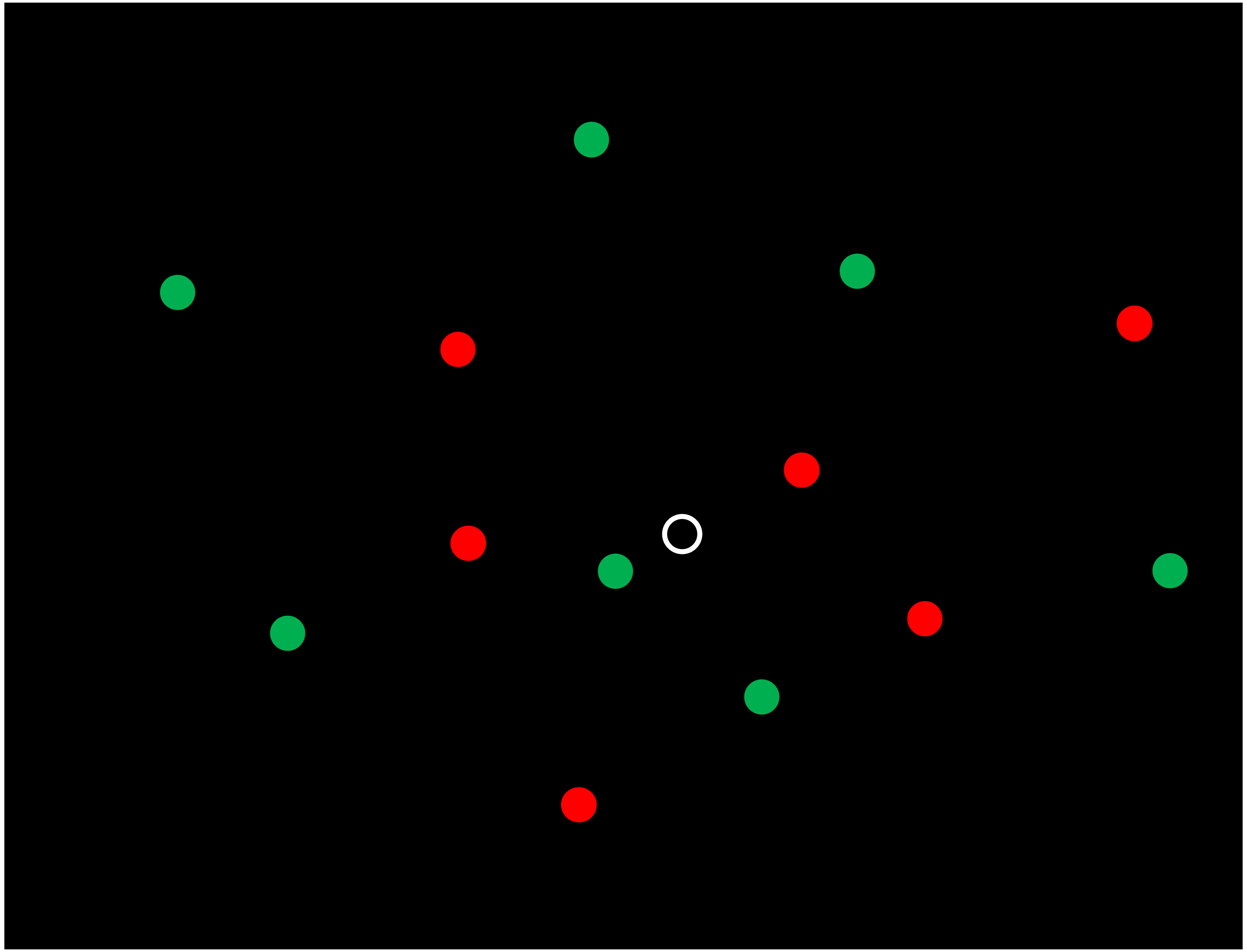
KNN CLASSIFICATION

Ratchainant Thammasudjarit, Ph.D.

