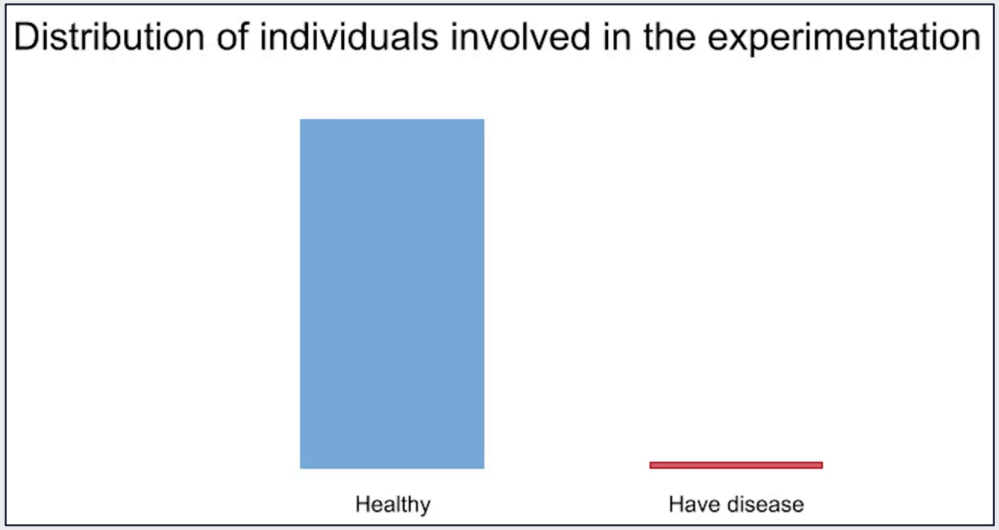
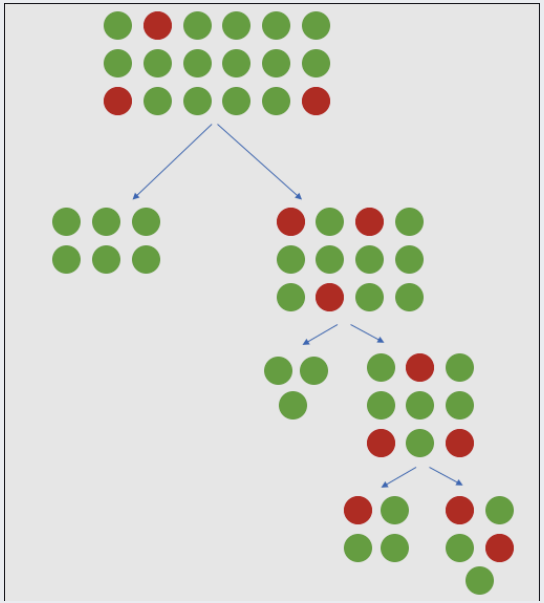
Learning from Imbalanced Datasets

Imbalanced classes are a common problem in machine learning classification, where there's a disproportionate ratio of observations in each class.



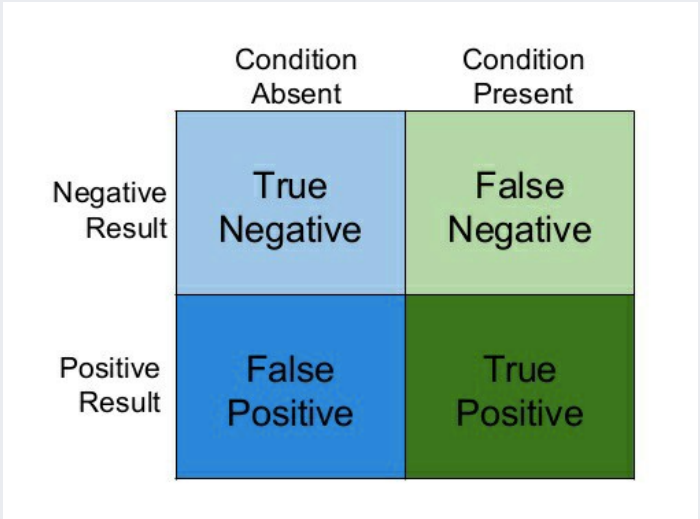
- *The Accuracy Paradox is the paradoxical finding that accuracy is not a good metric for predictive models when classifying in predictive analytics.*
- *This is because a simple model may have a high level of accuracy but be too crude to be useful.*

Tree based Model



The top of the decision tree is likely to learn splits which separate out the majority class into pure groups at the expense of learning rules which separate the minority class.

