PRINCIPLES OF MACHINE LEARNING IN MEDICINE

Evaluation of Binary Classifiers

09/05/2022

Neslihan Bayramoglu

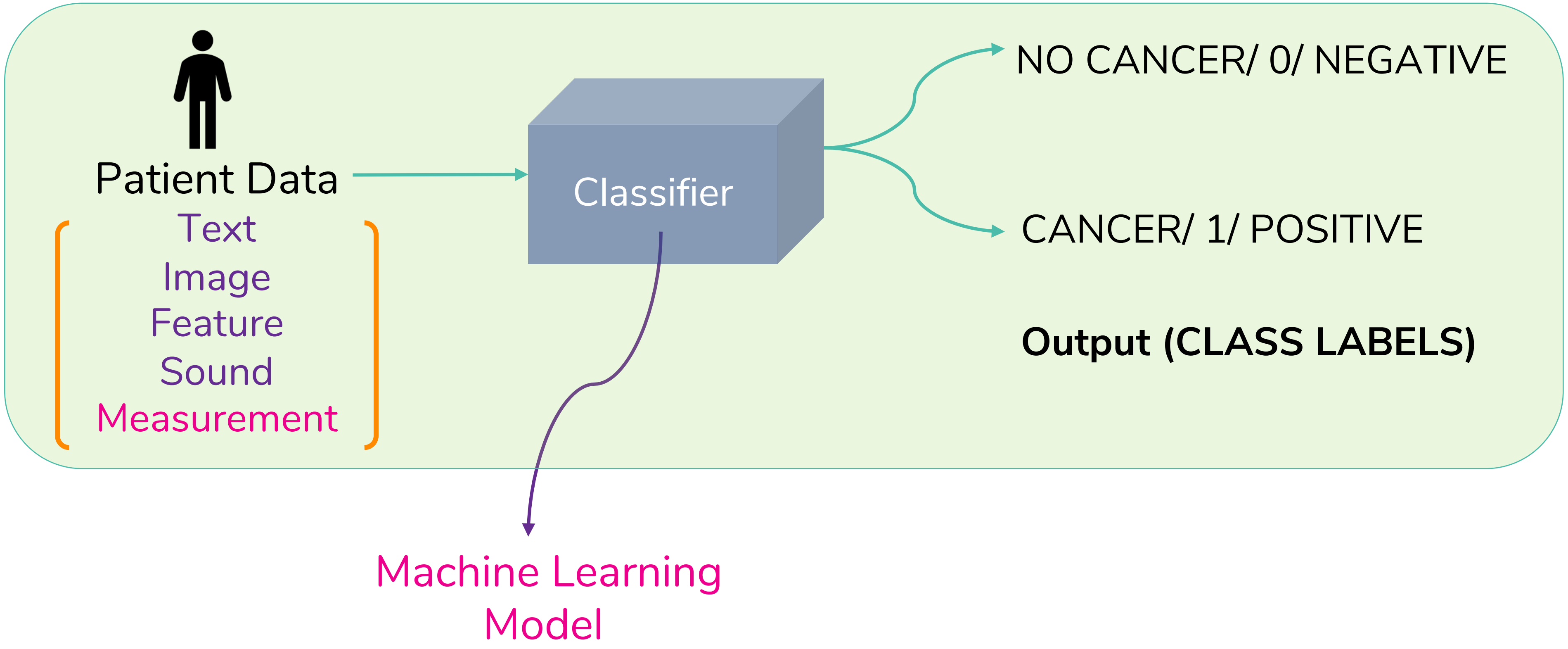
Docentship Demo Lecture

In this lecture

- Overview of
- Binary classifiers
- Discrete vs probabilistic classifiers
- Comparing different machine learning algorithms
- Performance metrics
- Accuracy
- Confusion matrix
- ROC
- Precision Recall

Binary Classifiers

- classifying the data into two groups
- **a large number of medical studies are based on classification models**



Types of Binary Classifiers

Classification Model

Hard Classifier
Class Label

Soft Classifier
Probability Output

e.g. for Binary classification (class labels: 0, 1)

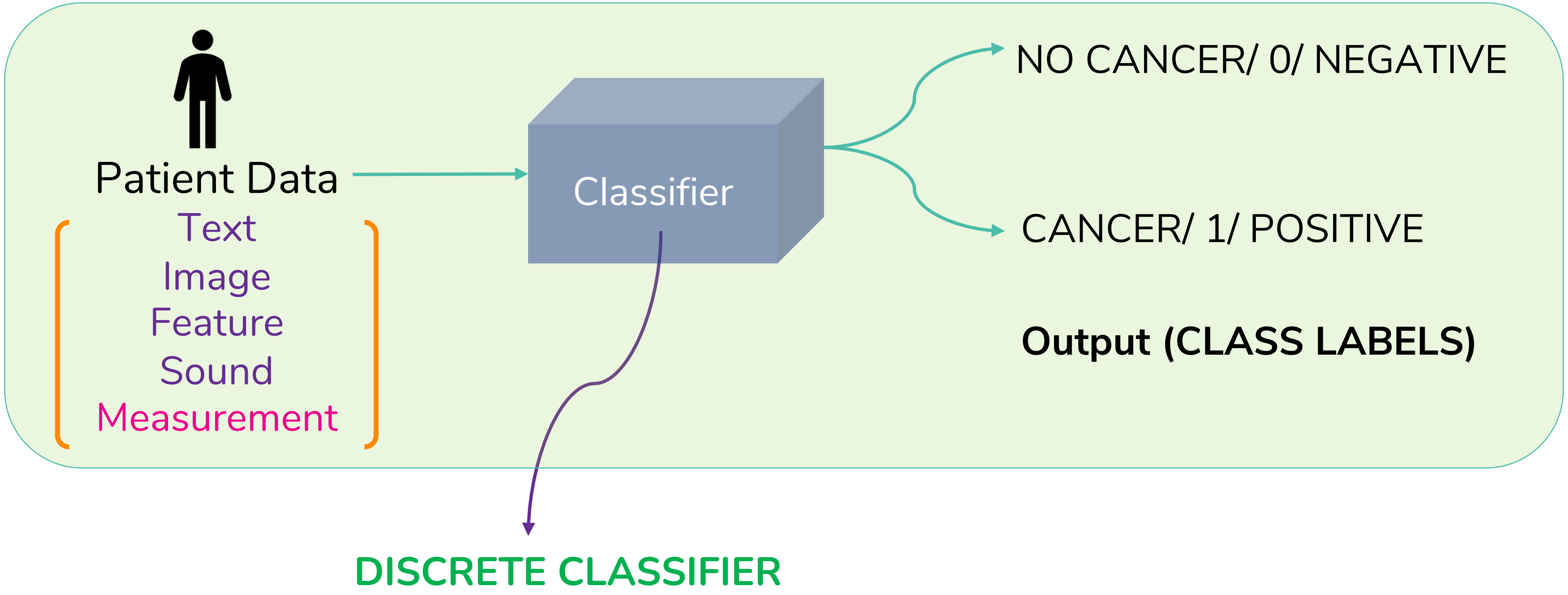
Output: 0 or 1

e.g. for Binary classification (class labels: 0,1)

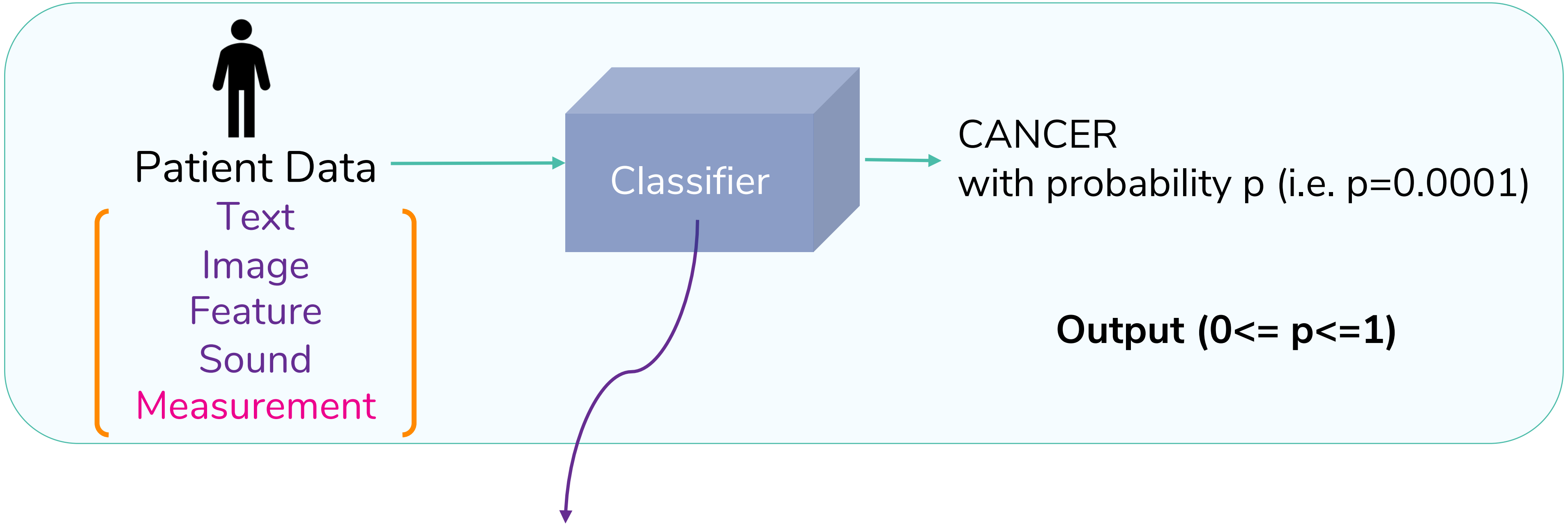
Output:
P(input=class 0)= p
P(input=class 1)= 1-p

