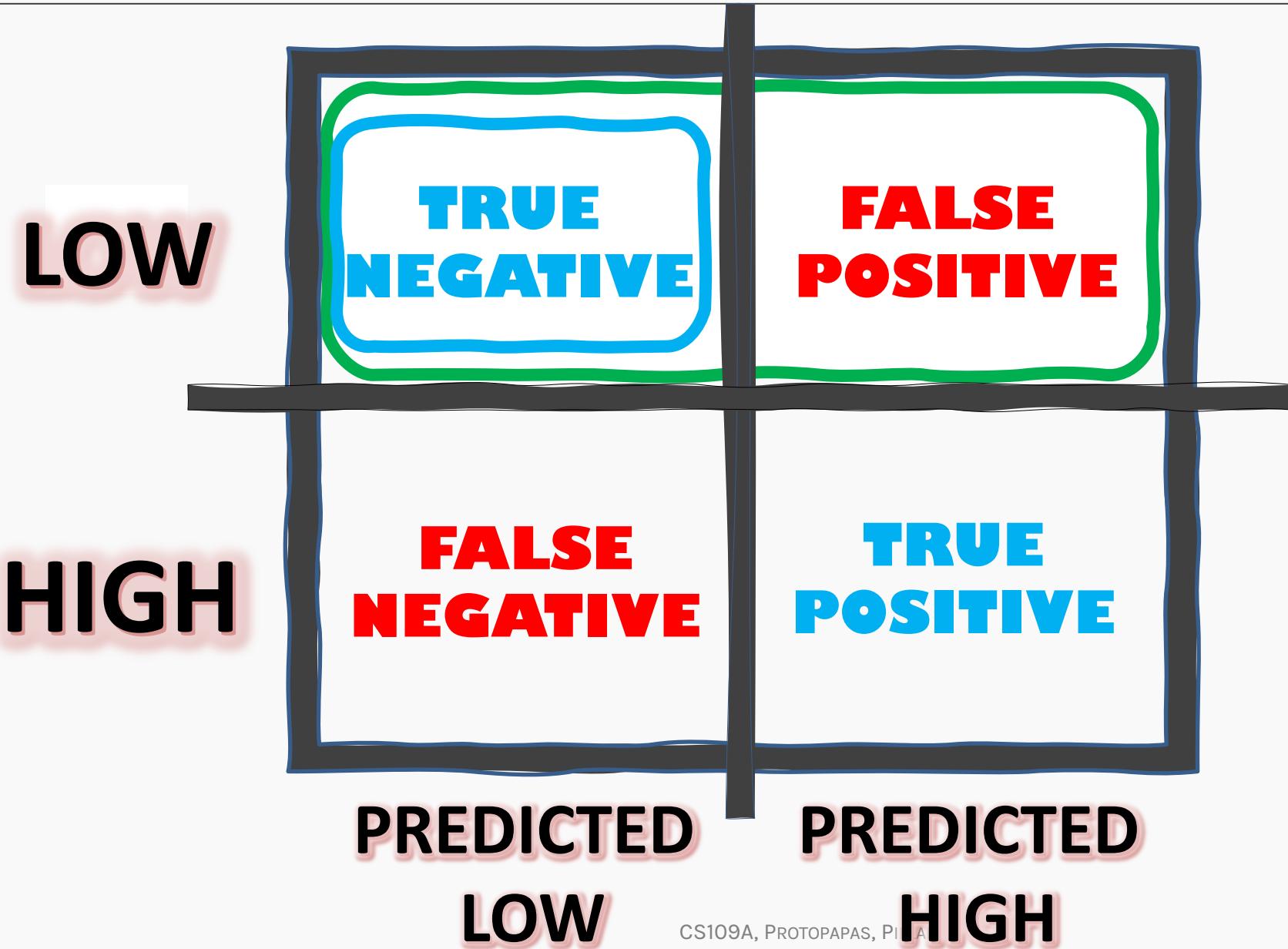


$$\frac{TP}{TP + FN}$$

# Model Comparison – Logistic vs kNN

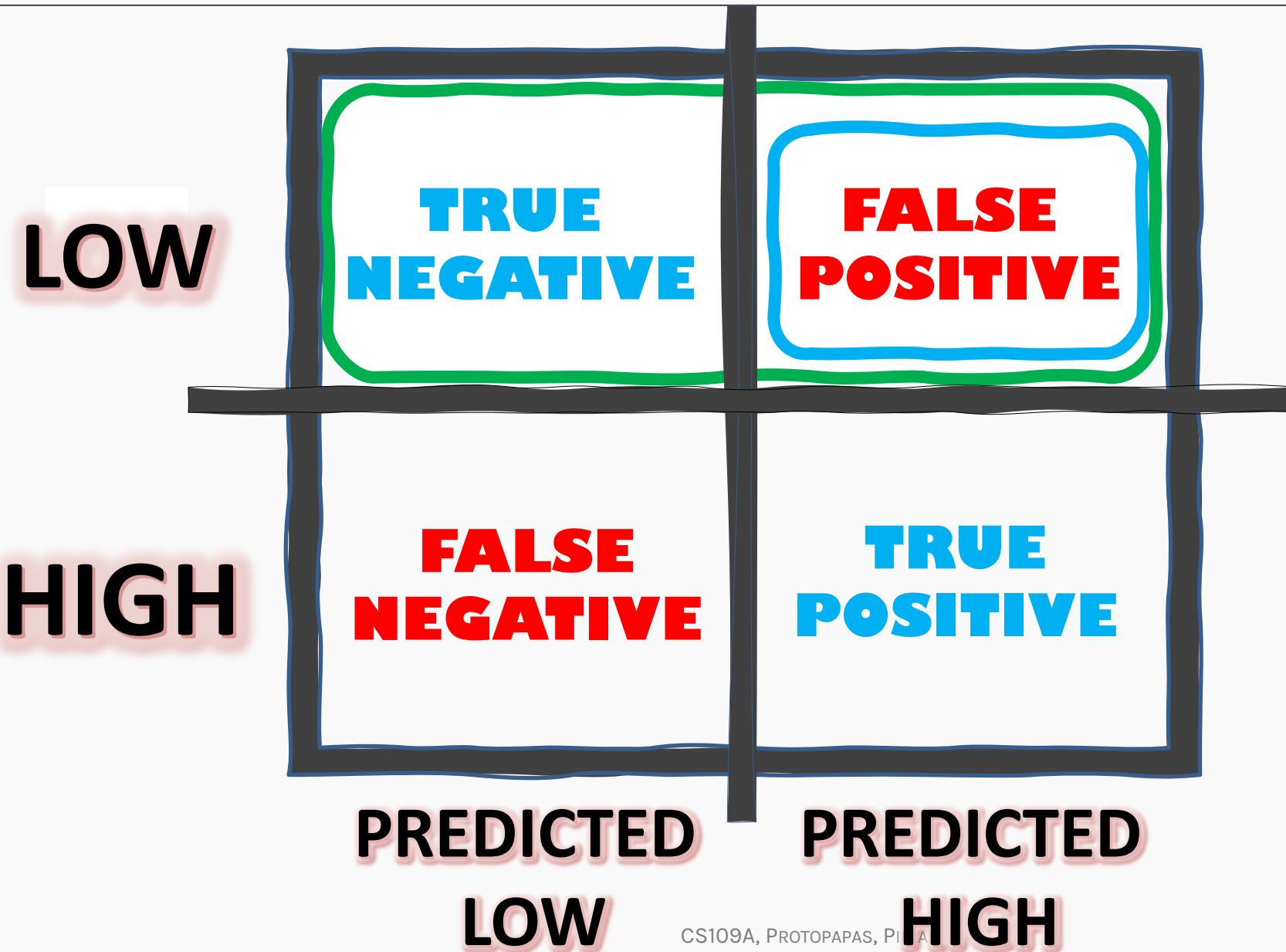
Classification Metric	Formula	Logistic Regression	kNN Classification
Accuracy	$\frac{TN + TP}{TP + TN + FP + FN}$		
Sensitivity (Recall)	$\frac{TP}{TP + FN}$		
Specificity			
Precision			
F1 score			