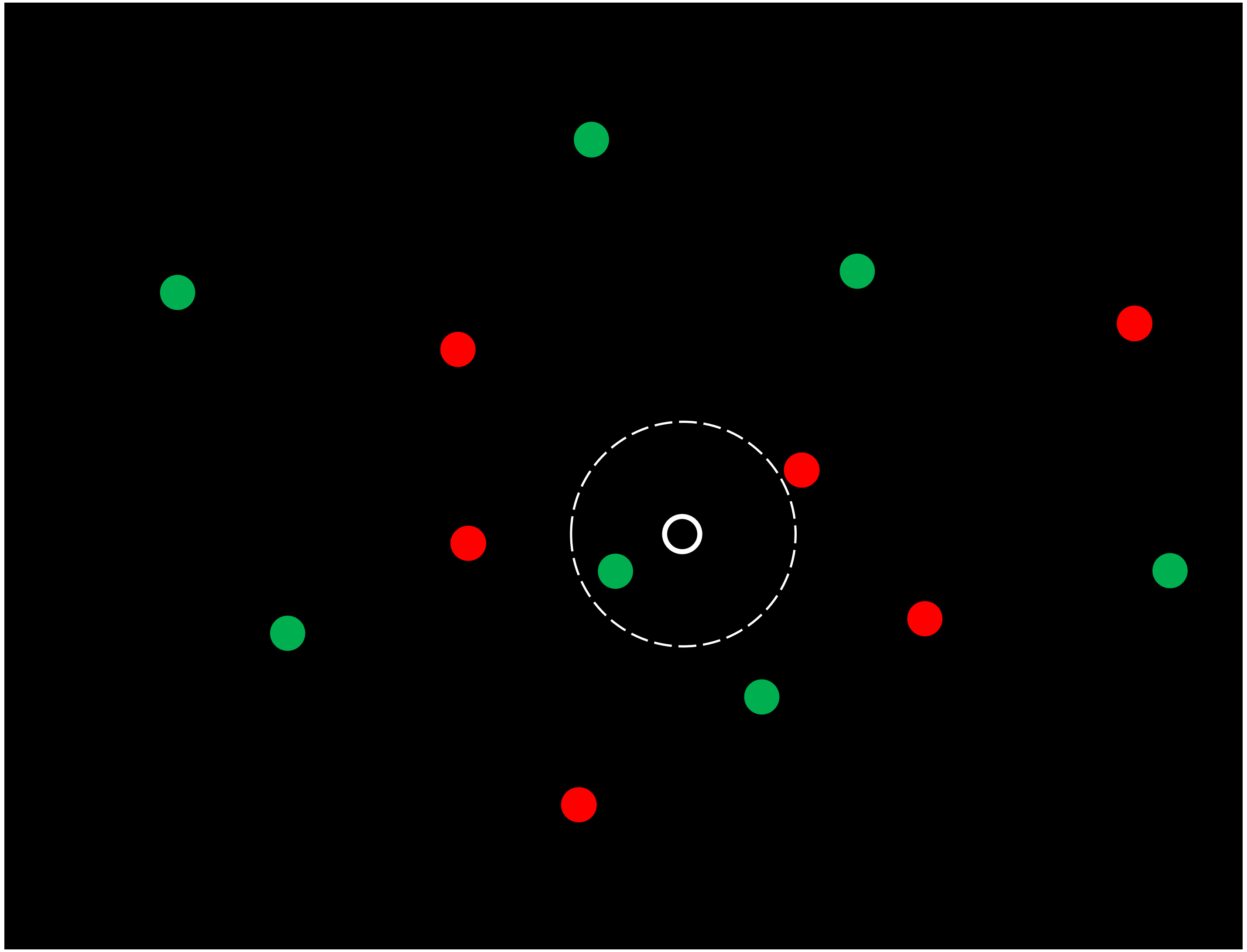
- Peer Pressure



Concepts

Which color should be for the white data point

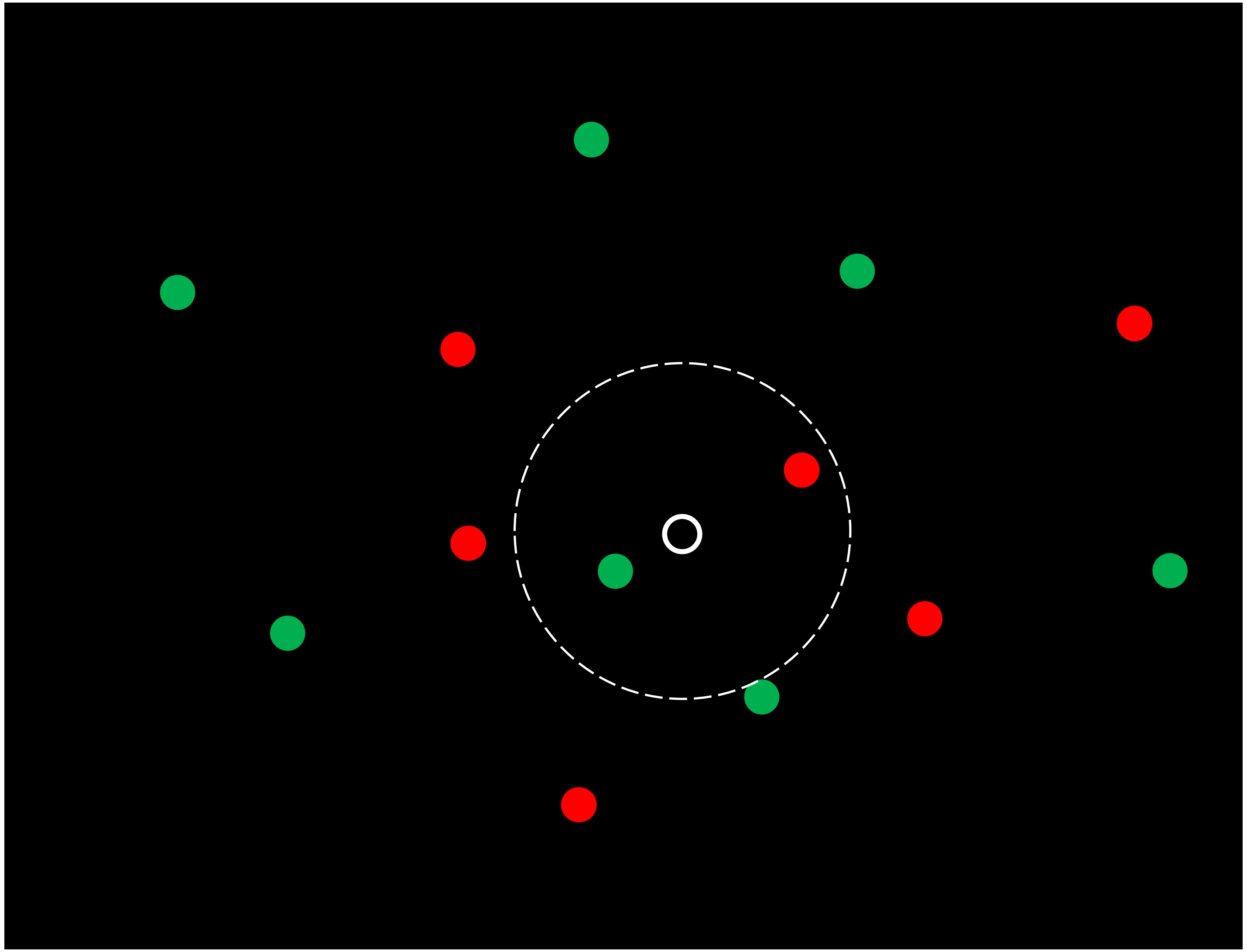
- 1 nearest data point
  - Green is majority: Prediction is green



Concepts

Which color should be for the white data point

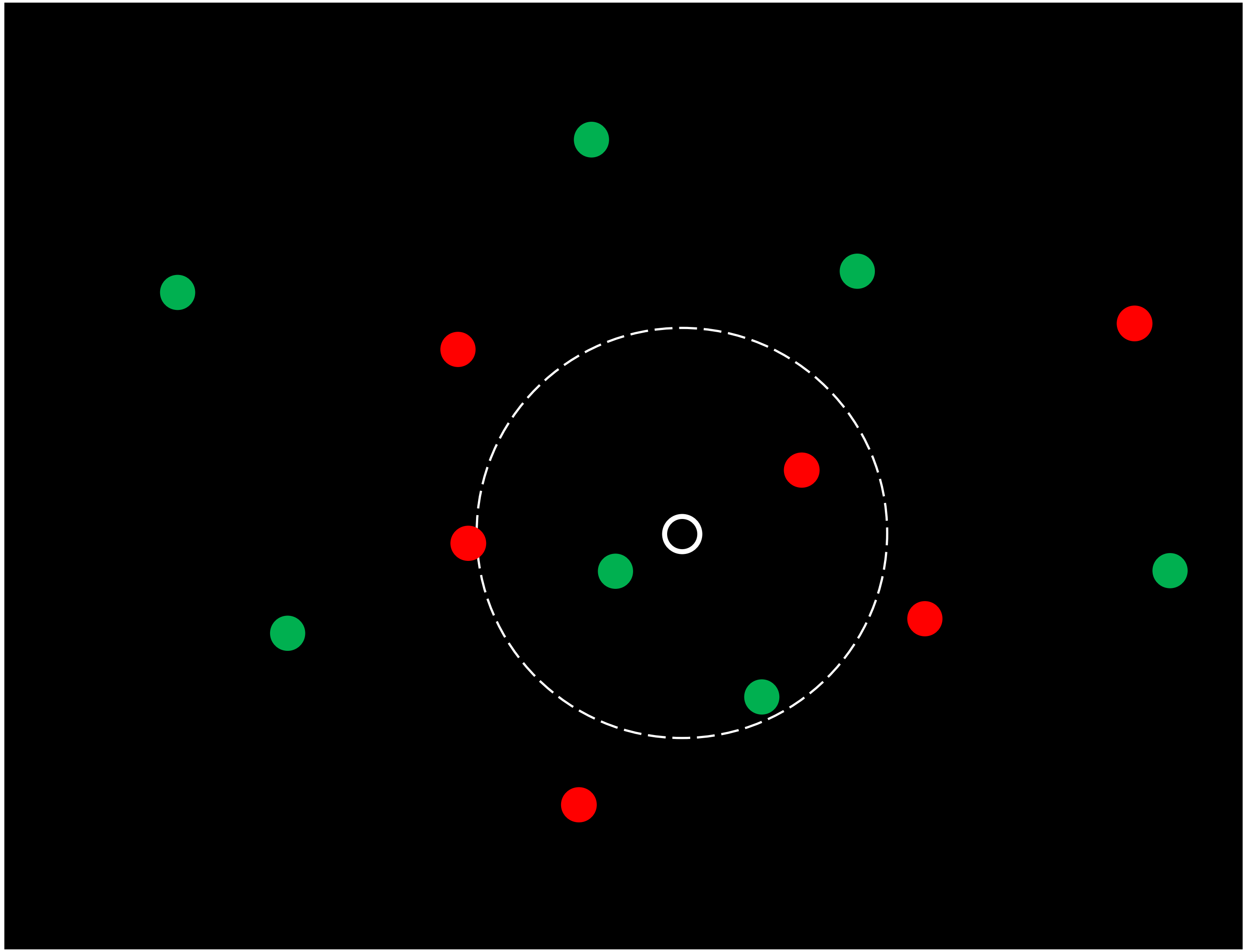
- 2 nearest data points
    - Cannot make decision



