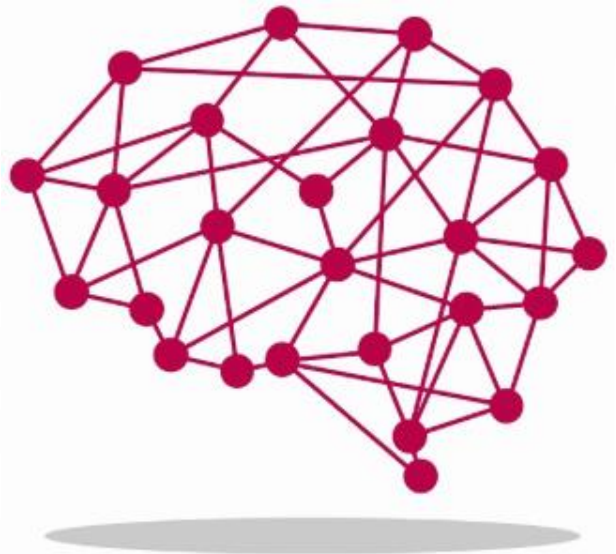
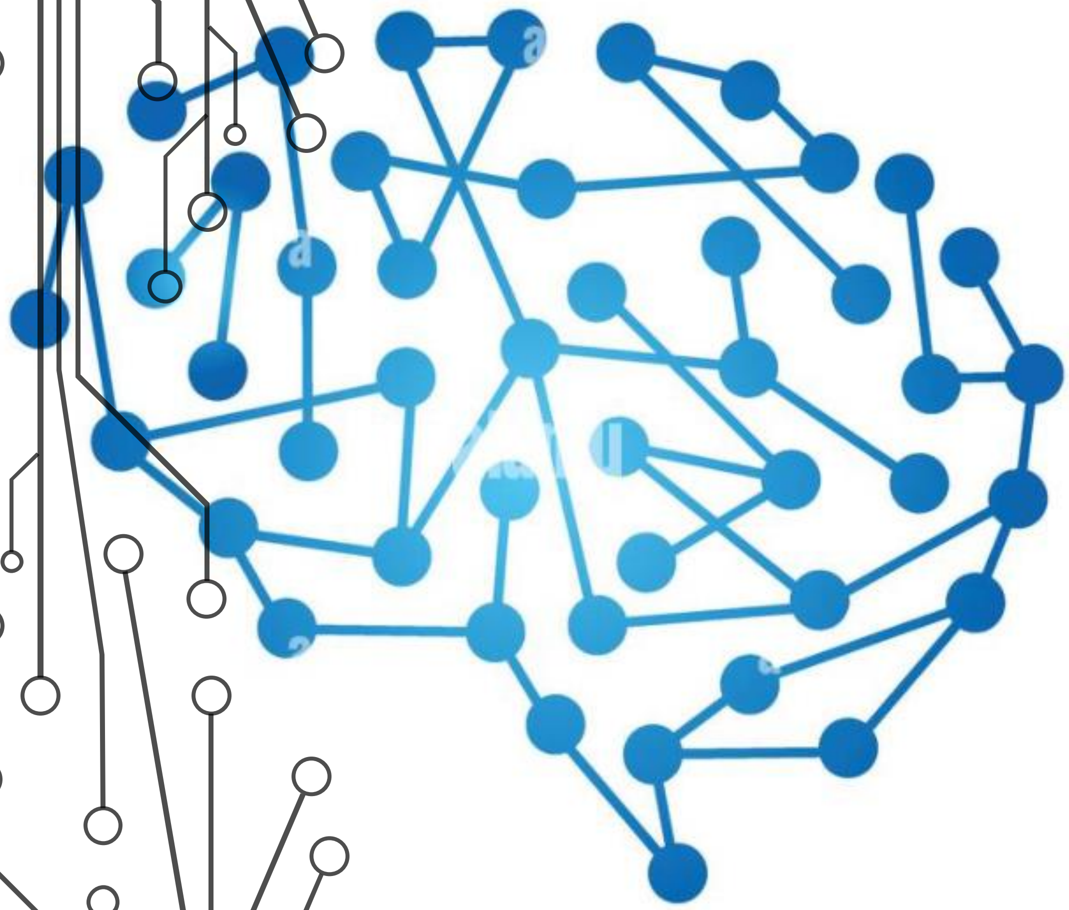
C379 EMERGING TECHNOLOGIES