# Specificity/True Negative Rate



$$\frac{TN}{TN + FP}$$

# False Positive Rate

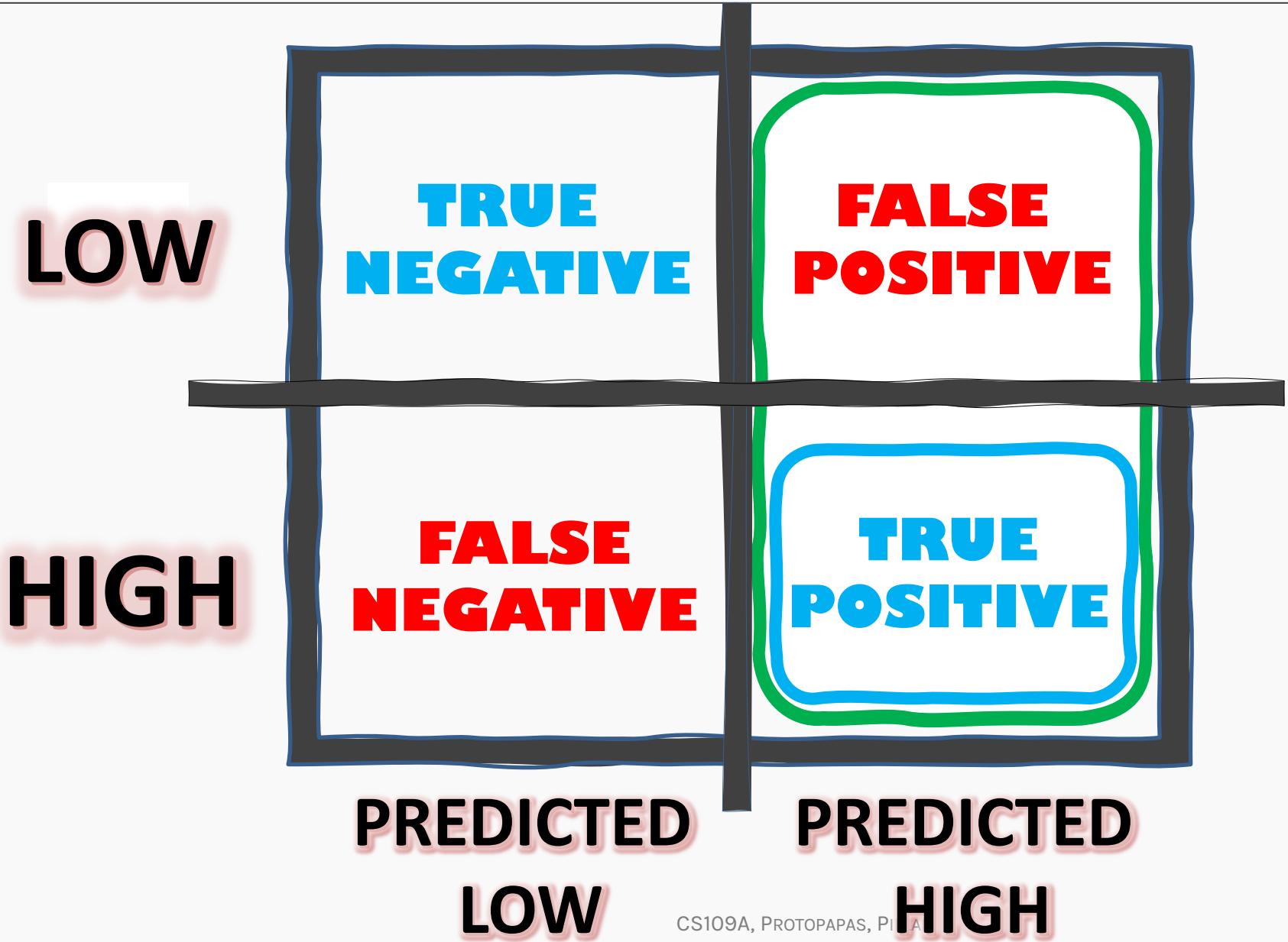


$$\frac{FP}{TN + FP}$$

# Model Comparison – Logistic vs kNN

Classification Metric	Formula	Logistic Regression	kNN Classification
Accuracy	$\frac{TN + TP}{TP + TN + FP + FN}$		
Sensitivity (Recall)	$\frac{TP}{TP + FN}$		
Specificity	$\frac{TN}{TN + FP}$		
Precision			
F1 score			