Threshold the output to obtain hard decisions

Hard (Discrete) Classifier

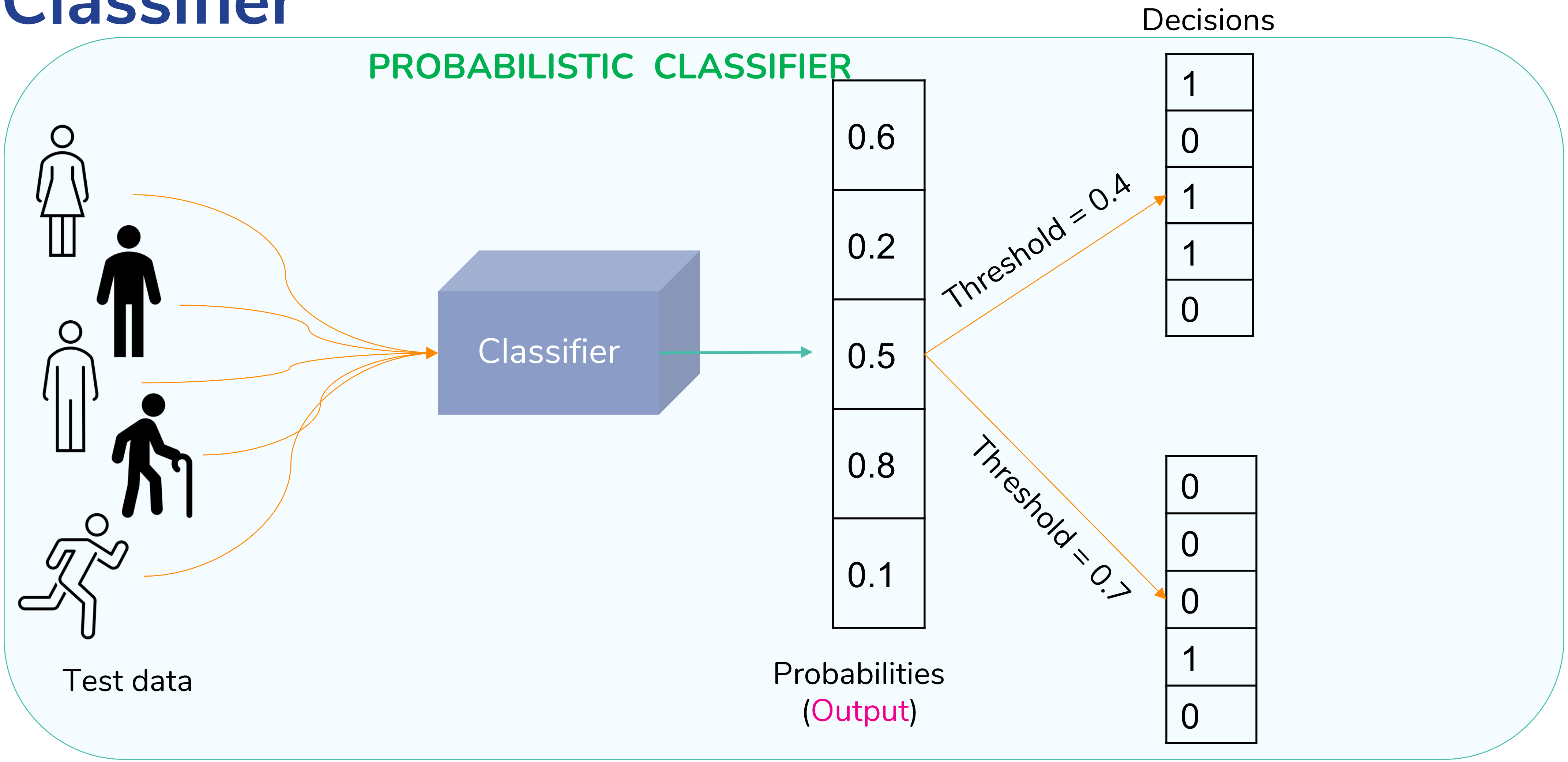


Soft (Probabilistic) Classifier

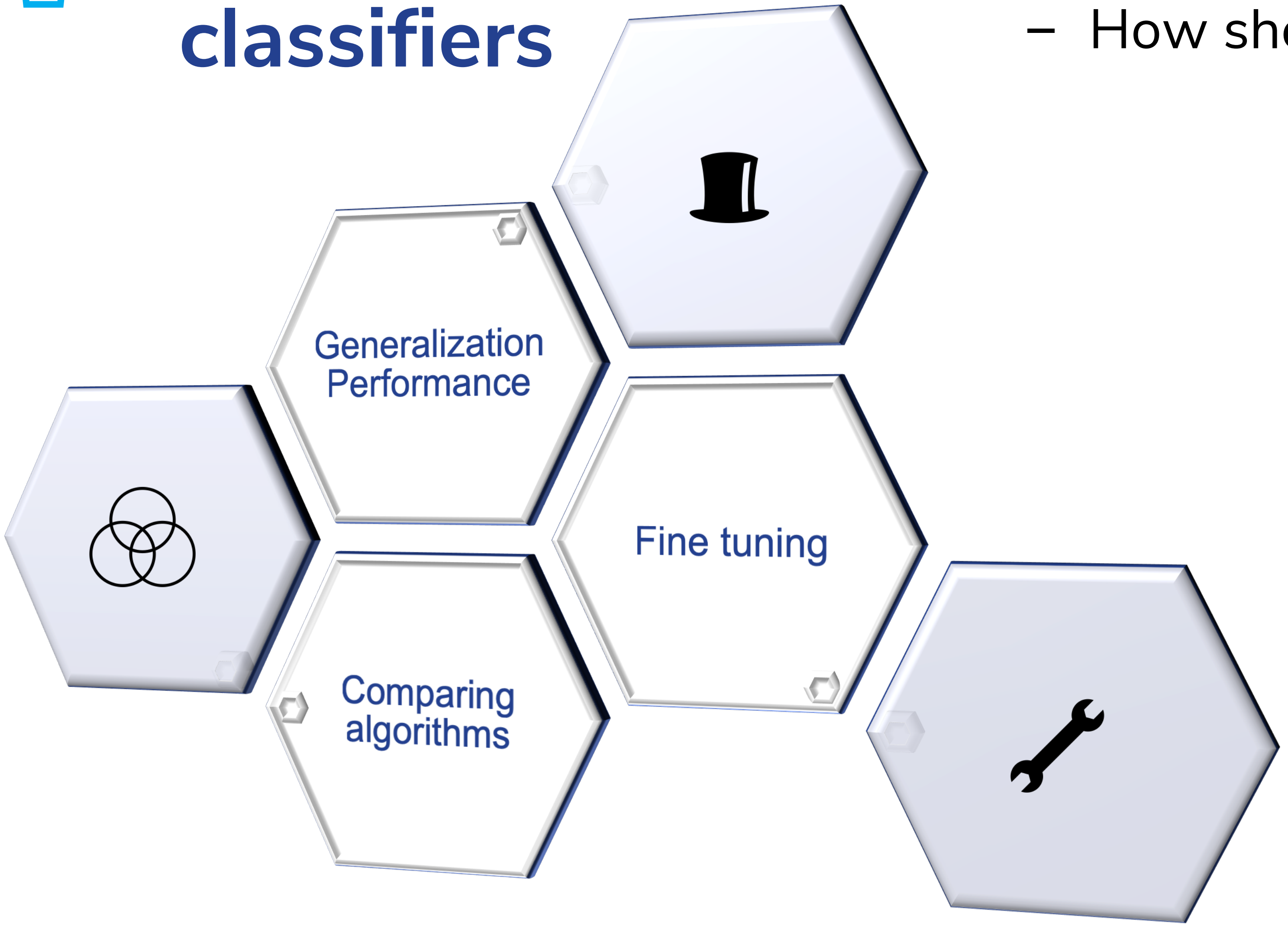


PROBABILISTIC CLASSIFIER

Soft (Probabilistic) Classifier



Performance of classifiers



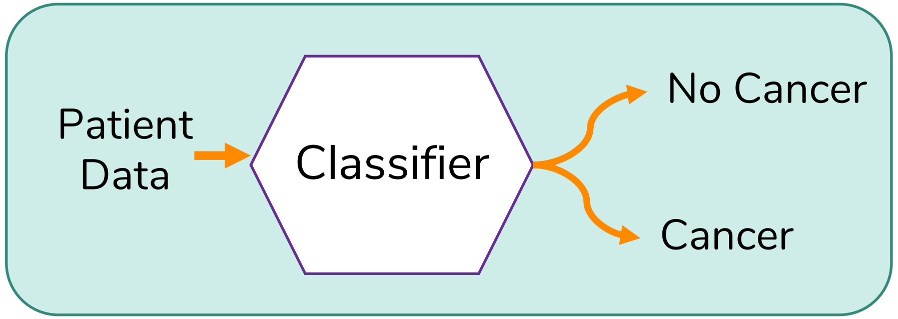
- The most important task
- How should we evaluate