Concepts

Which color should be for the white data point

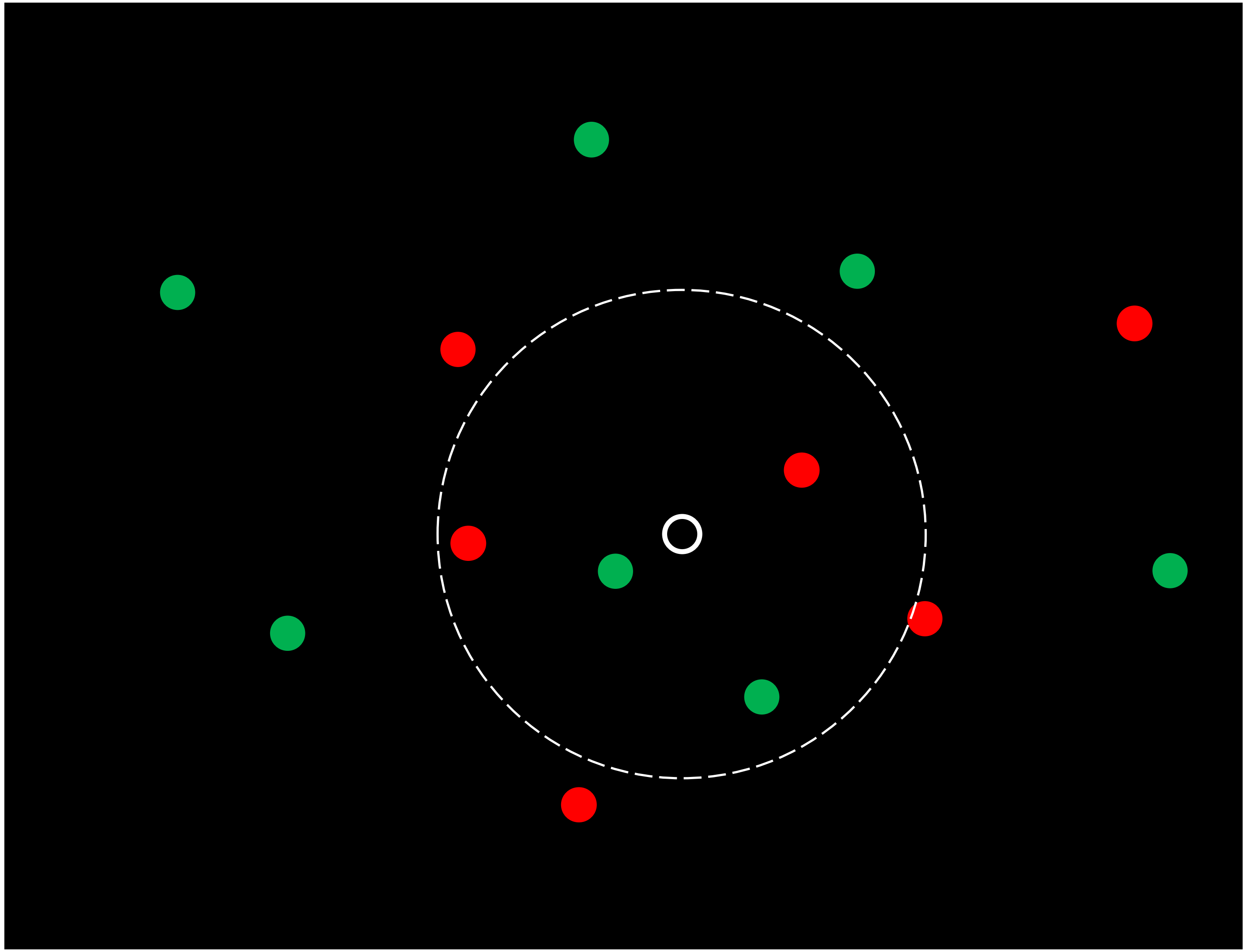
- 3 nearest data points
  - Green is majority: Prediction is green



Concepts

Which color should be for the white data point

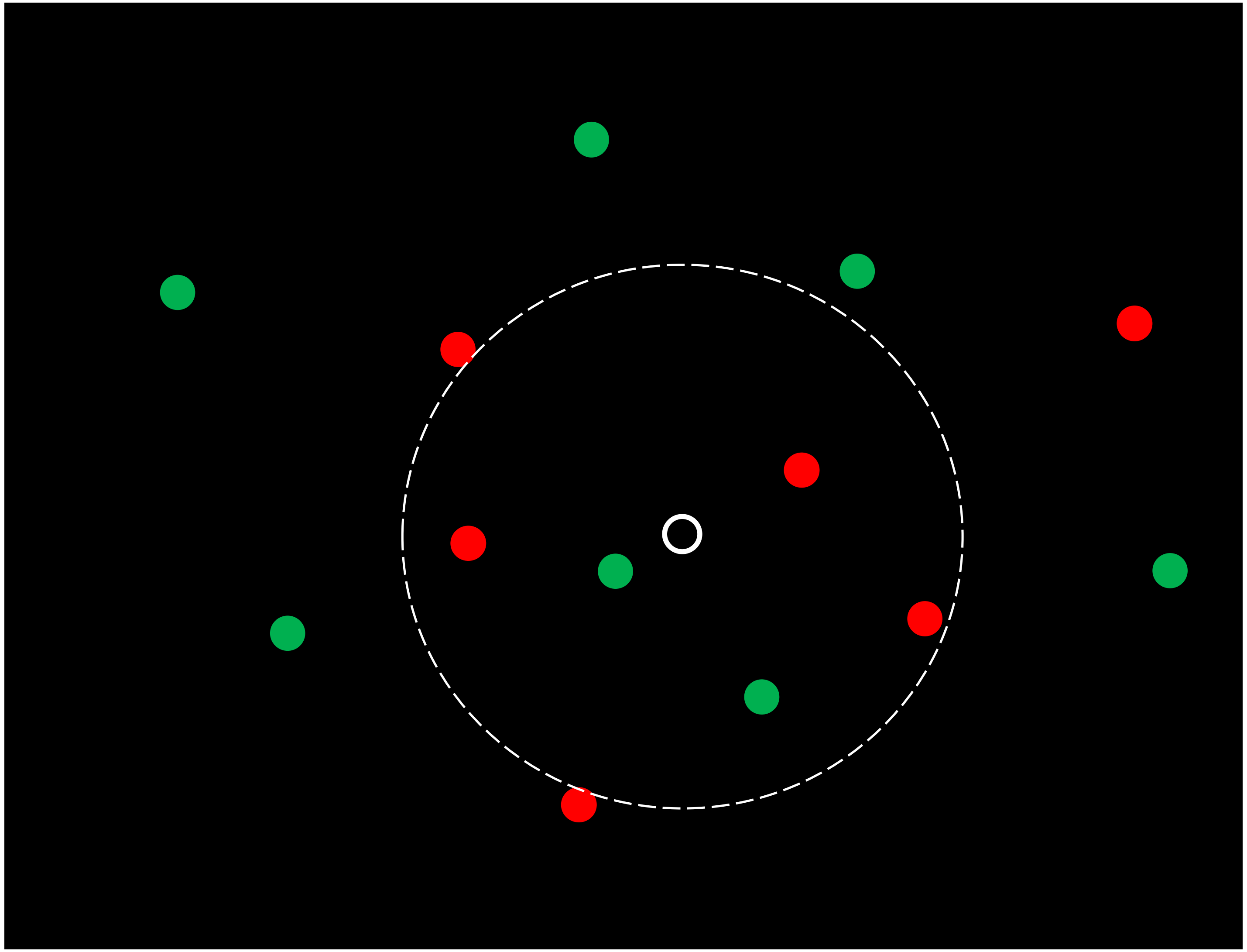
- 4 nearest data points
  - Cannot make decision