# Precision



$$\frac{TP}{TP + FP}$$

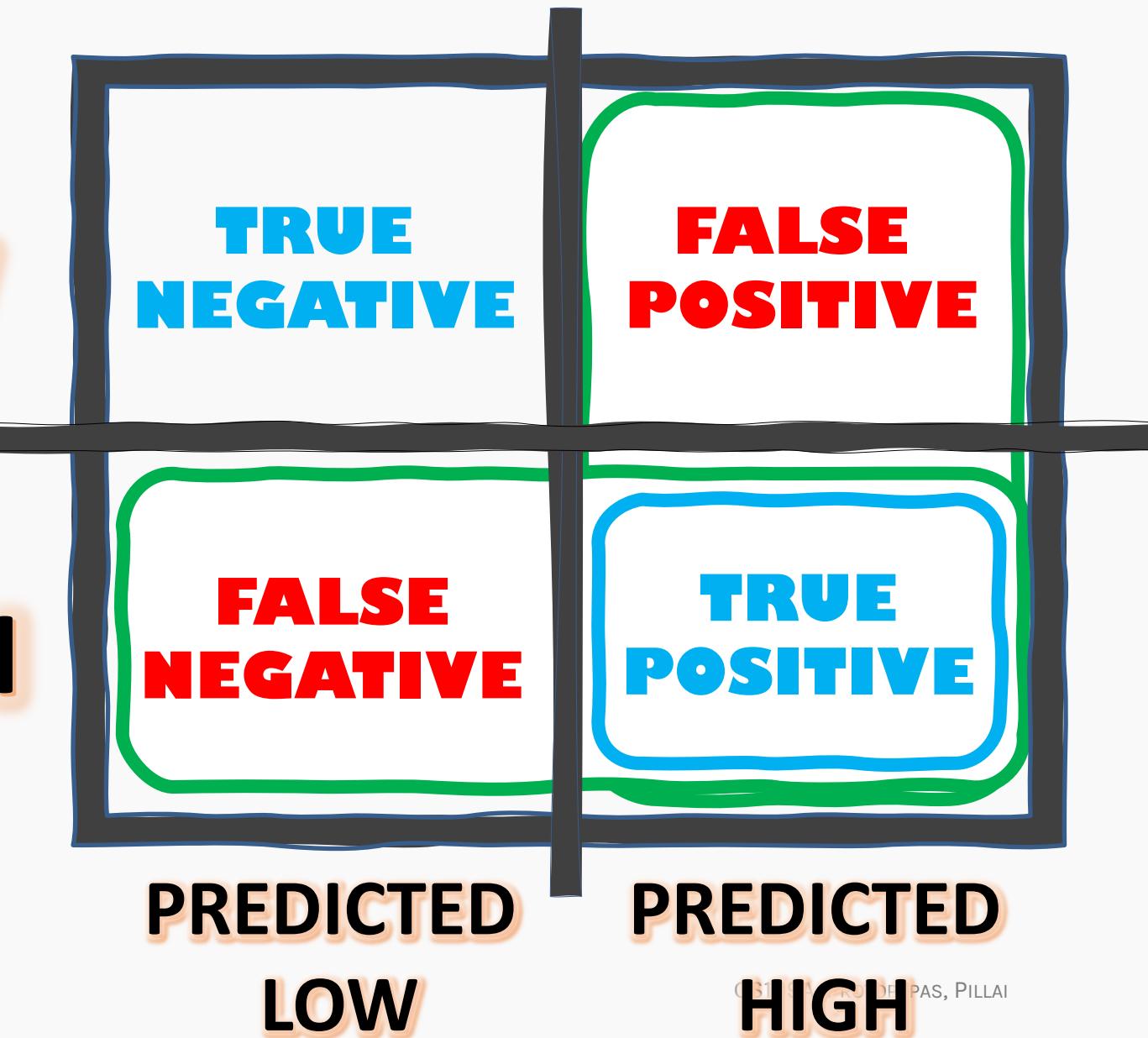
# Model Comparison – Logistic vs kNN

Classification Metric	Formula	Logistic Regression	kNN Classification
Accuracy	$\frac{TN + TP}{TP + TN + FP + FN}$		
Sensitivity (Recall)	$\frac{TP}{TP + FN}$		
Specificity	$\frac{TN}{TN + FP}$		
Precision	$\frac{TP}{TP + FP}$		
F1 score			