Types of misclassifications



False positives and false negatives

In the case of the accuracy paradox, we can see that all of our mistakes were predicting positive (meaning that there is a fire) as not fire, since all the predictions we made were only not fire.

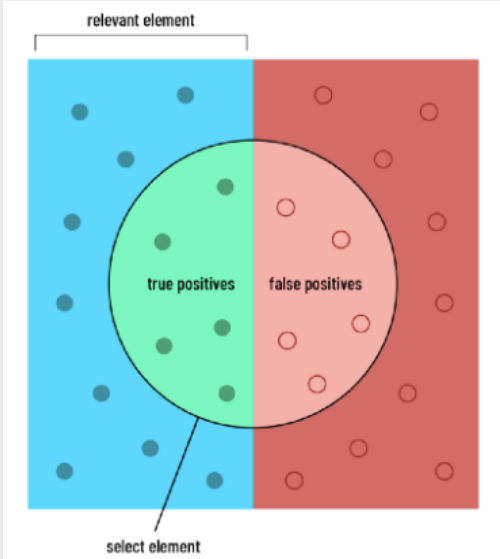
Metrics

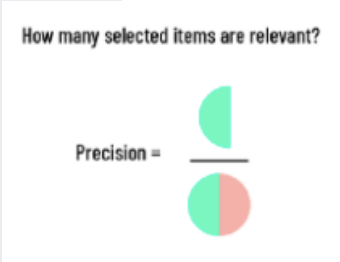
- **Precision**: What proportion of positive identifications was actually correct?

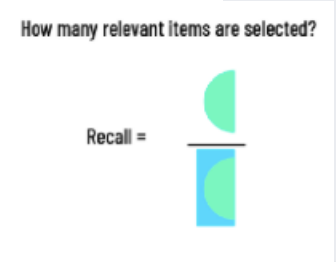
$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Recall:** What proportion of actual positives was identified correctly?

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$







F1 score

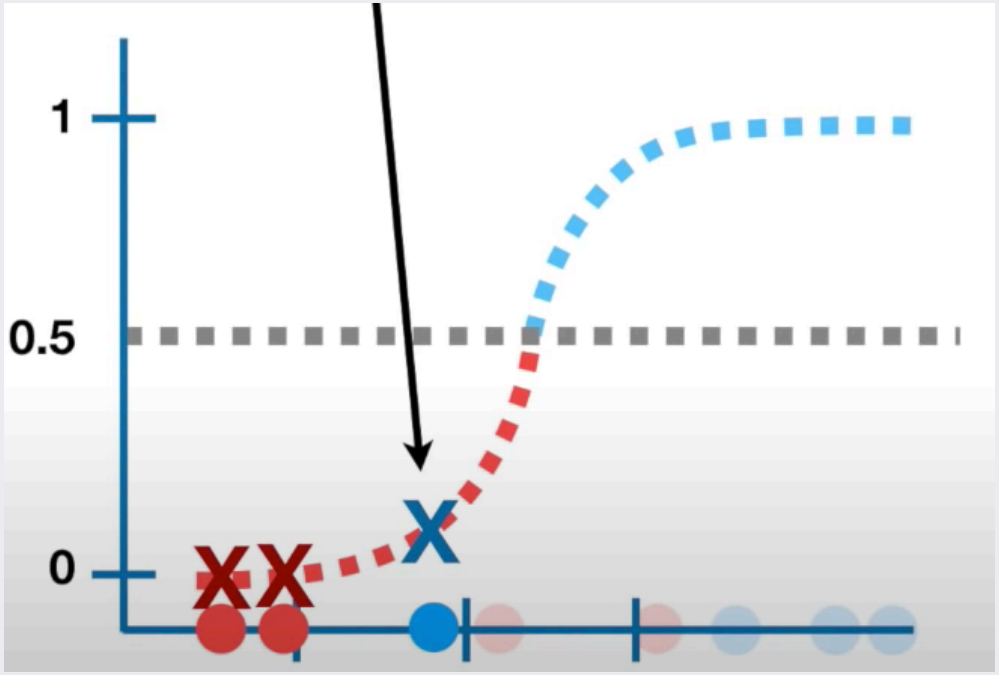
The F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