Concepts

Which color should be for the white data point

- 5 nearest data points
    - Red is majority: Prediction is red



Concepts

Which color should be for the white data point

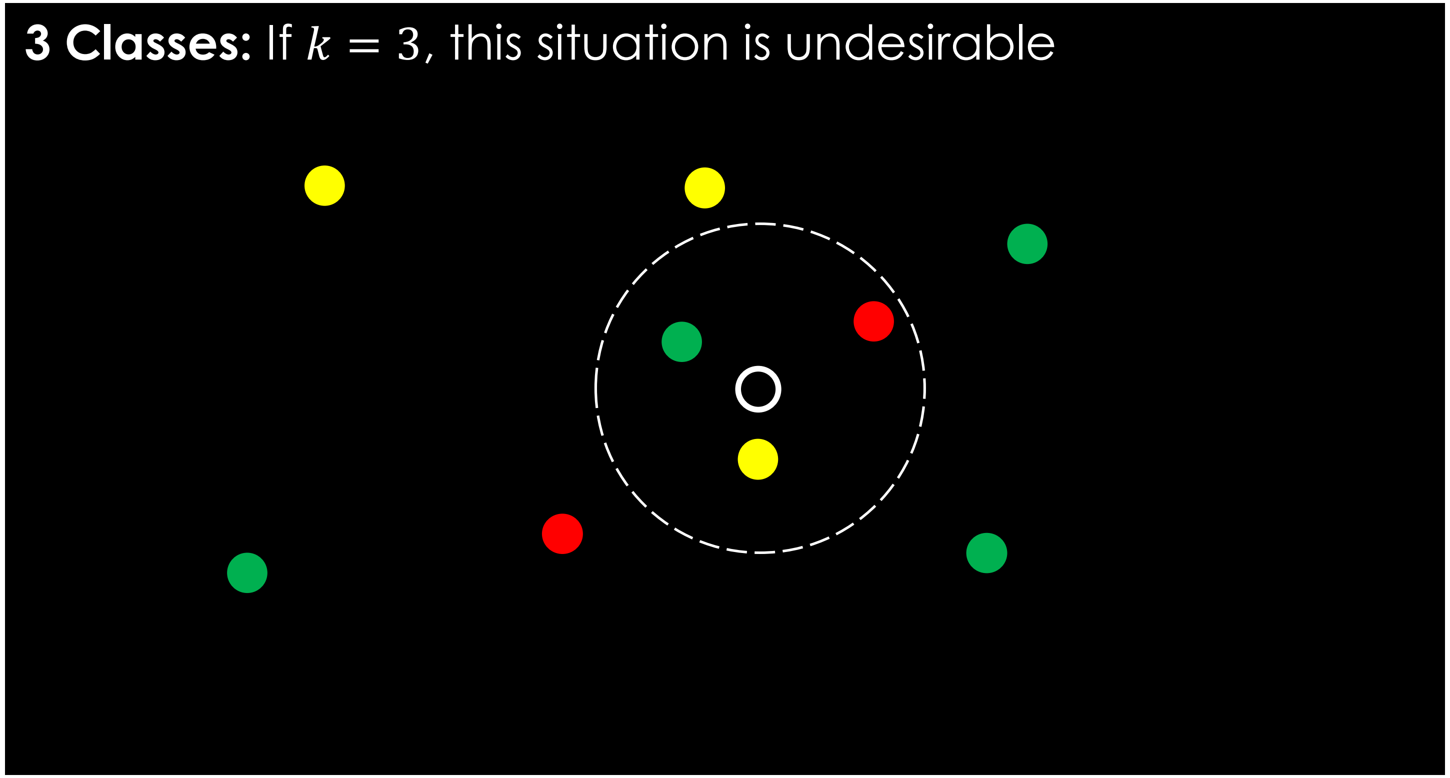
Concepts

The number of nearest neighbor k plays important role for prediction

In common practice, the k is set to be some odd number

- For binary classification
  - k is recommended to be the odd number

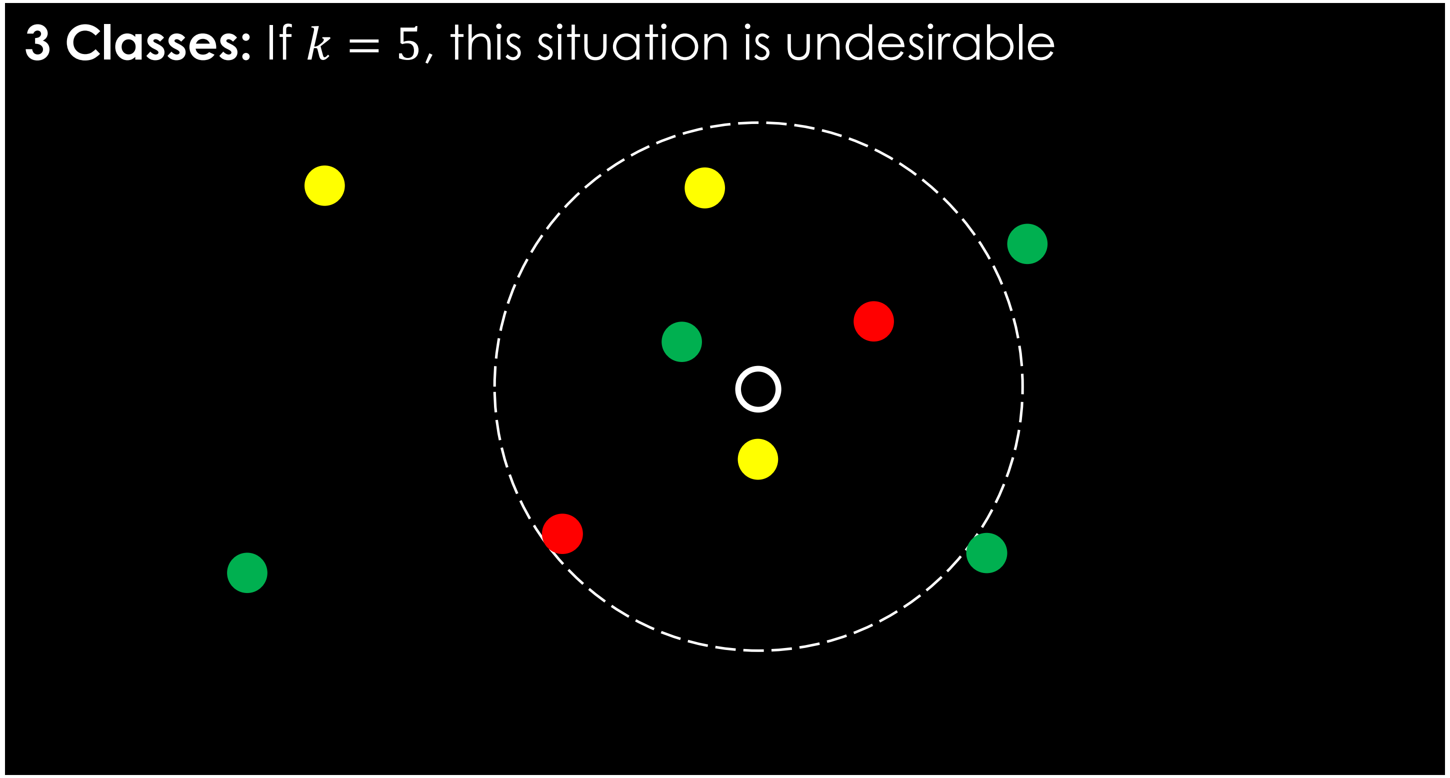
- For multi-classification
  - k is recommended to be the odd number and at least $2C + 1$



3 Classes: If $k = 3$, this situation is undesirable

- For binary classification
  - k is recommended to be the odd number

- For multi-classification (C classes)
  - k is recommended to be the odd number and at least $2C + 1$