# F1-score

LOW  
HIGH



$$F1 \text{ score} = \frac{(2 * Precision * Recall)}{Precision + Recall}$$

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

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

# Model Comparison – Logistic vs kNN

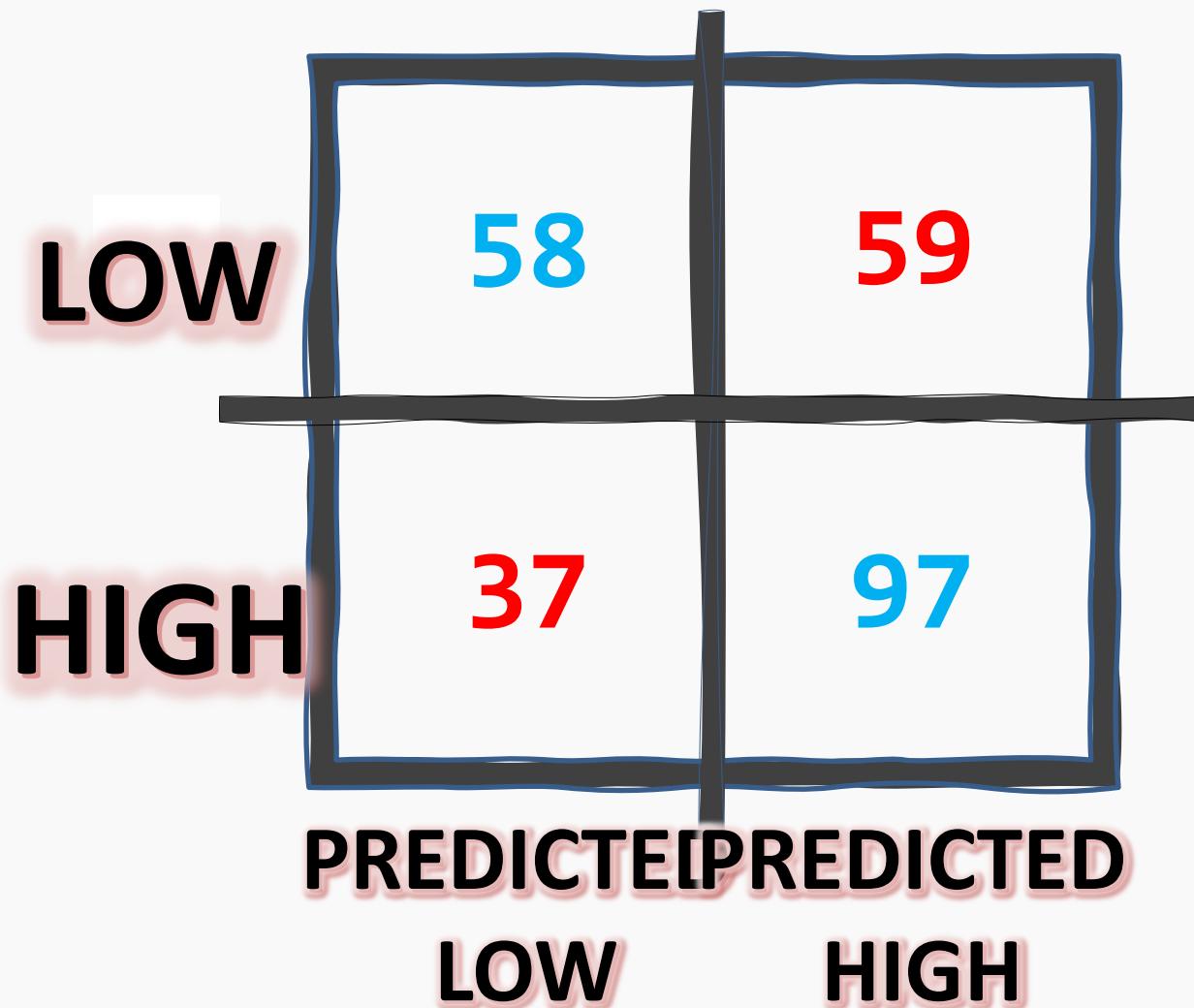
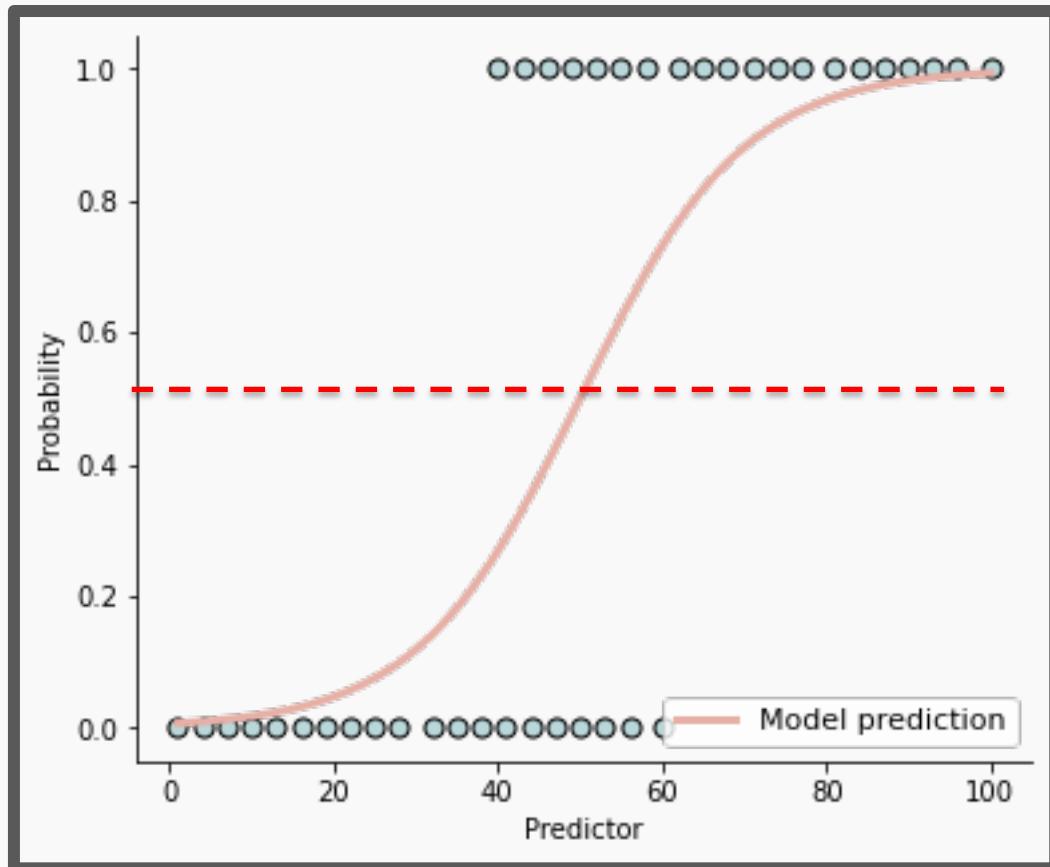
Classification Metric	Formula	Logistic Regression	kNN Classification
Accuracy	$\frac{TN + TP}{TP + TN + FP + FN}$		
Sensitivity (Recall)	$\frac{TP}{TP + FN}$		
Specificity	$\frac{TN}{TN + FP}$		
Precision	$\frac{TP}{TP + FP}$		
F1 score	$\frac{(2 * Precision * Recall)}{Precision + Recall}$		