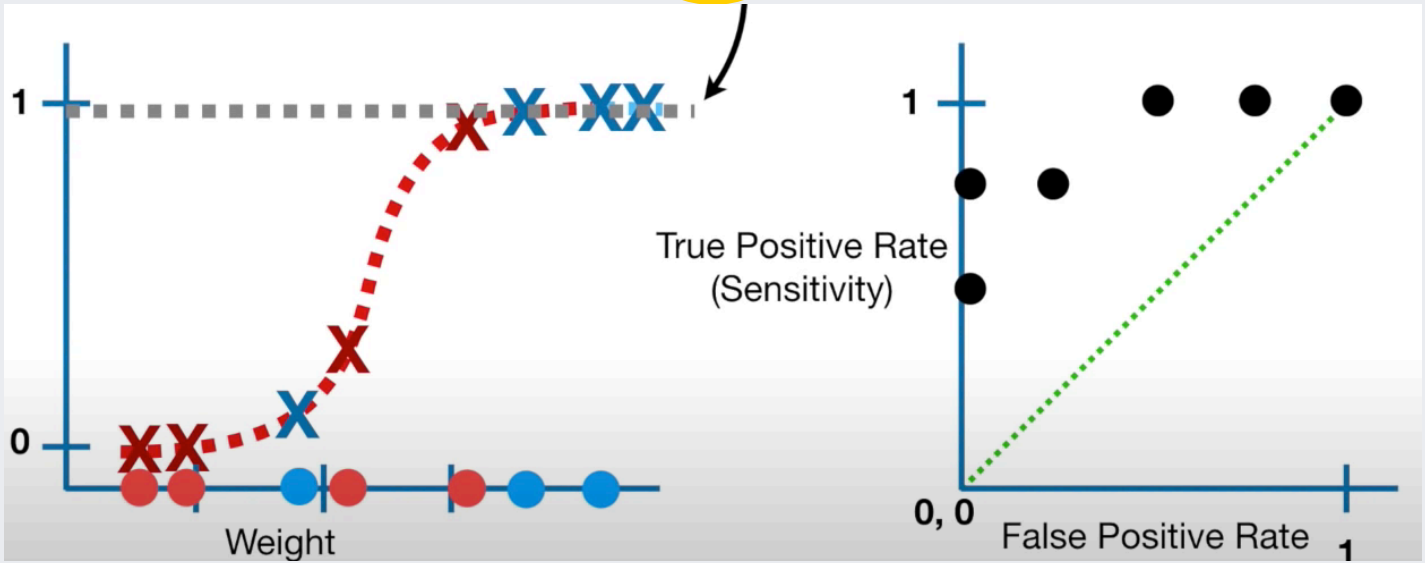
$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