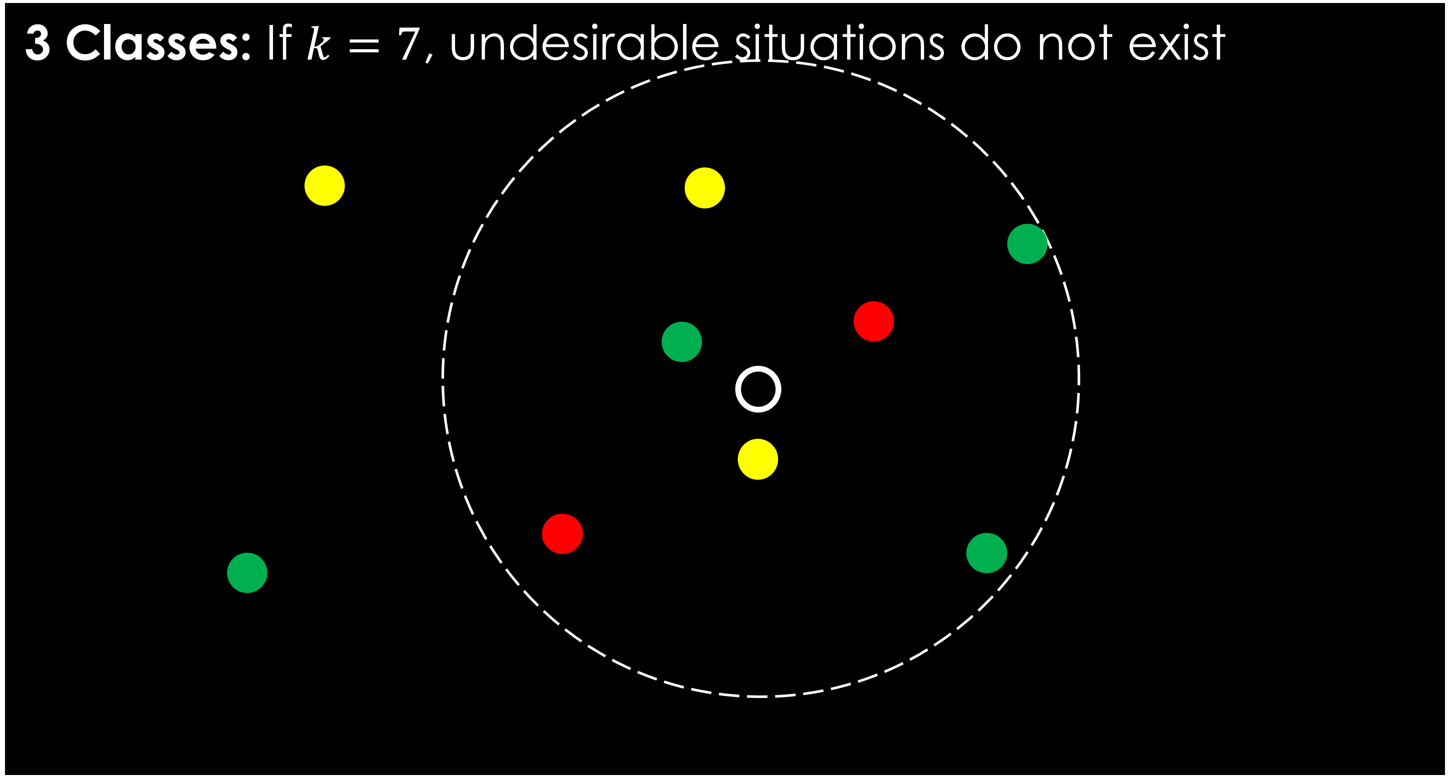
Concepts

The number of nearest neighbor k plays important role for prediction

In common practice, the k is set to be some odd number

- For binary classification
  - k is recommended to be the odd number

- For multi-classification
  - k is recommended to be the odd number and at least $2C + 1$



3 Classes: If $k = 7$, undesirable situations do not exist

Concepts

The number of nearest neighbor k plays important role for prediction

In common practice, the k is set to be some odd number

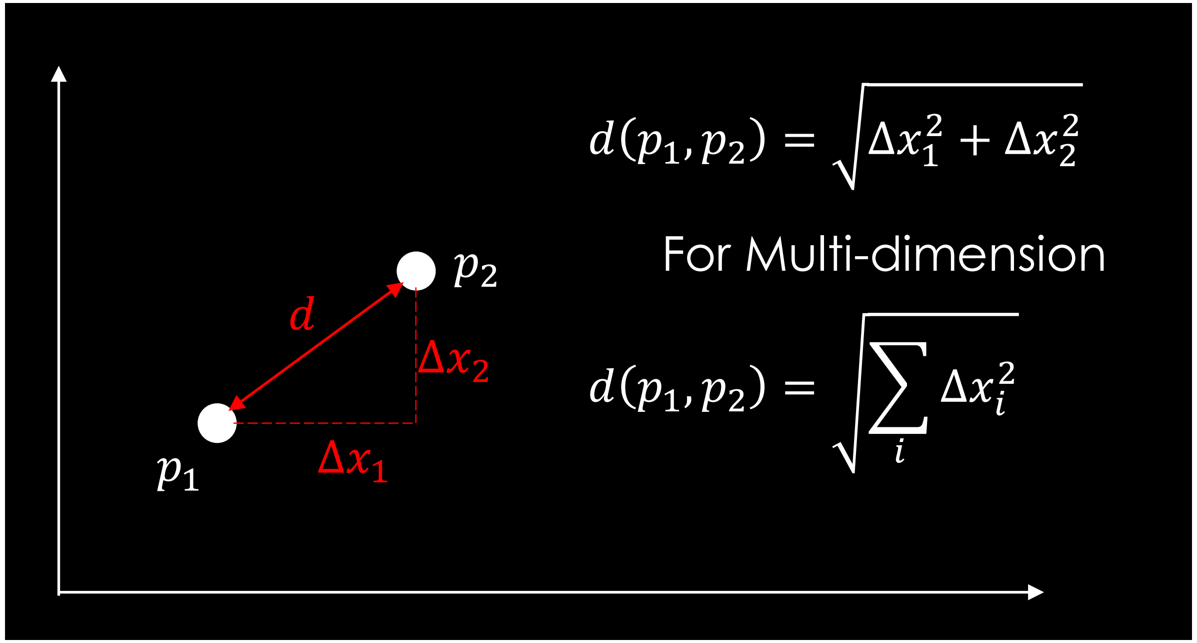
Nearest Neighbors

Determining the k nearest neighbors relies on the **similarity measure**

There are several similarity measures

- Euclidean Distance (0 to ∞)
  - 0: Exactly the same
  - ∞: Completely different

- In practice, ∞ of Euclidean distance is relatively impossible
- Magnitude does matter
- All features are recommended to have the same scale



$$d(p_1, p_2) = \sqrt{\Delta x_1^2 + \Delta x_2^2}$$

For Multi-dimension

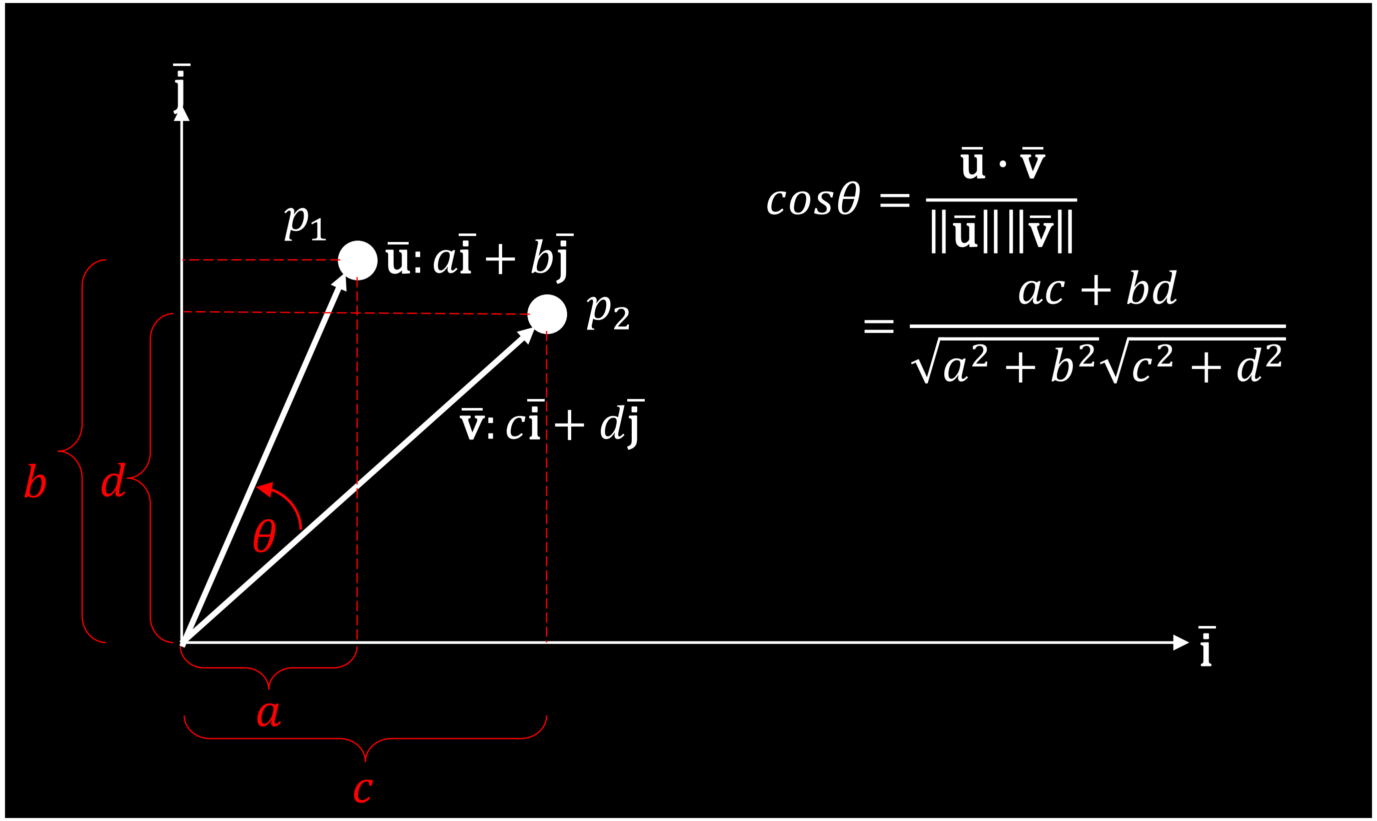
$$d(p_1, p_2) = \sqrt{\sum_i \Delta x_i^2}$$

Nearest Neighbors

Determining the k nearest neighbors relies on the **similarity measure**

There are several similarity measures

- Cosine similarity (-1 to 1)
  - 1: Similar
  - 0: Unable to detect similarity
  - -1: Unsimilar
  - Magnitude does not matter