# Bayes threshold



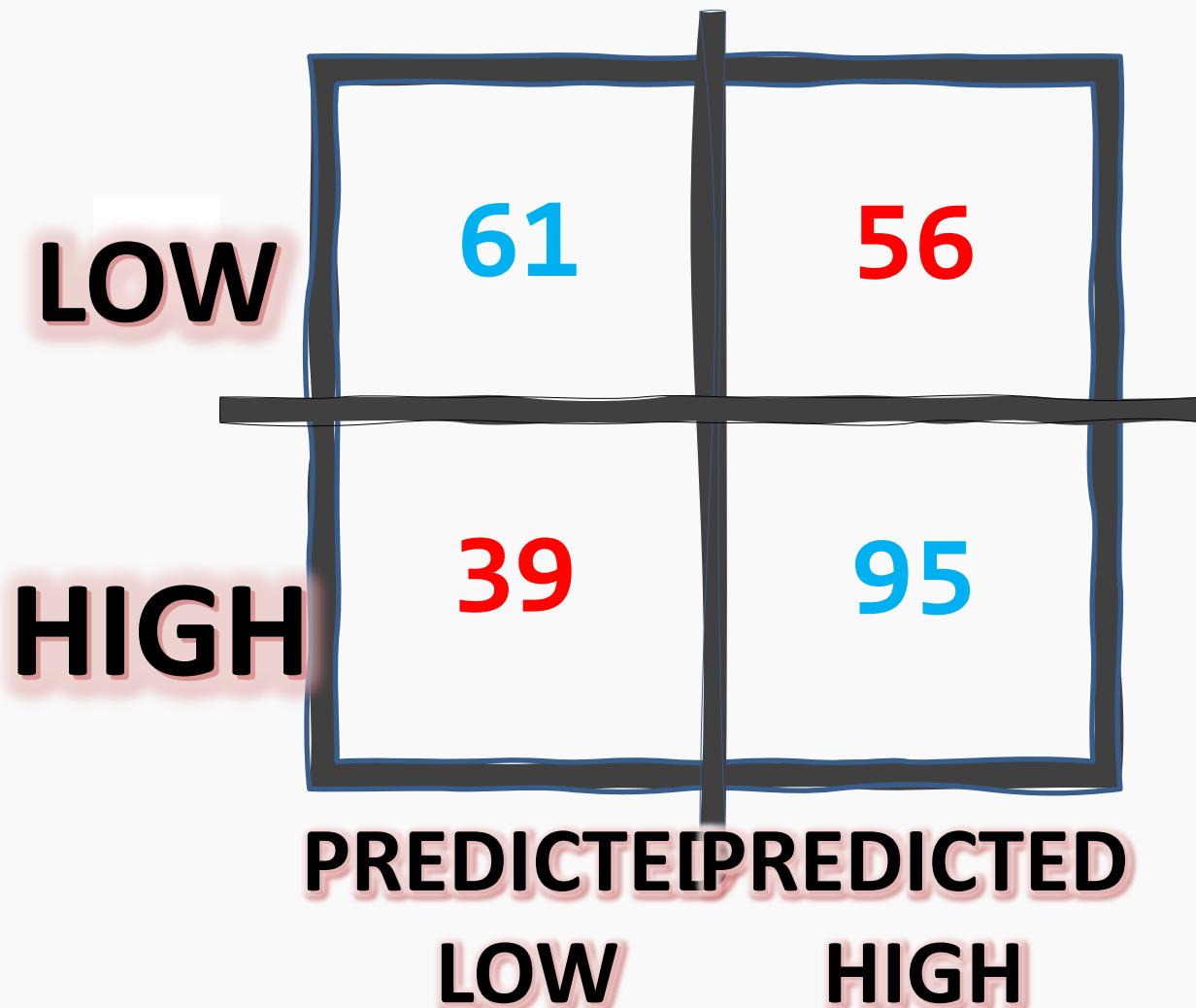
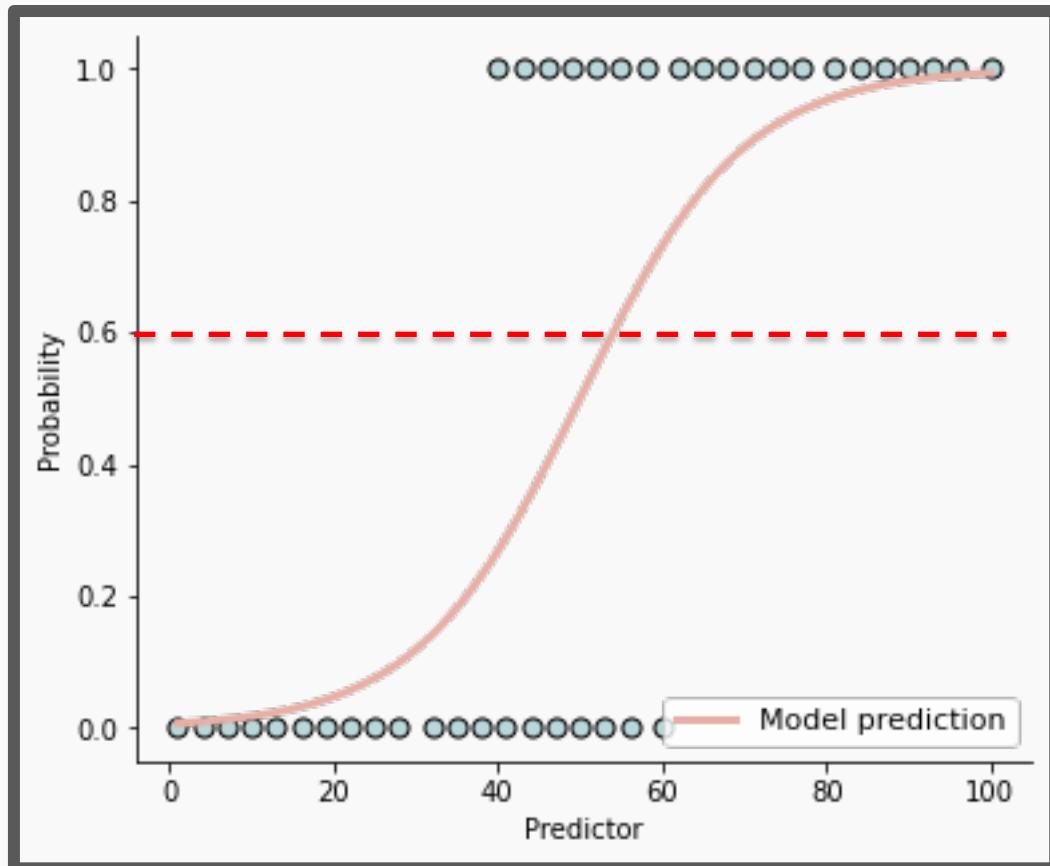
# Bayes Threshold