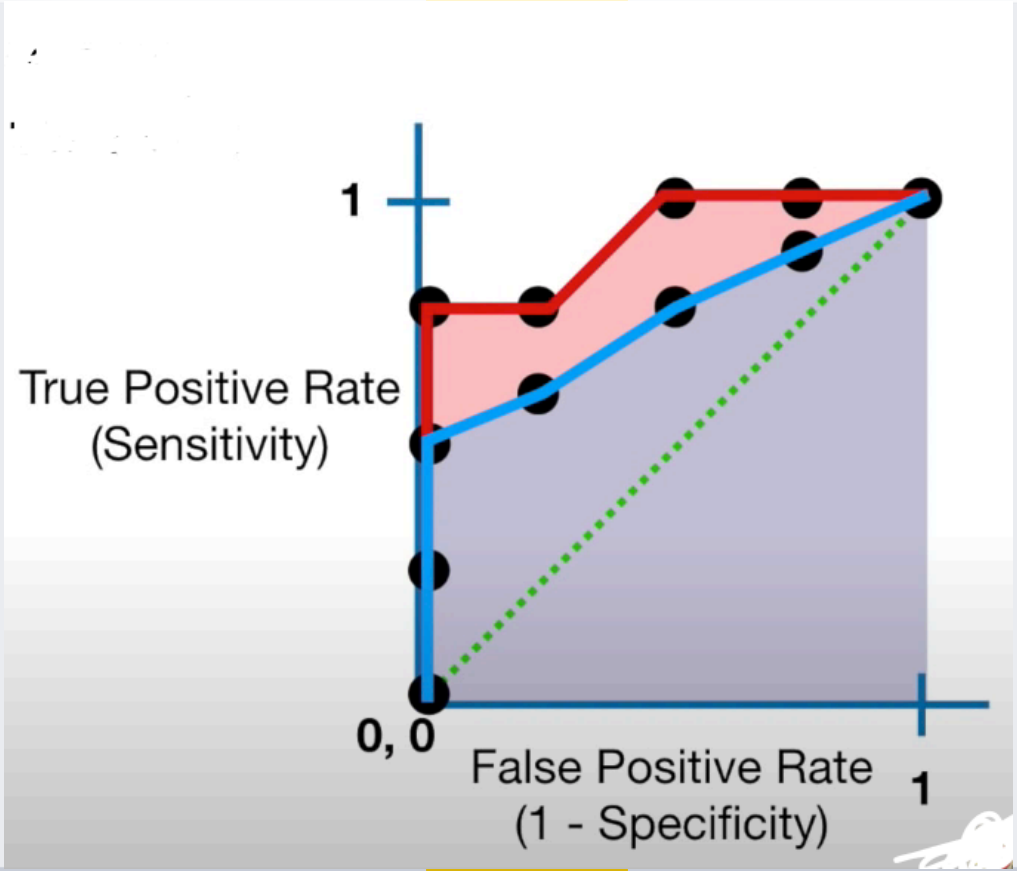
ROC curve

visualizes an algorithm's ability to discriminate the positive class from the rest of the data.