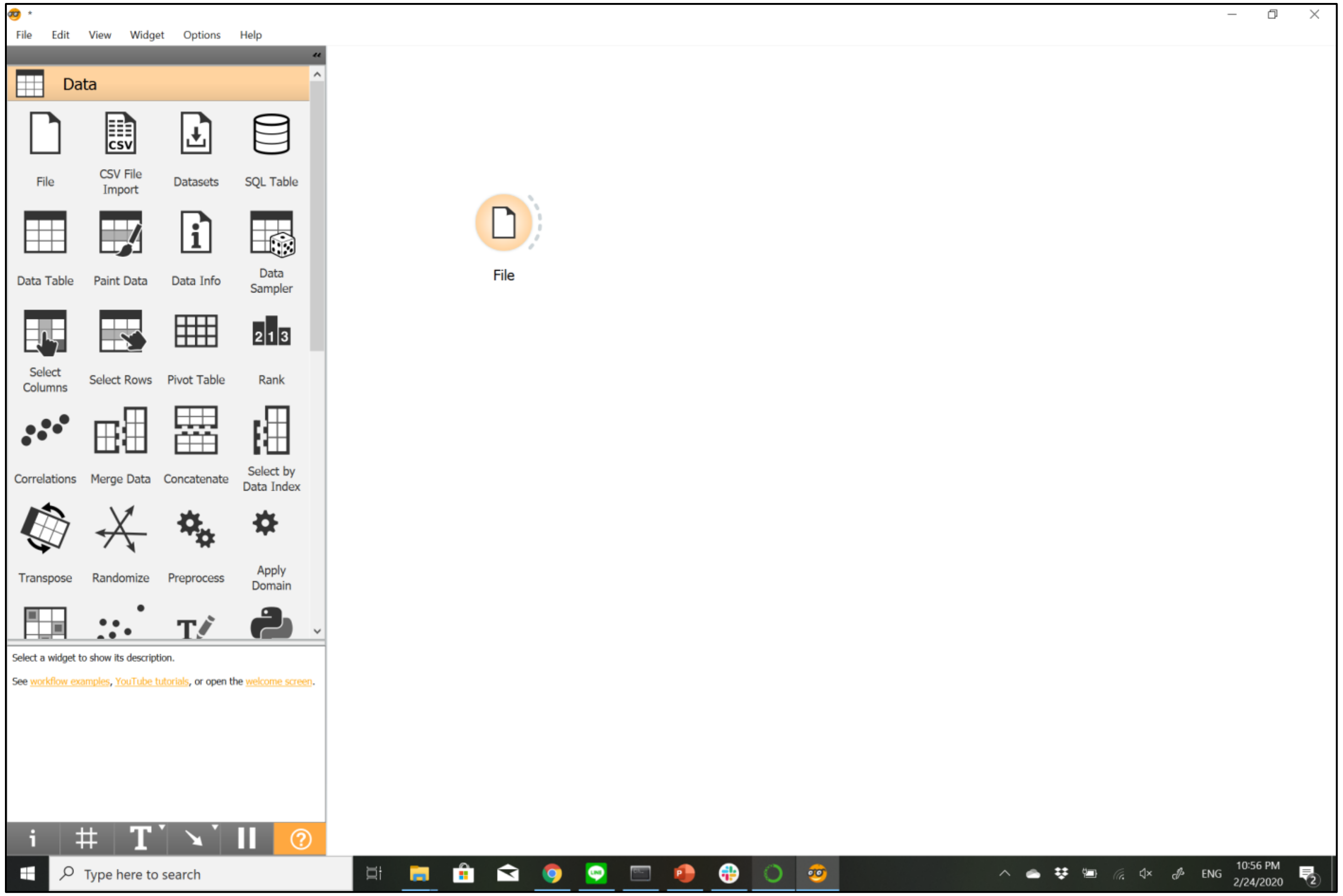
$\bar{j}$
p_1
$\bar{u}: a\bar{i} + b\bar{j}$
p_2
$\bar{v}: c\bar{i} + d\bar{j}$
b
d
θ
$\bar{i}$
a
c

$$cos\theta = \frac{\bar{u} \cdot \bar{v}}{\|\bar{u}\| \|\bar{v}\|} = \frac{ac + bd}{\sqrt{a^2 + b^2}\sqrt{c^2 + d^2}}$$

KNN Classification in Orange

Build your model in 10 seconds

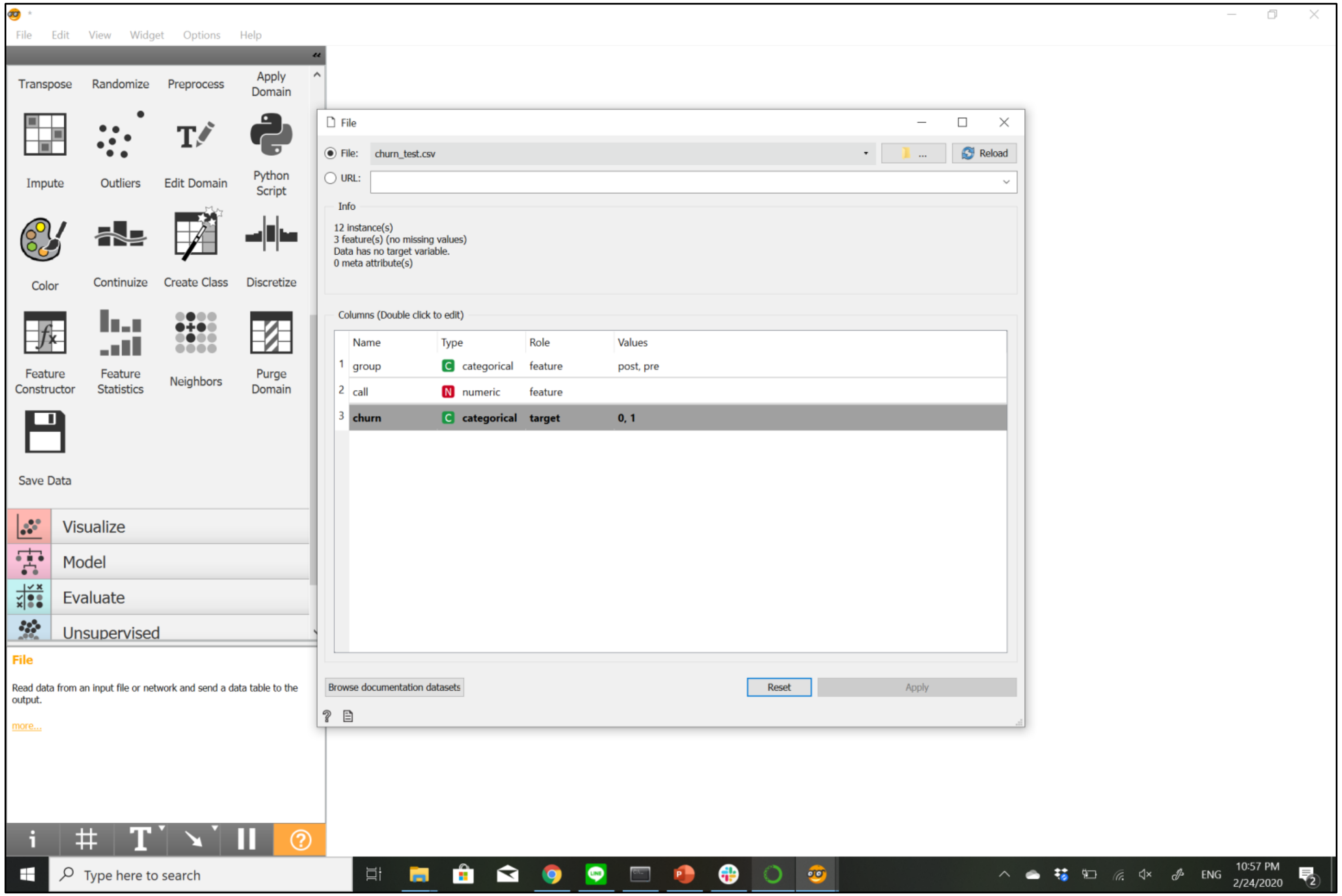
- Import data



Select a widget to show its description.

See workflow examples, YouTube tutorials, or open the welcome screen.

- Identify features and target



File

Read data from an input file or network and send a data table to the output.

more...