## Logistic Regression



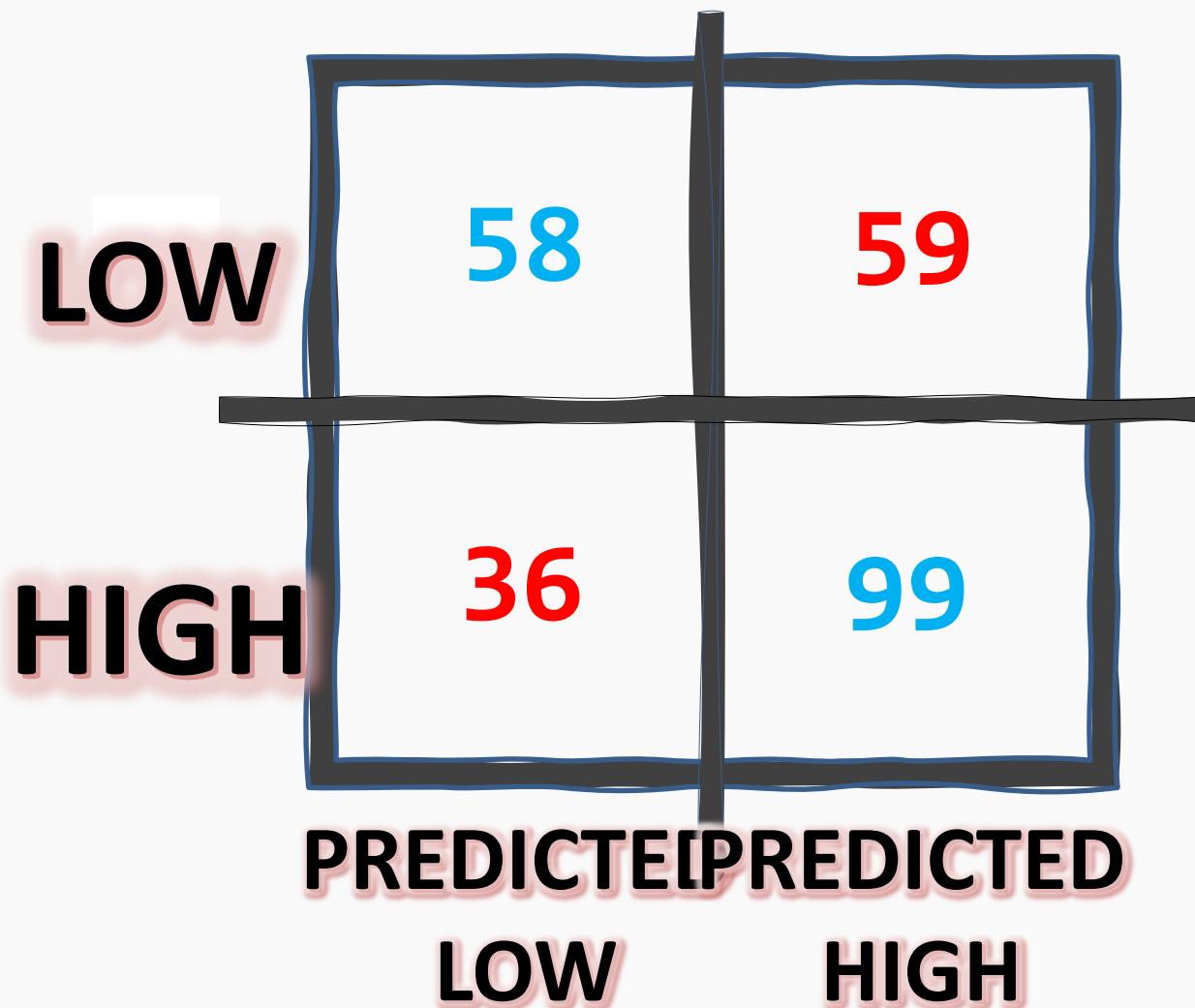
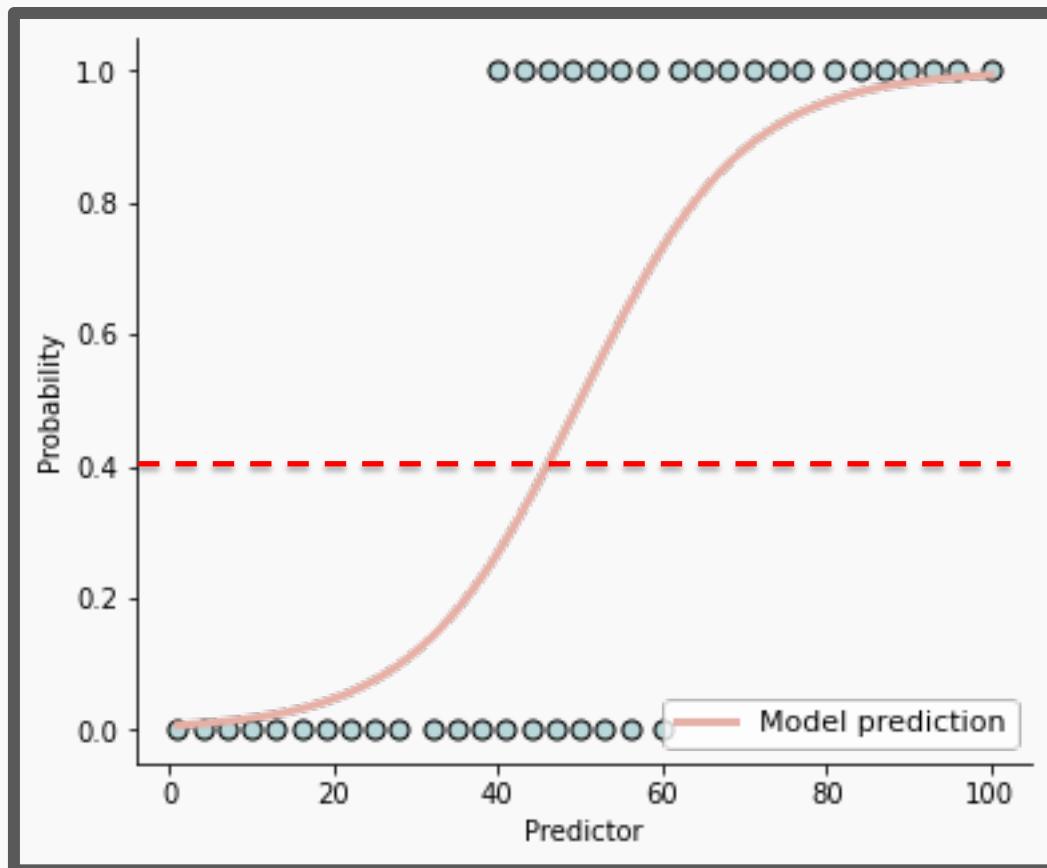
# Bayes Threshold