Performance Metrics

Compare predicted labels and true labels
OR
interpret the predicted probabilities

1. Confusion Martix
2. False positive rate | Type-I error
3. False negative rate | Type-II error
4. True negative rate | Specificity
5. Negative predictive value
6. False discovery rate
7. True positive rate | Recall | Sensitivity
8. Positive predictive value | Precision
9. Accuracy
10. F beta score
11. F1 score
12. F2 score
13. Cohen Kappa
14. Matthews correlation coefficient
15. ROC curve
16. ROC AUC score
17. Precision-Recall curve
18. PR AUC | Average precision
19. Log loss
20. Brier score
21. Cumulative gain chart
22. Lift curve | Lift chart
23. Kolmogorov-Smirnov plot
24. Kolmogorov Smirnov statistics

Confusion matrix



True positive (TP
False positive (FP)- Type 1 error
True negative (TN)
False negative (FN) – Type2 error

ACTUAL

If patient have cancer or not

PREDICTION *what our model predicted*

	have cancer	doesn't have cancer
have cancer	number of **TP**	number of **FP**
doesn't have cancer	number of **FN**	number of **TN**

Image: https://www.analyticsvidhya.com/blog/2020/10/confusion-matrix-is-no-more-a-confusion/

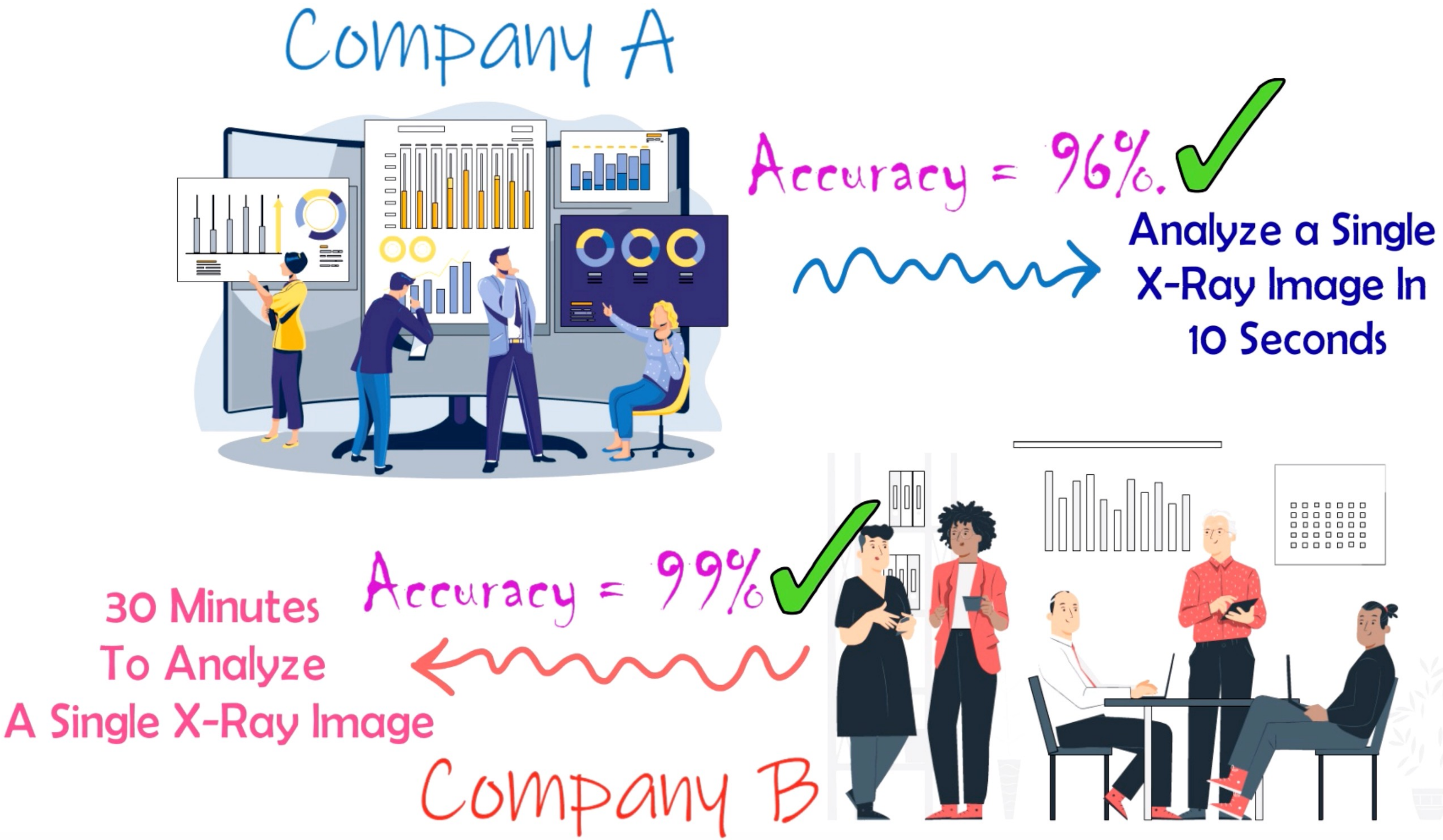
Accuracy

		ACTUAL *If patient have cancer or not*	
		have cancer	doesn't have cancer
PREDICTION *what our model predicted*	have cancer	number of **TP**	number of **FP**
	doesn't have cancer	number of **FN**	number of **TN**

$$ACCURACY = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} = \frac{TP + TN}{TP + TN + FN + FP}$$

Image: https://www.analyticsvidhya.com/blog/2020/10/confusion-matrix-is-no-more-a-confusion/

What would you do?



What Would You Do Now?

What should you do?

– Should not decide yet

Image: https://www.123rf.com/

Fair Comparison

– **Use the same test set**

- Otherwise test would be biased

– **Collect a diverse and big data**

- Test data should be representative of the real life problem

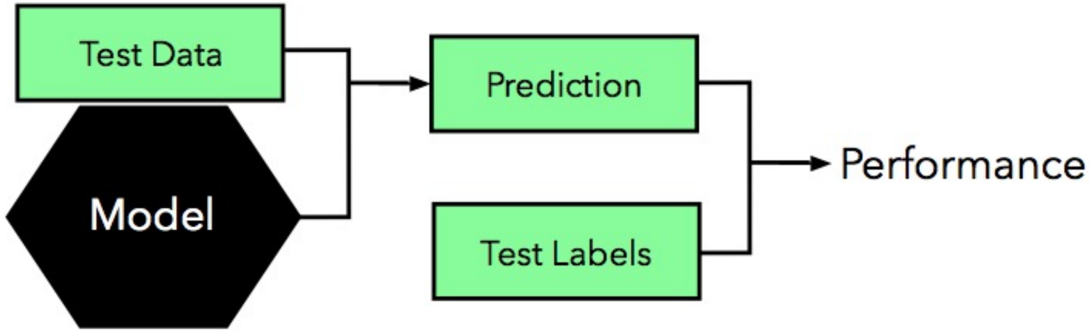
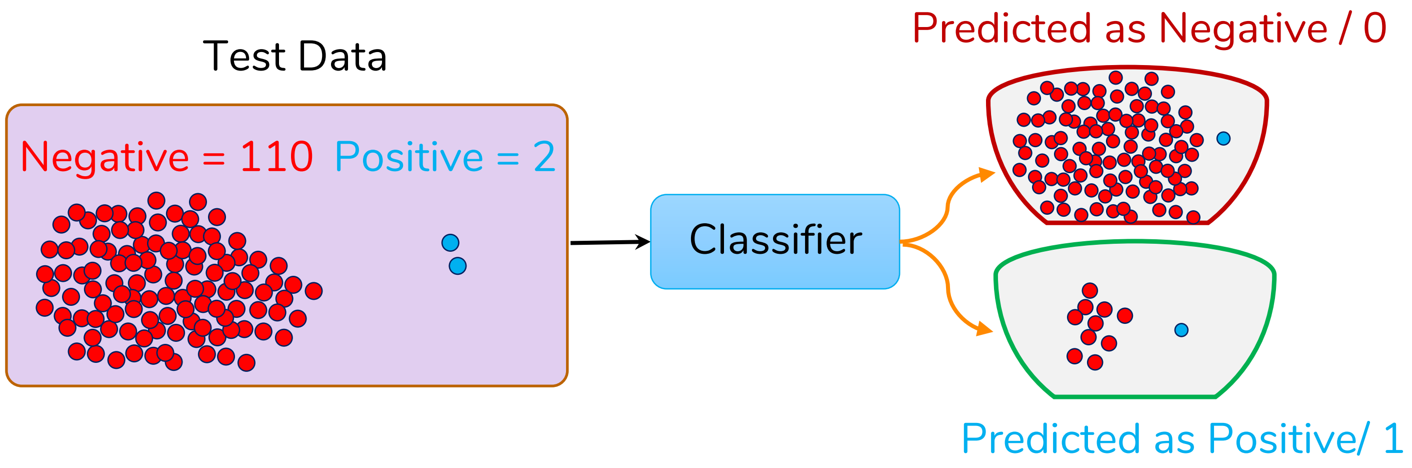