File

● File: churn_test.csv

○ URL:

Info

12 instance(s)
3 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

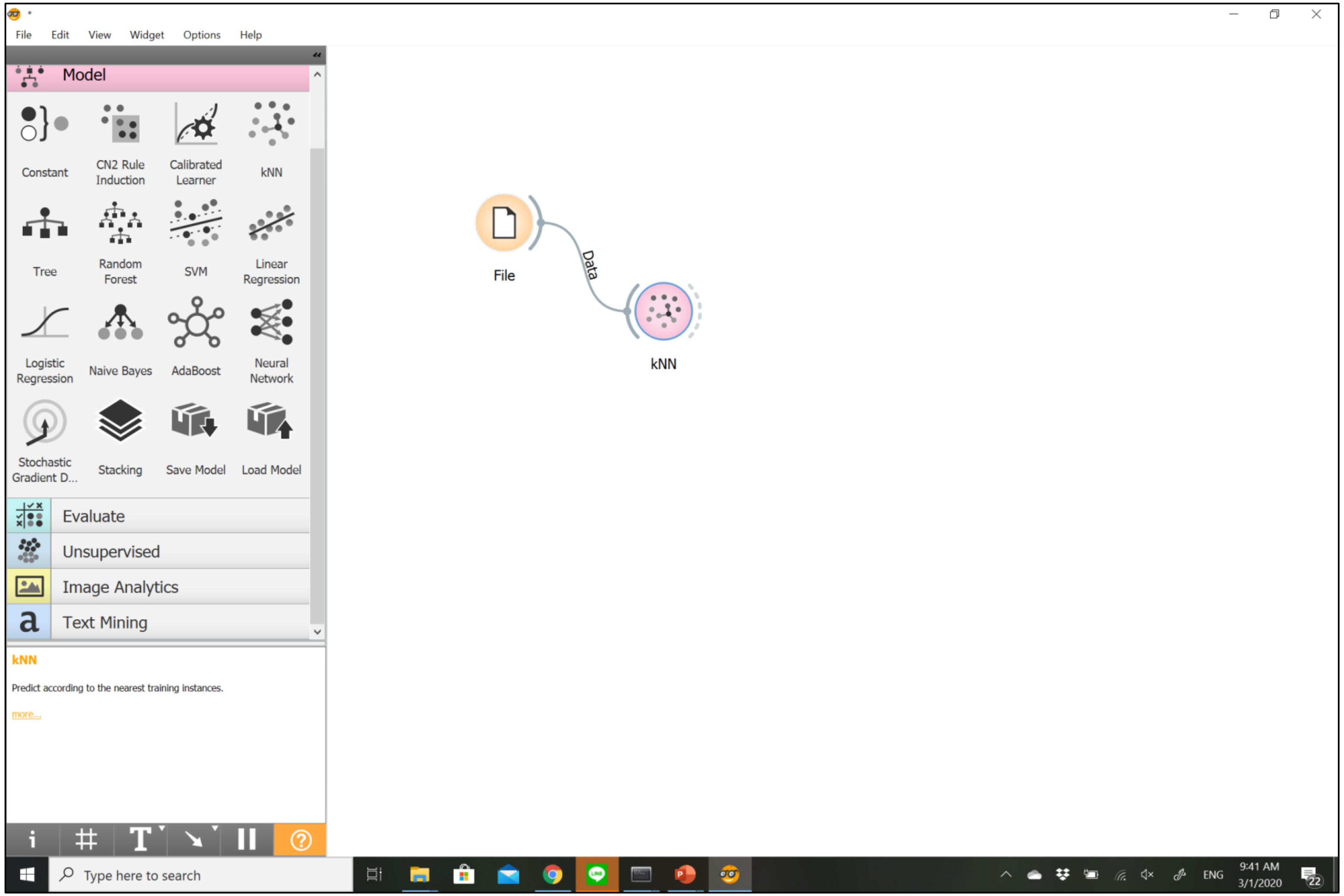
Columns (Double click to edit)

	Name	Type	Role	Values
1	group	C categorical	feature	post, pre
2	call	N numeric	feature	
3	churn	C categorical	target	0, 1

KNN Classification in Orange

Build your model in 10 seconds

- Add model



kNN

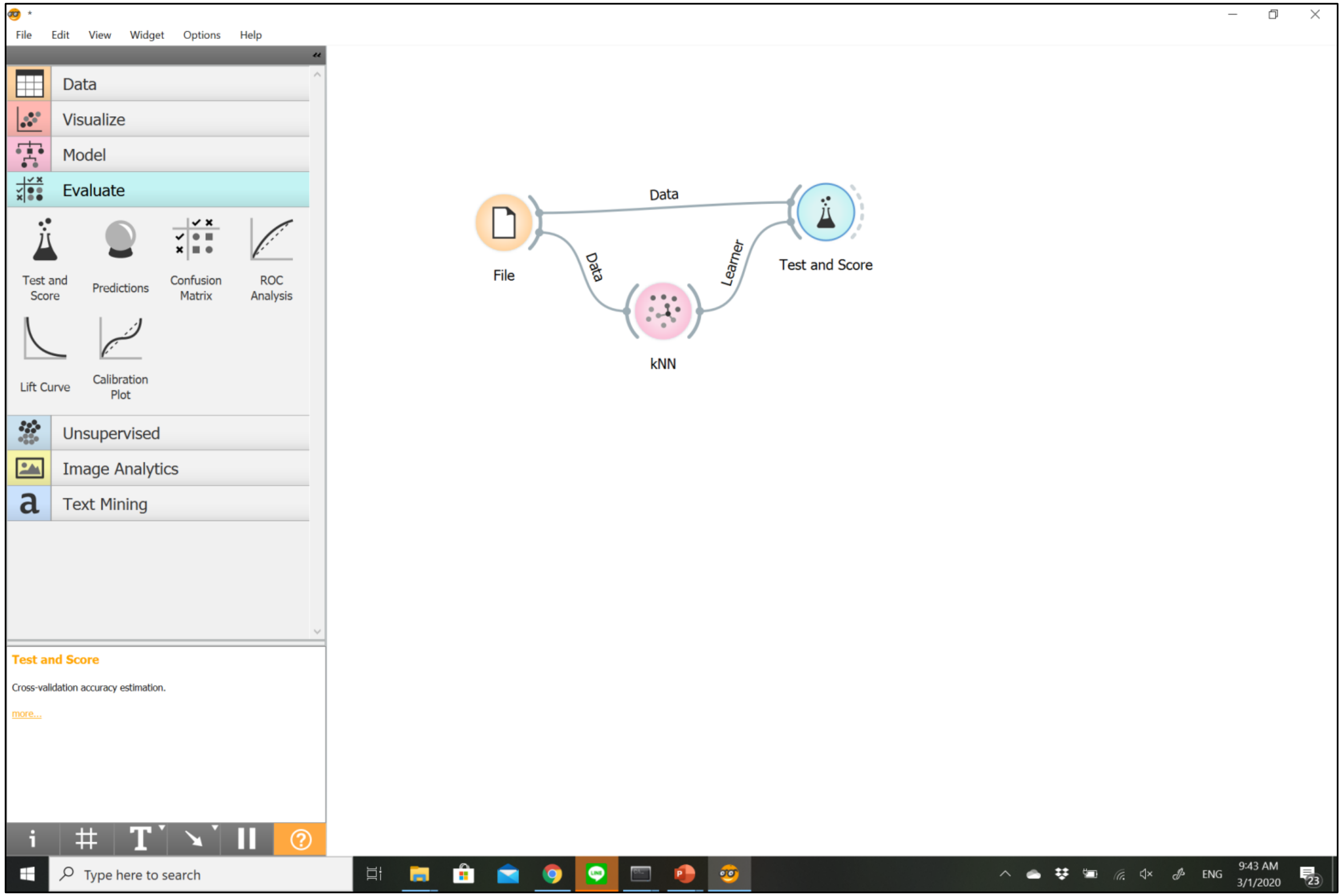
Predict according to the nearest training instances.

more...