## Logistic Regression



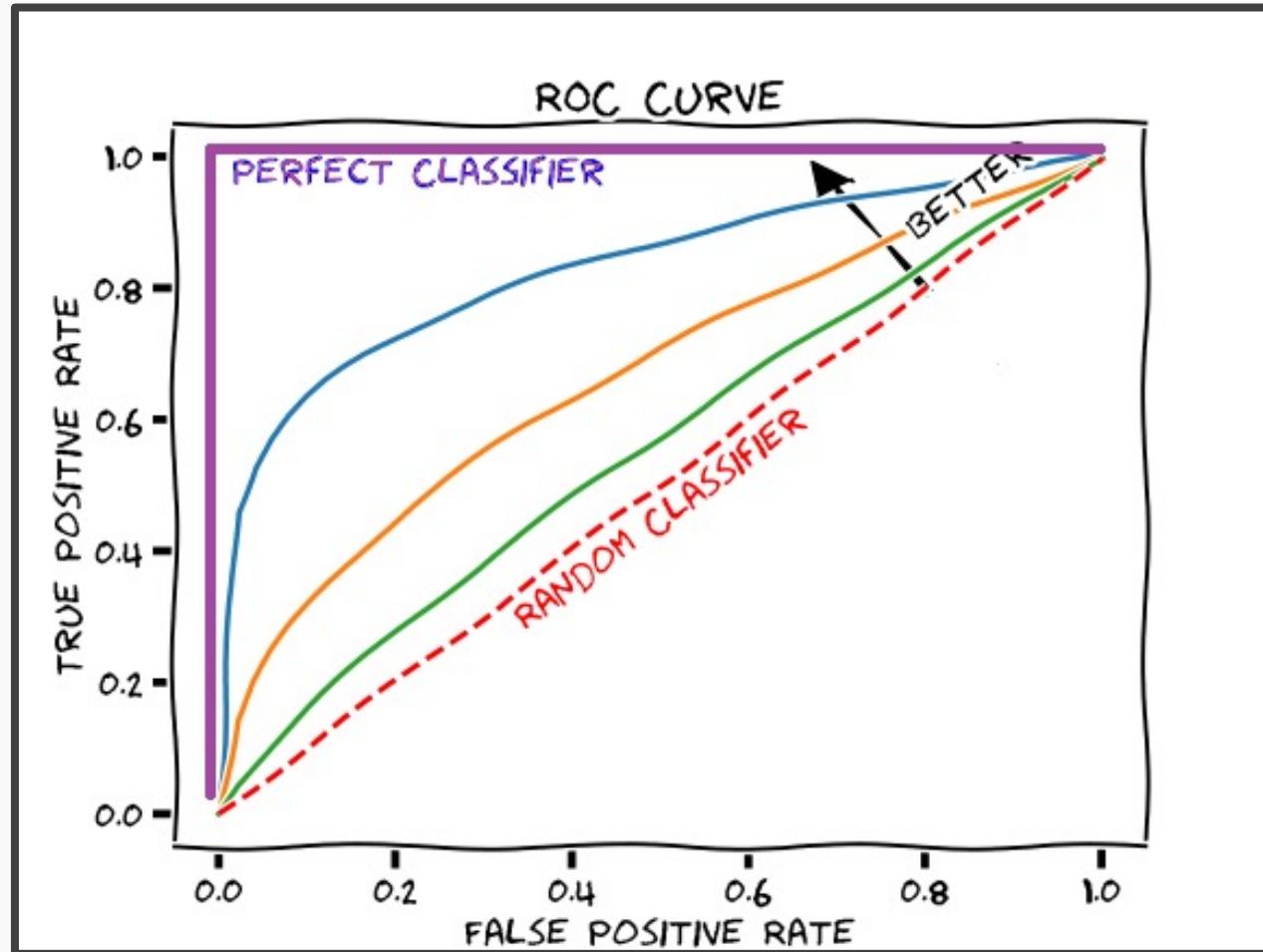
# Bayes Threshold

## Logistic Regression

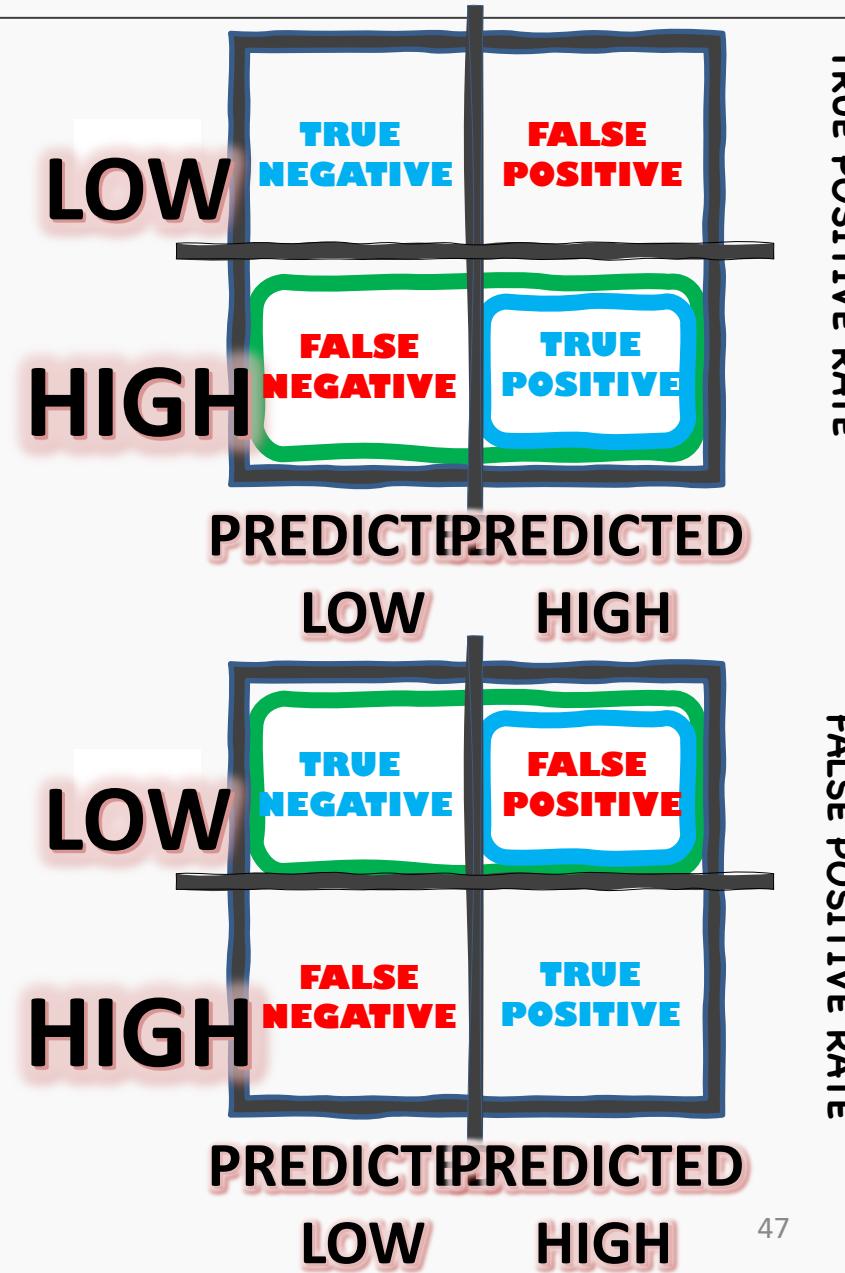
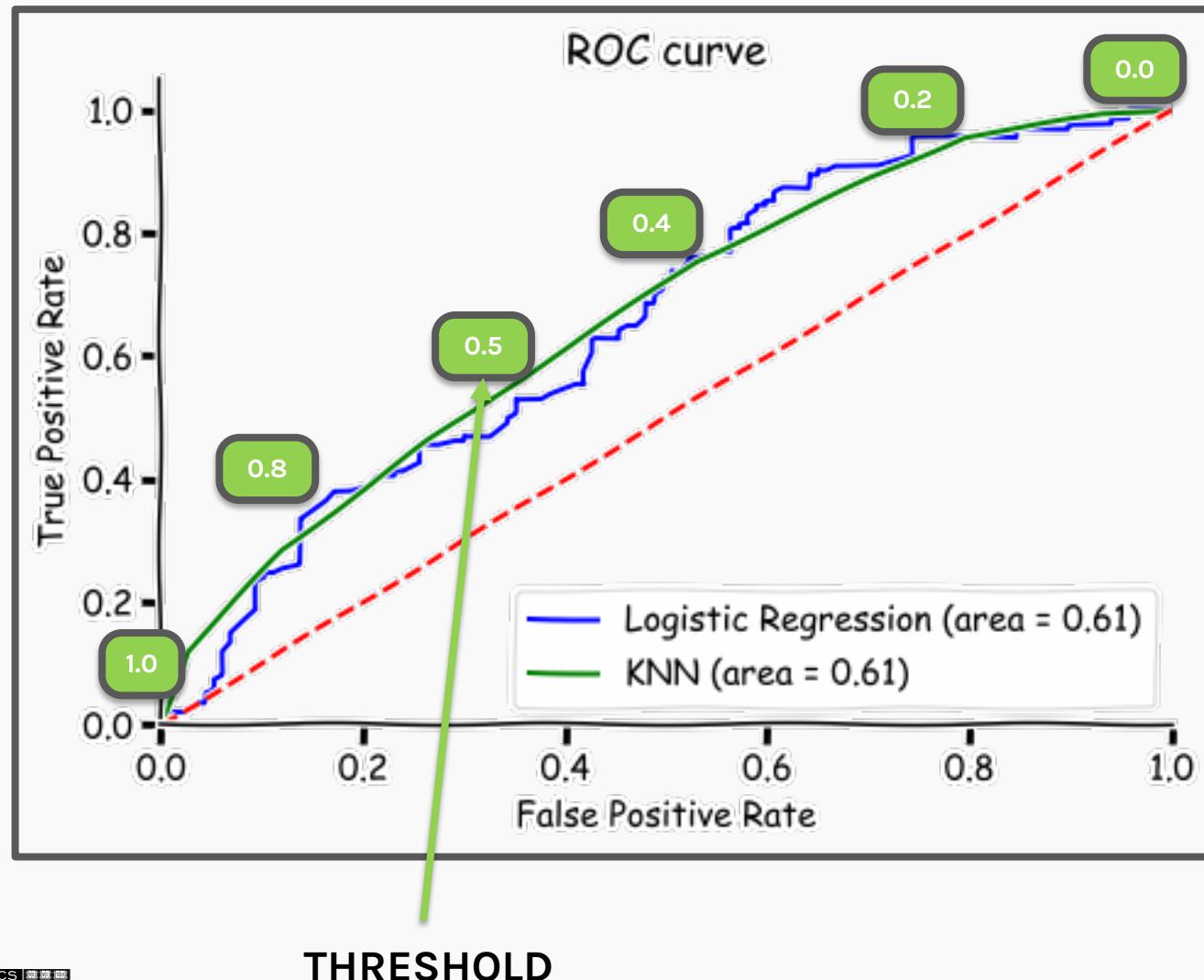