Image: https://sebastianraschka.com/

ACCURACY PARADOX



Negative = No pneumonia
Positive = pneumonia

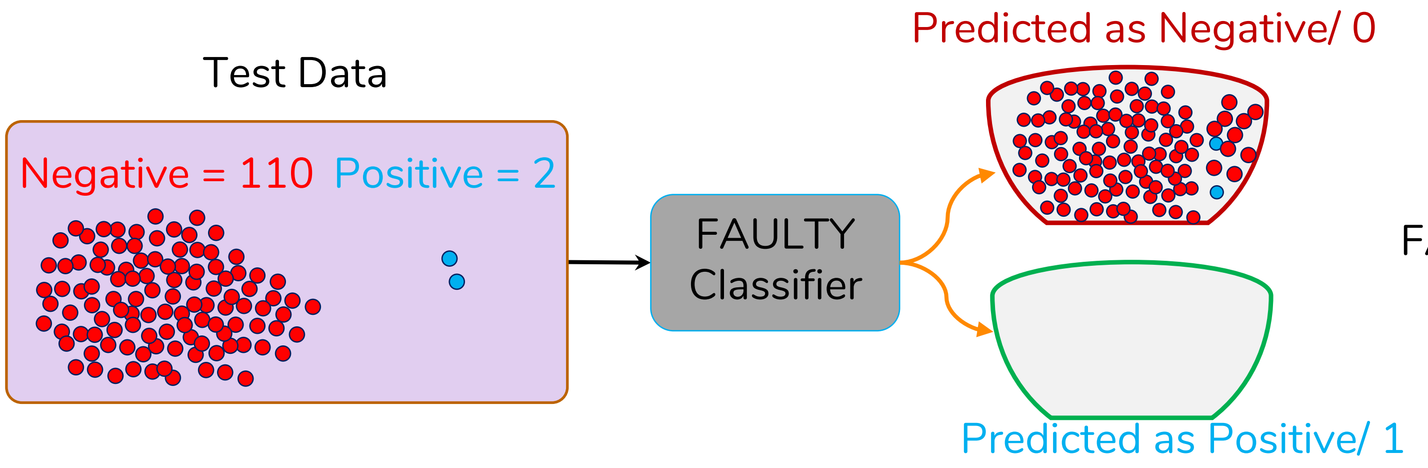
		Actual	
		Positive	Negative
Predicted	Positive	1	10
Predicted	Negative	1	100

Confusion Matrix

False Positive Rate	0.091
Accuracy	0.901

Recall/True Positive Rate (FP/N)	0.5
Precision (TP/(TP+FP))	0.091

ACCURACY PARADOX



FAULTY MODEL

Negative = No pneumonia
Positive = pneumonia

		Actual	
		Positive	Negative
Predicted	Positive	0	0
	Negative	2	110

Accuracy= 110/112=0.98

Confusion Matrix

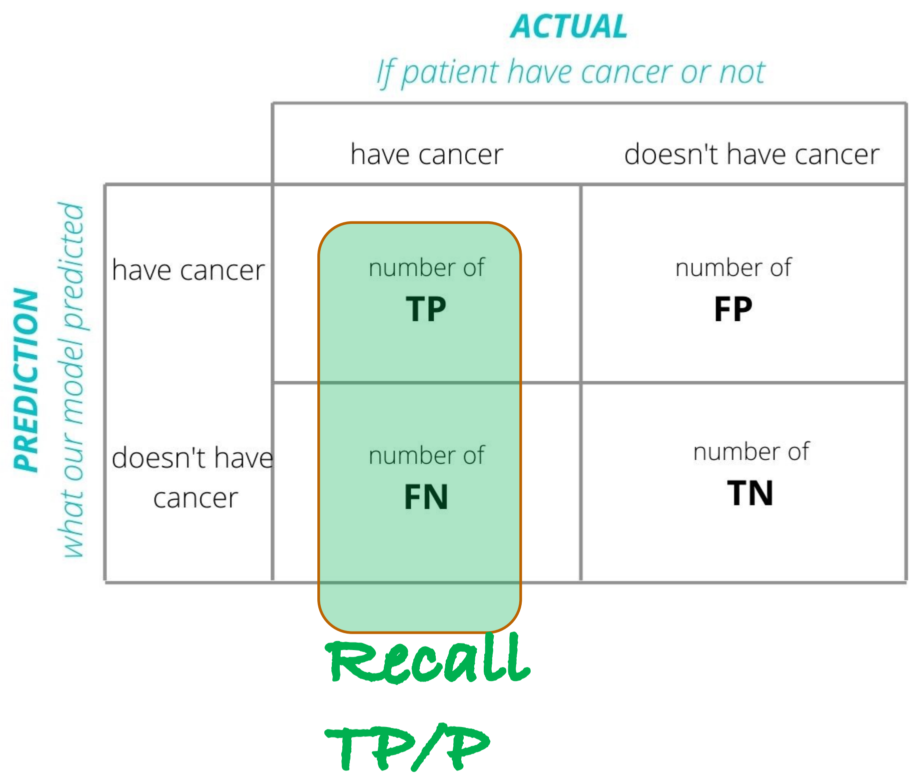
More Metrics Derived from Confusion Matrix

- Sensitivity (Recall or True positive rate)
- Specificity (True negative rate)
- False positive rate (FPR)
- Precision

- Recall – Specificity → Balanced Accuracy
- Recall – FPR → ROC AUC
- Precision –Recall → PR AUC

Recall & True Negative Rate

- Sensitivity (Recall or True positive rate)
- Specificity (True negative rate)