		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	3	1
	Is Not Obese	1	3




1
0
Weight
True Positive Rate
(Sensitivity)
1
0, 0
False Positive Rate
1

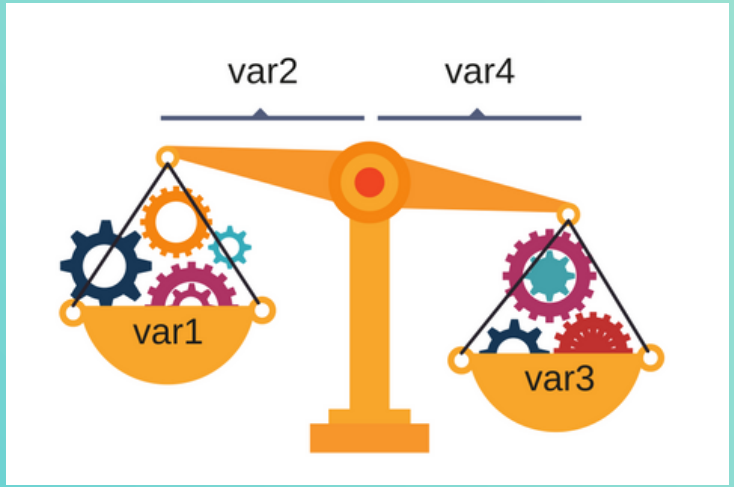

1
True Positive Rate
(Sensitivity)
0, 0
False Positive Rate
(1 - Specificity)
1

Ways to address the class imbalance

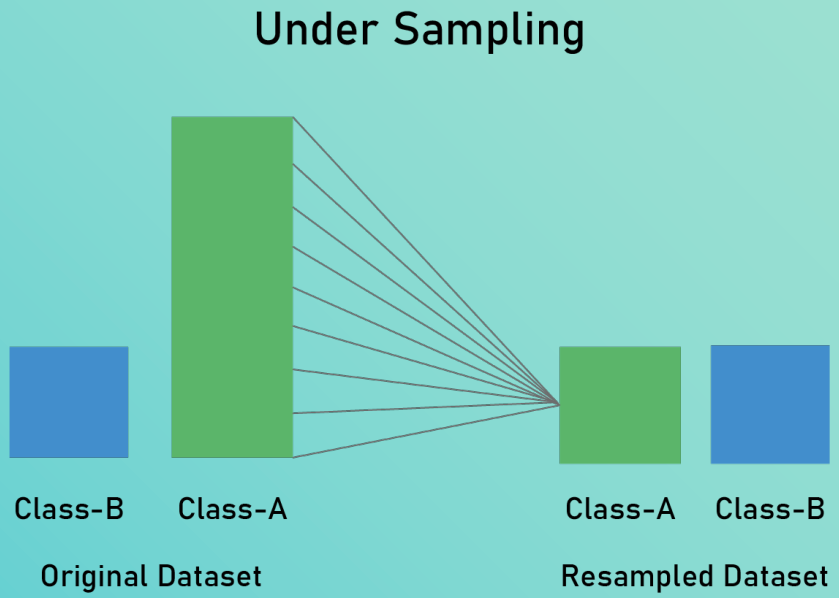
Class weight, Under & Over Sampling

Class weight

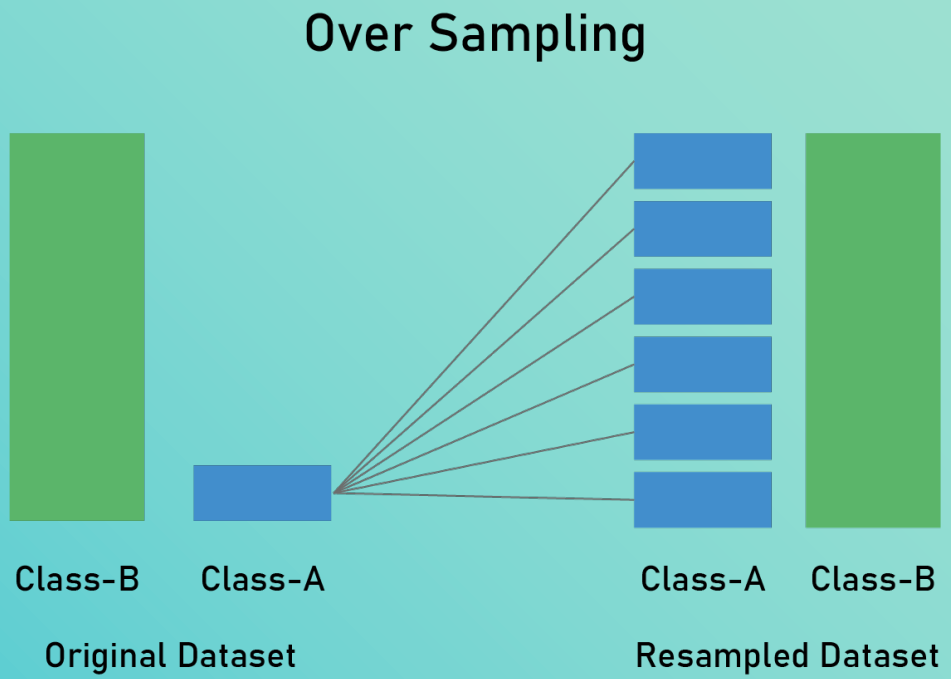
One of the simplest ways to address the class imbalance is to simply provide a weight for each class which places more emphasis on the minority classes such that the end result is a classifier which can learn equally from all classes.