# Receiver Operating Characteristic curve (ROC)

- The ROC curve was first developed by radar engineers during World War II for detecting enemy objects in battlefields.
- The ROC curve is created by plotting the **true positive rate (TPR)** against the **false positive rate (FPR)** at various threshold settings.
- If used correctly, ROC curves are a very powerful tool as a statistical performance measure in detection/classification theory.



# ROC curve for various thresholds



Two models

---

## Logistic Regression



v/s

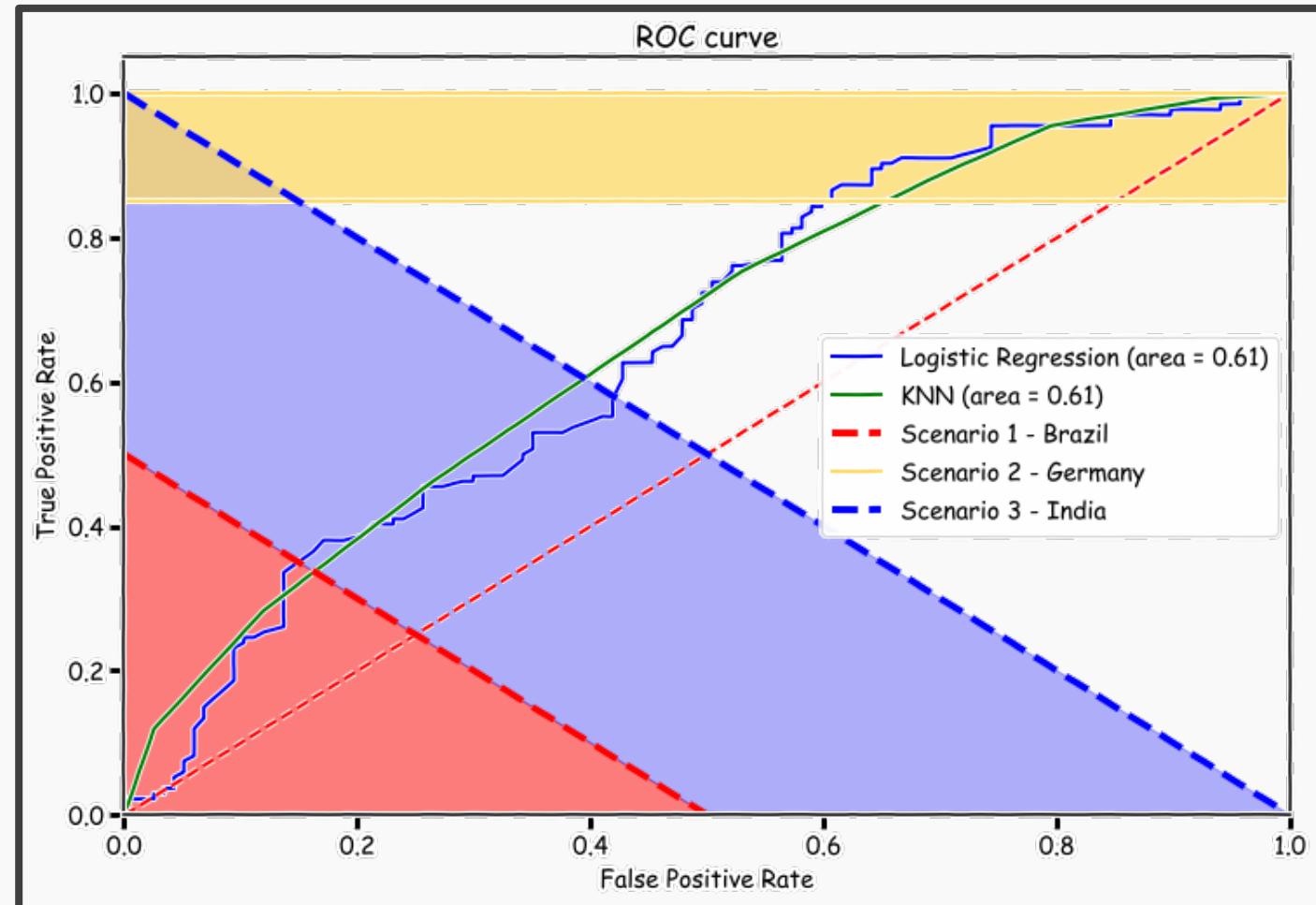
## kNN Classification



# Choice of Classifier

Based on the constraints we have the following choice of classifier:

- BRAZIL: Logistic regression with a high threshold
- GERMANY: Logistic regression with a low threshold
- INDIA: kNN classifier with a moderate threshold



The choice of classifier depends on the constraints and the threshold value.