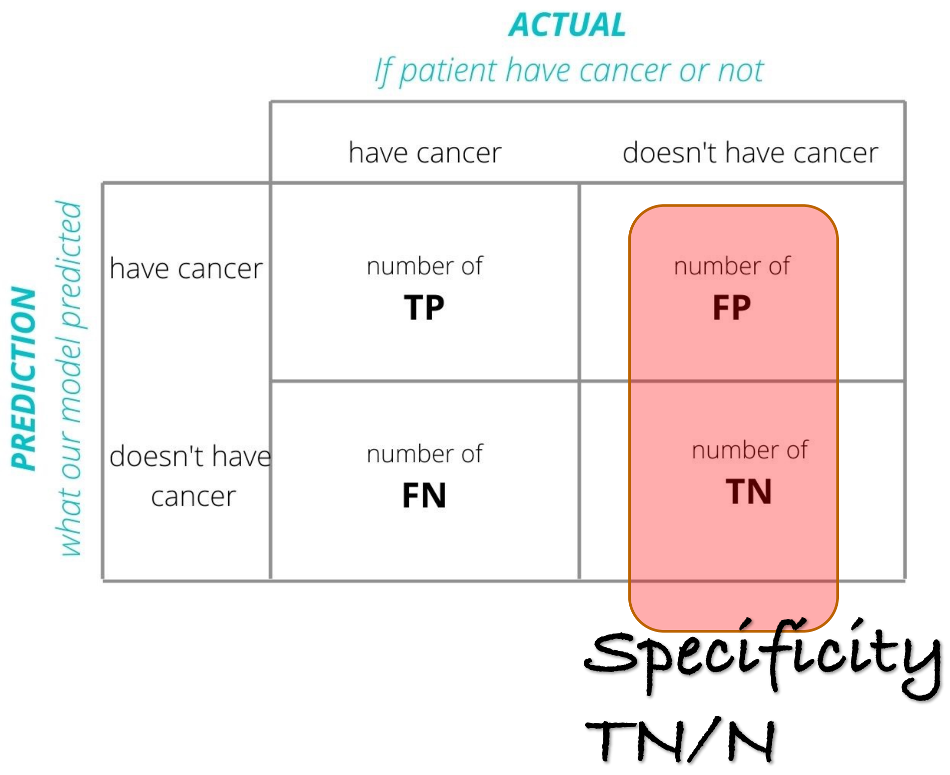
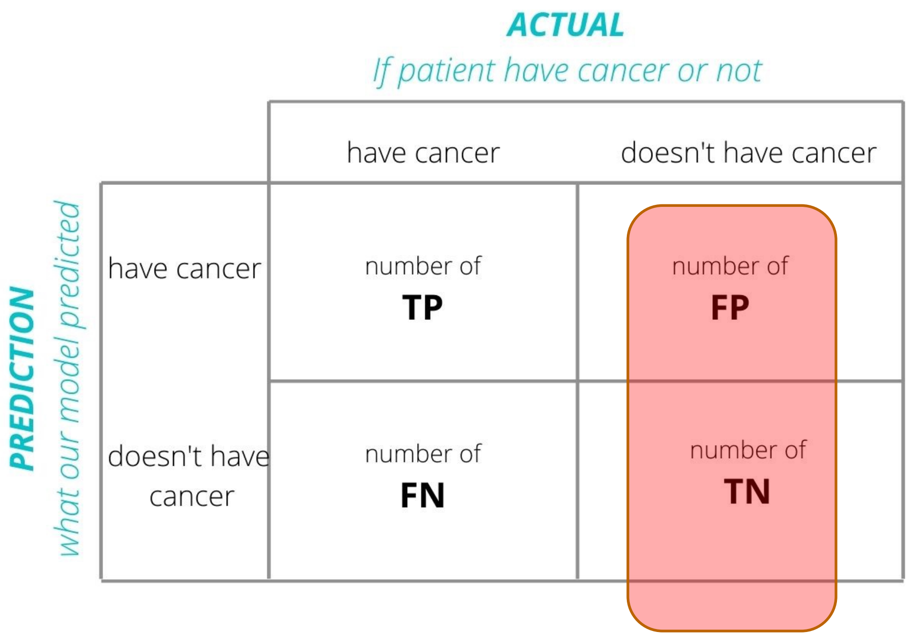
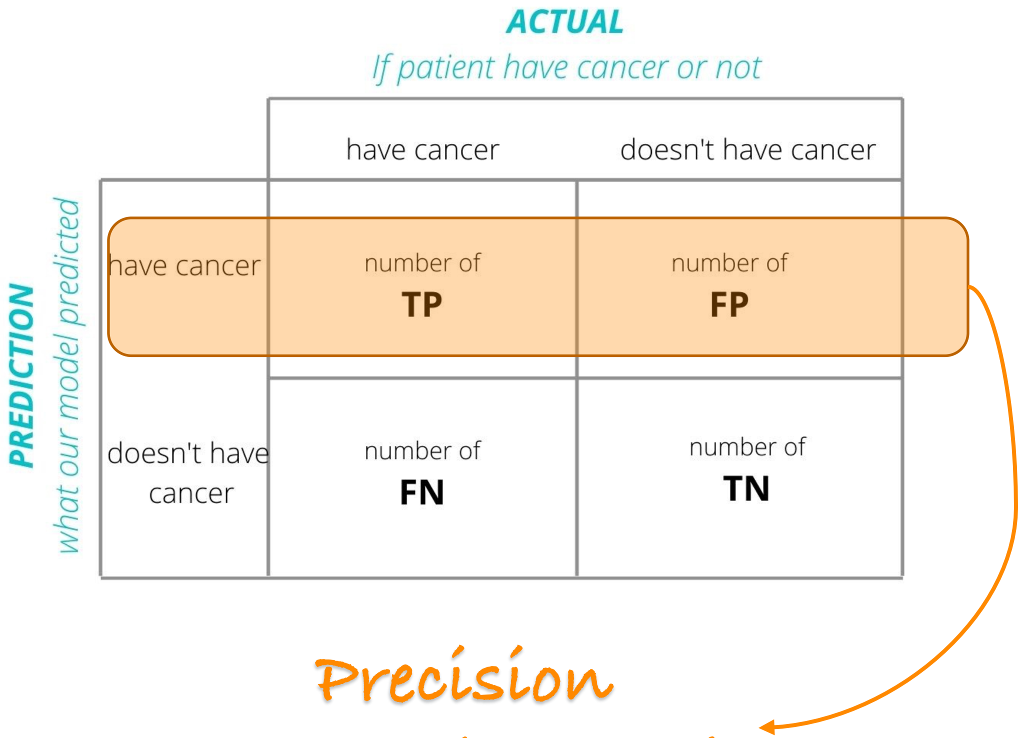


Image: https://www.analyticsvidhya.com/blog/2020/10/confusion-matrix-is-no-more-a-confusion/

False Positive Rate & Precision



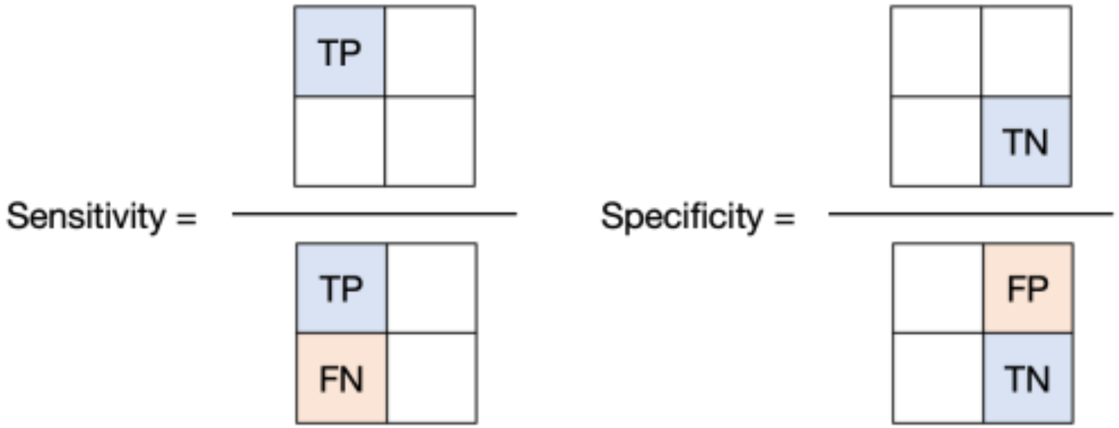
False Positive Rate
FP/N



Precision
TP/(TP+FP)

Image: Wikipedia
Image: https://www.analyticsvidhya.com/blog/2020/10/confusion-matrix-is-no-more-a-confusion/

Balanced accuracy



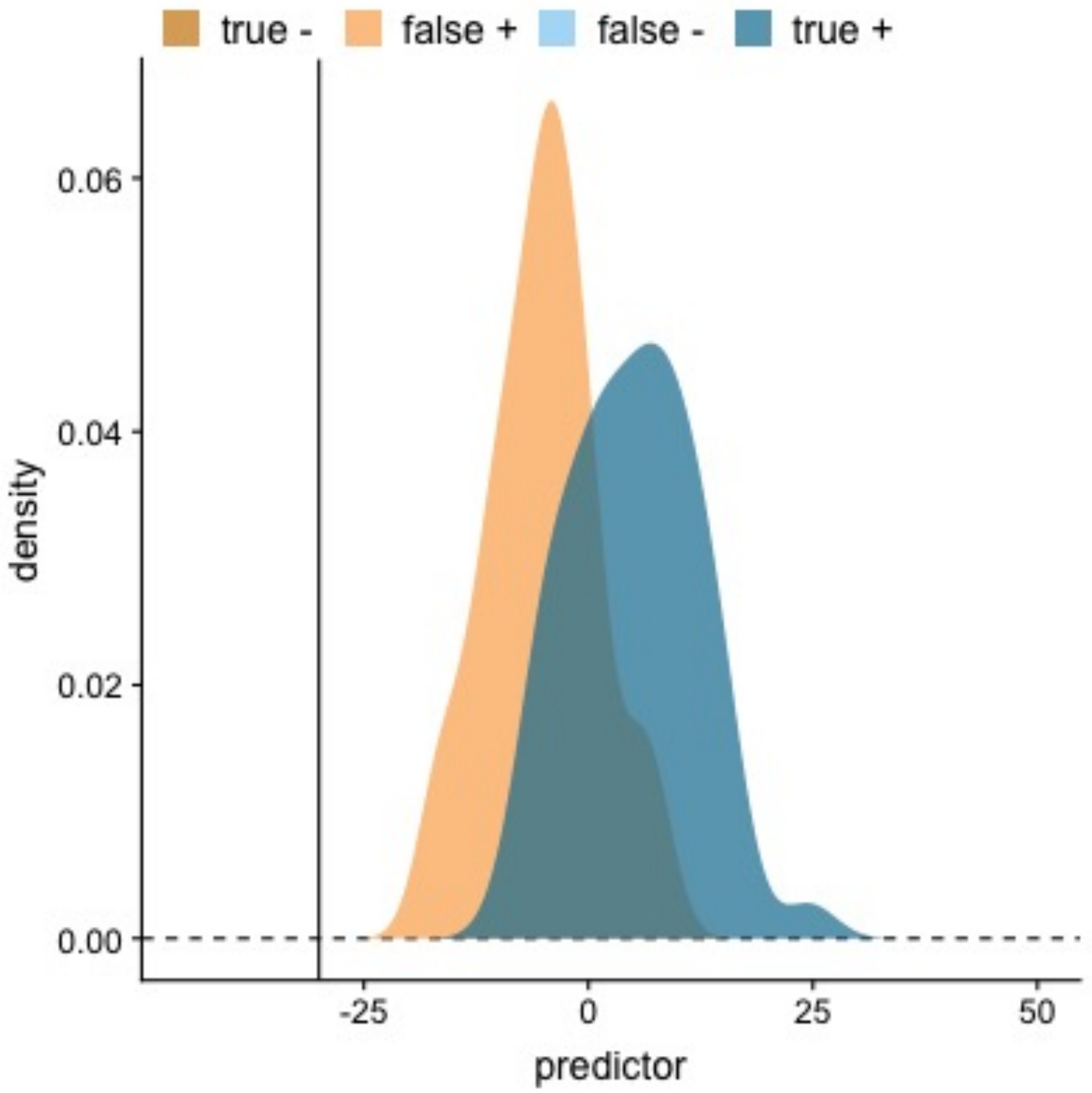
$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2}$$