KNN Classification in Orange

Build your model in 10 seconds

- Evaluate Model



Test and Score

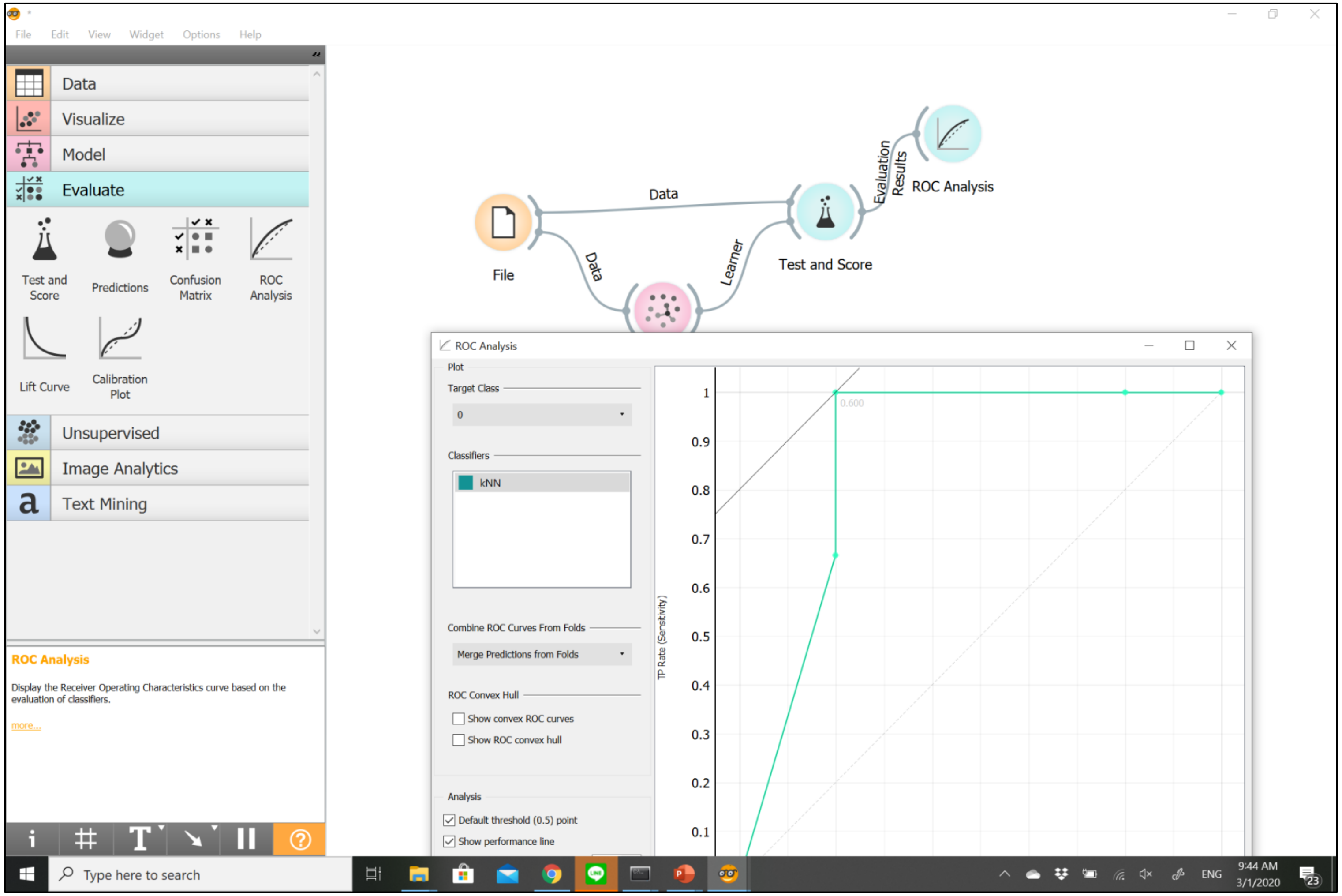
Cross-validation accuracy estimation.

more...

KNN Classification in Orange

Build your model in 10 seconds

- ROC Plot



ROC Analysis

Display the Receiver Operating Characteristics curve based on the evaluation of classifiers.

more...

ROC Analysis

Plot

Target Class: 0

Classifiers: kNN

Combine ROC Curves From Folds: Merge Predictions from Folds

ROC Convex Hull

☐ Show convex ROC curves

☐ Show ROC convex hull

Analysis

☑ Default threshold (0.5) point

☑ Show performance line

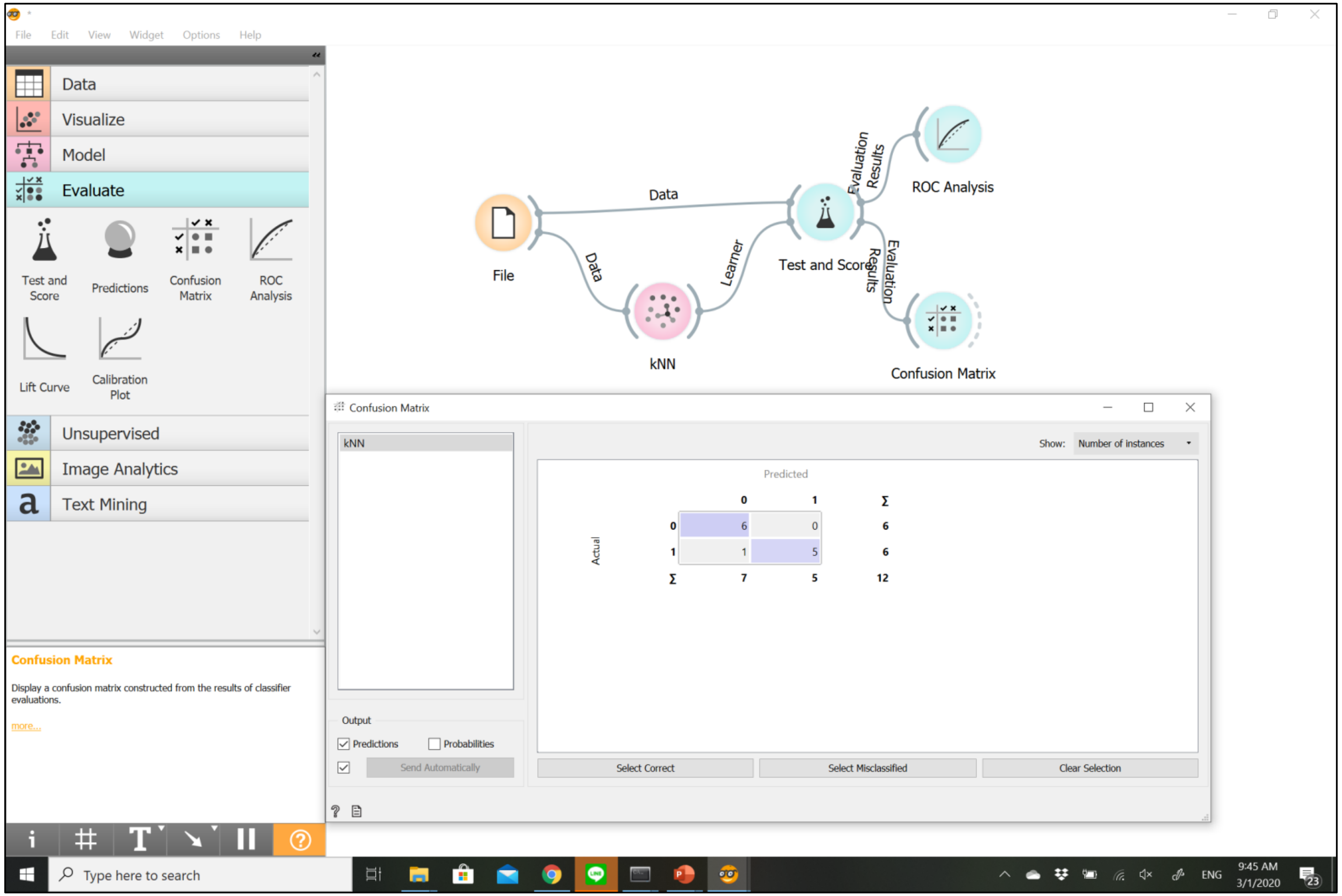
KNN Classification in Orange

Build your model in 10 seconds

KNN Classification in Orange

Build your model in 10 seconds

- Confusion Matrix



Confusion Matrix

Display a confusion matrix constructed from the results of classifier evaluations.

more...

Confusion Matrix

Show: Number of instances

Predicted

Actual	0	1	Σ
0	6	0	6
1	1	5	6
Σ	7	5	12

Output

☑ Predictions □ Probabilities

☑ Send Automatically

- Build your model using Kaggle dataset

Exercise

Our sample data is very small for demonstration purpose

Now it is the time to work with the dataset churn prediction of telecom in Kaggle Competition