Under sampling majority class



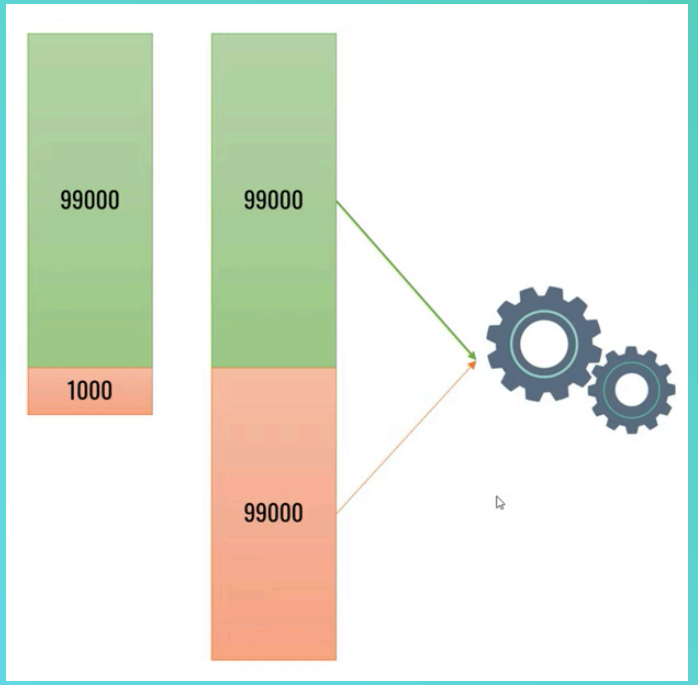
Over sampling majority class



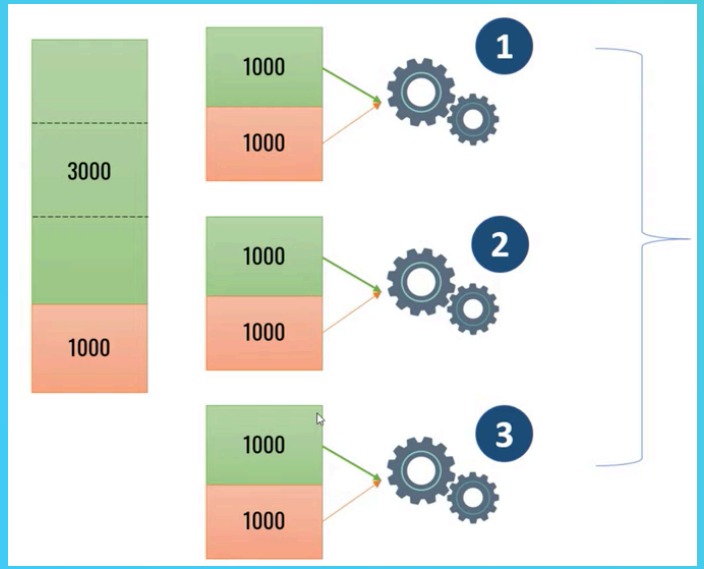
Over sampling using SMOTE

Synthetic minority over sampling technique

Generates synthetic examples using K nearest neighbors algo.



Ensemble Voting

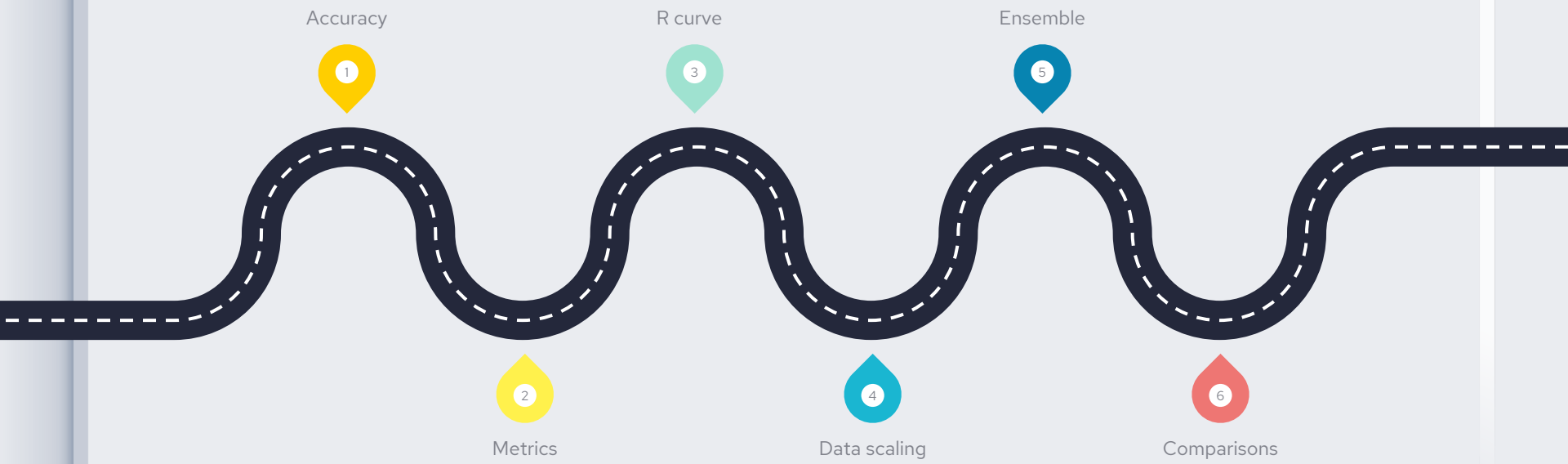


Majority Voting

ANN model – 100 epoch

	Without sampling	Uunder sampling	Over sampling	SMOTE	Ensemble
Precision	**0.63**	**0.73**	**0.75**	**0.81**	**0.46**
Recall	**0.45**	**0.69**	**0.84**	**0.81**	**0.84**
F1 score	**0.53**	**0.71**	**0.79**	**0.81**	**0.60**

Summary


Accuracy
1
R curve
3
Ensemble
5
2
Metrics
4
Data scaling
6
Comparisons

Thank You

Questions ?