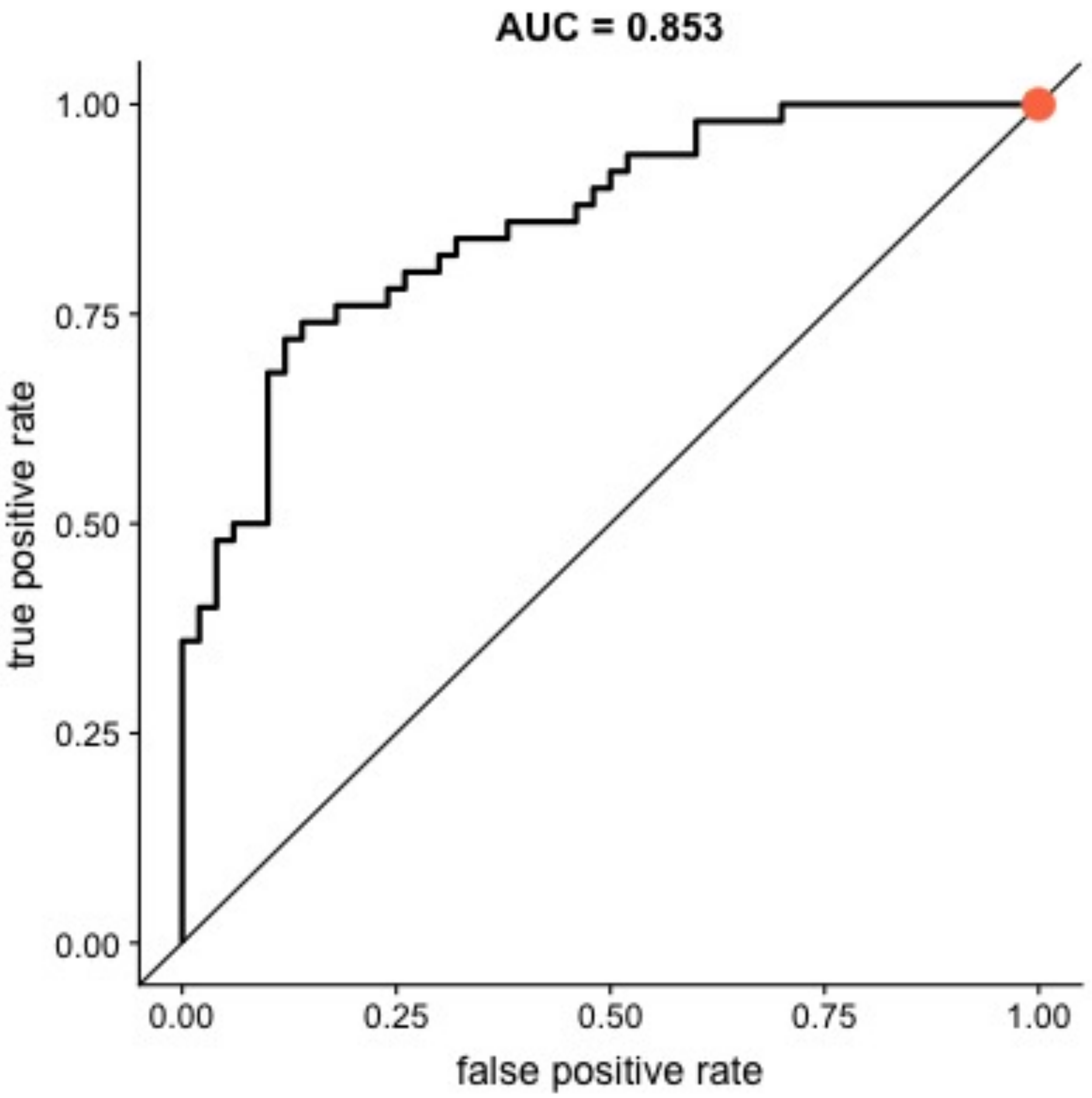
		Actual	
		Positive	Negative
Predicted	Positive	1	10
	Negative	1	100

Recall/True Positive Rate	0.5
False Positive Rate	0.091
Precision	0.091
Accuracy	0.901
Balanced Accuracy	**0.45**

https://statisticaloddsandends.wordpress.com/2020/01/23/what-is-balanced-accuracy/

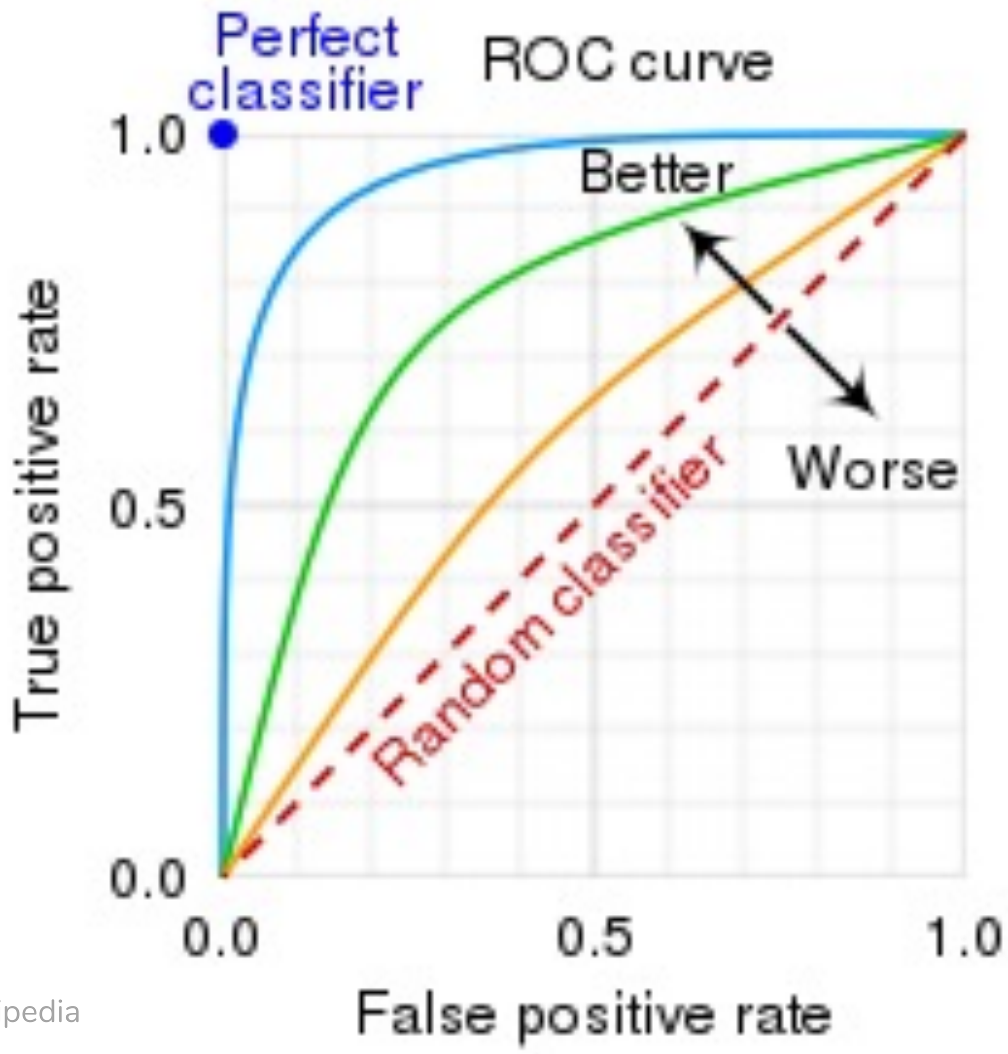
ROC Curve –ROC-AUC





https://github.com/dariyasydykova/open_projects

ROC Curves and ROC AUC



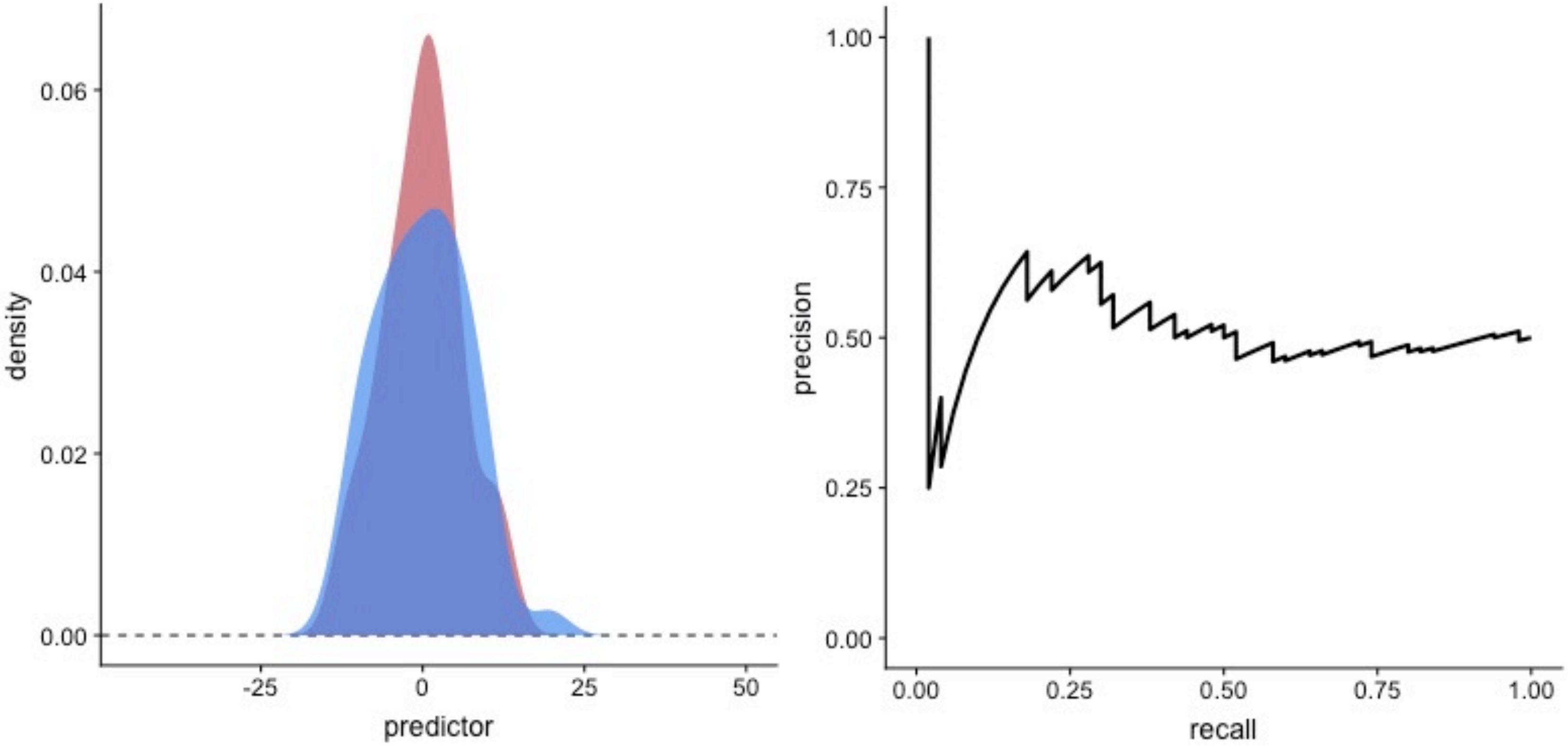
wikipedia

- A receiver operating characteristic curve, or **ROC curve**: Recall(TPR) vs FPR
- The **ROC-AUC** : Area under the ROC curve → summarizes classifier performance
- **ROC-AUC=0.5 → random classifier**
- **ROC-AUC= 1.0 → perfect classifier**
- More informative than accuracy for imbalanced data
- **Excessively optimistic for highly imbalanced set**

(# of negative samples)>>(# of positive samples)

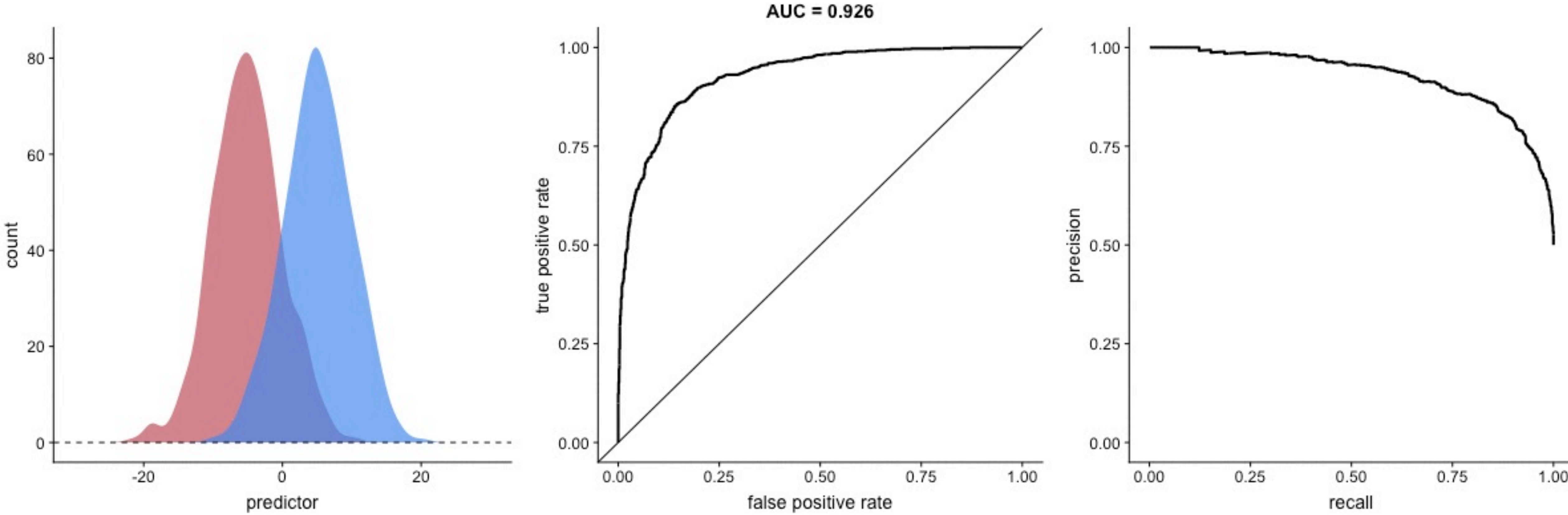
Precision-Recall(PR) Curve and Area Under PR Curve

- For soft-classifiers
- Precision vs Recall
- Area Under the Precision-Recall Curve: summarizes the PR Curve (AP: Average Precison, AUCPR, AUPRC)



ROC vs PR Curve

- When data is imbalanced, the ROC-AUC might not reflect the true performance of the classifier
- PR AUC would be the metric to use if the focus of the model is to identify correctly as many positive samples as possible.



https://github.com/dariyasydykova/open_projects

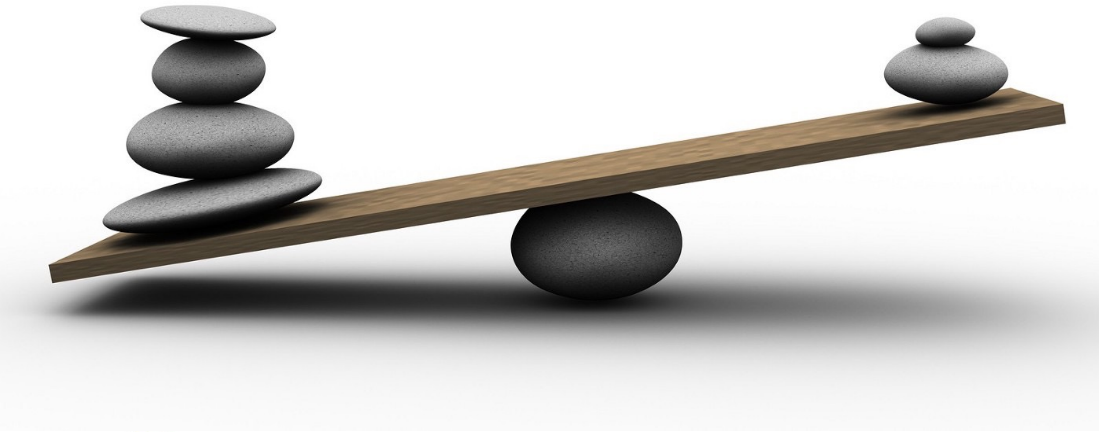
Summary

– Fair comparison

- Use same test data

– Metrics

- Accuracy (might be misleading)
- Balanced Accuracy
- ROC-AUC (if both classes are equally important)
- PR-AUC (if focusing to identify positive samples)



Evaluation of Binary Classifiers

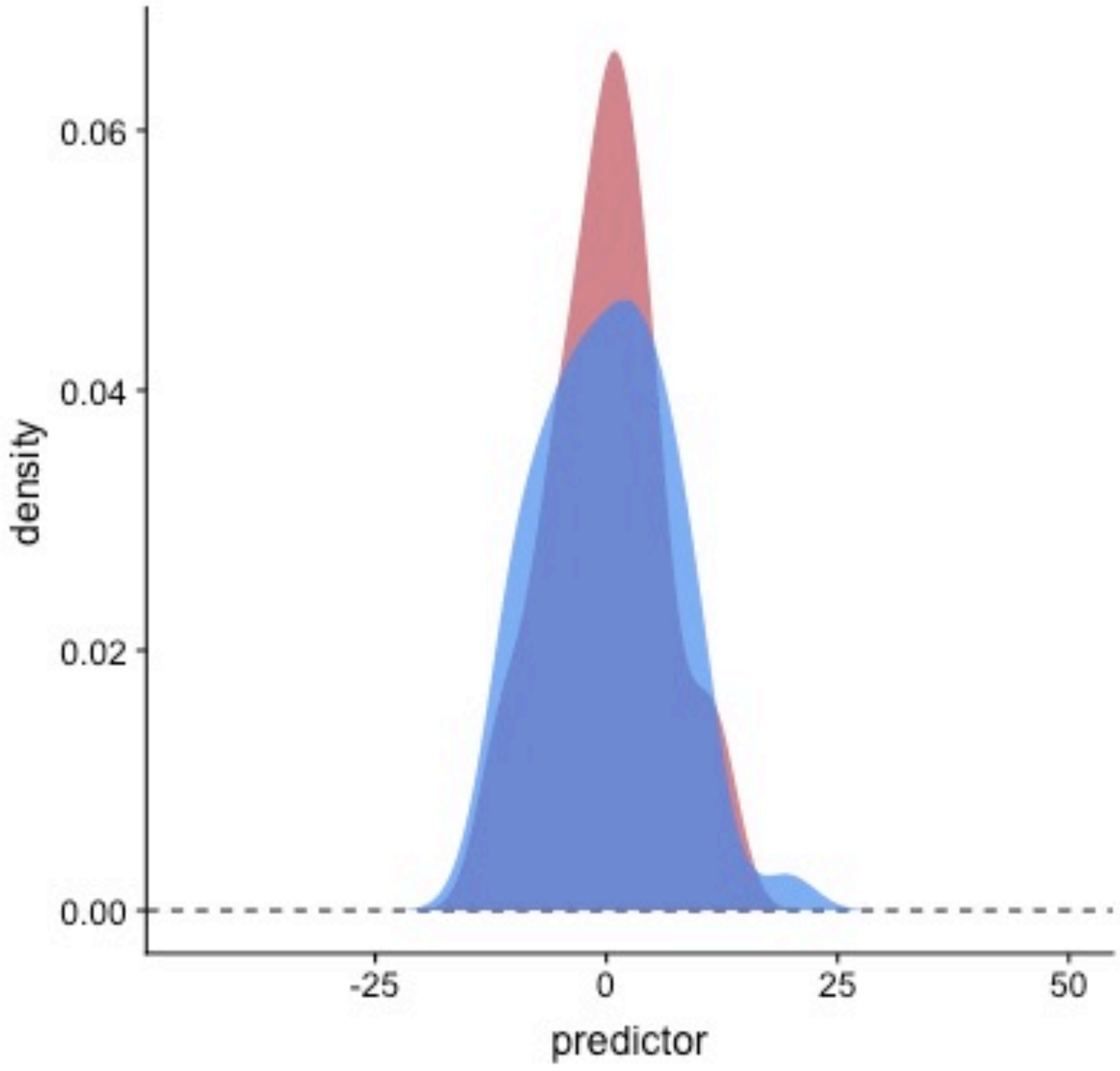
Neslihan Bayramoglu

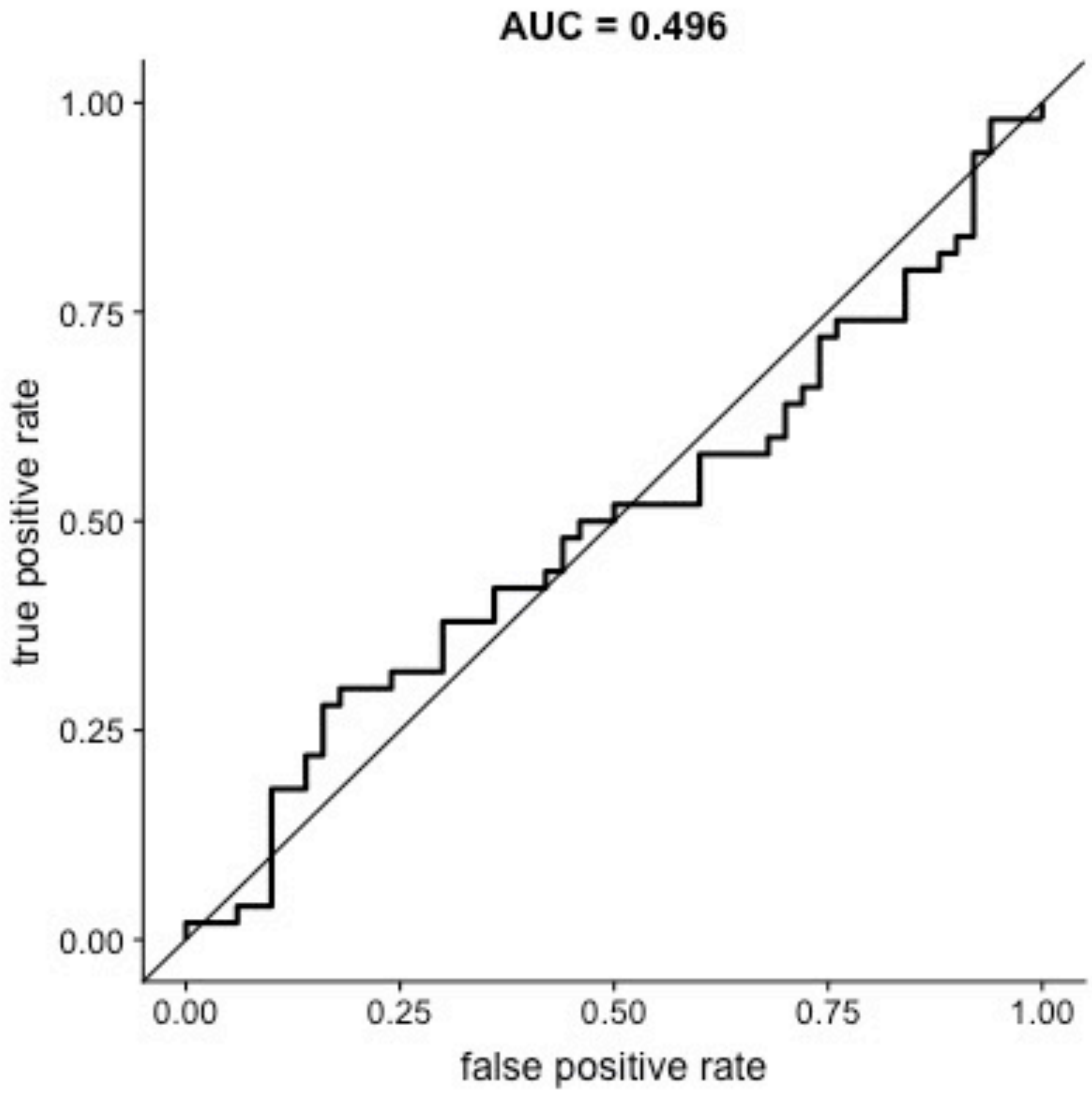
09/05/2022

neslihan.bayramoglu@oulu.fi

Image: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb

ROC Curve –ROC-AUC





https://github.com/dariyasydykova/open_projects

Binary Classifiers

- classifying the data into two groups
- **a large number of medical studies are based on classification models**