LESSON 17: MODEL EVALUATION & KNN



L17 LEARNING OBJECTIVES

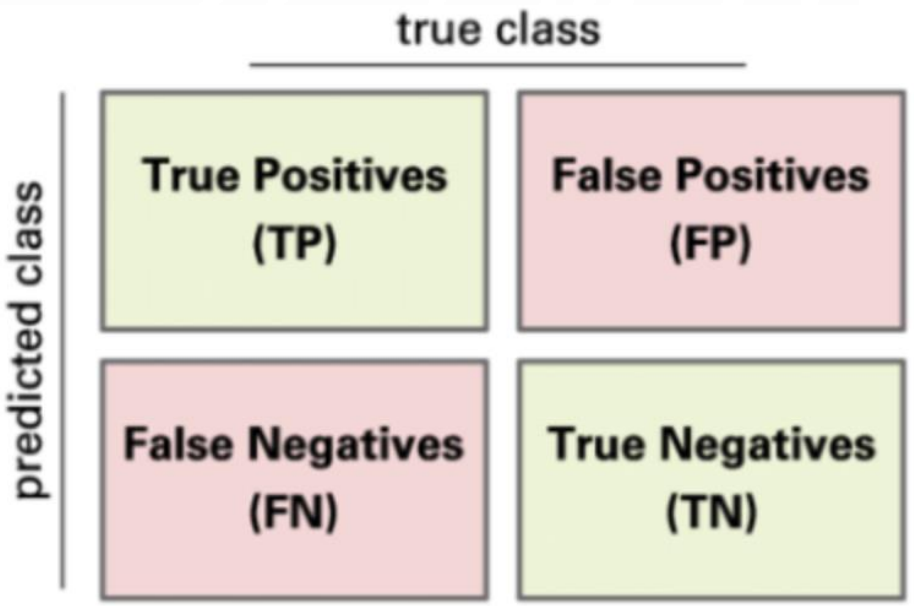
- Evaluate machine learning models using metrics such as accuracy, precision, recall, F1-score, and mean squared error.
- Describe KNN algorithms used to perform classification tasks in AI
- Apply KNN algorithms to train a predictive model
- Test and evaluate the KNN model



MODEL EVALUATION

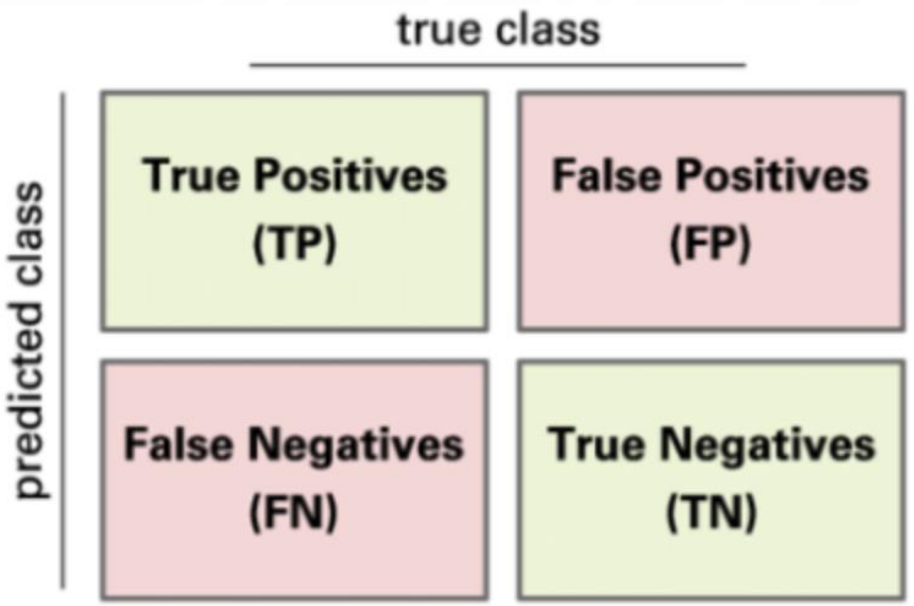
CONFUSION MATRIX

- It is a table with 4 different cases that represent the 4 possible outcomes in a binary classification problem. The model makes a correct prediction have 2 possible cases:
    - **True Positive (TP):** The model predicts the positive class and the observation actually belongs to the positive class.
    - **True Negative (TN):** The model predicts the negative class and the observation actually belongs to the negative class.



CONFUSION MATRIX

- On the other hand, when the model makes a mistake, again 2 possible cases can be happening:
  - **False Positive (FP):** The model predicts the positive class and the observation actually belongs to the negative class.
  - **False Negative (FN):** The model predicts the negative class and the observation actually belongs to the positive class.



PERFORMANCE METRICS

- There are some metrics that you can calculate to make more sense of these numbers above. The following are some of the most important metrics we can get from the quantities observed in the confusion matrix:
  - **Accuracy:** Proportion of cases correctly identified by the classifier (*N* represents the total number of observation in the testing dataset):

$$\boldsymbol{accuracy = \frac{TP + TN}{N}}$$

PERFORMANCE METRICS

- **Precision:** Proportion of correct positive predictions. In this problem, this is the proportion of cases when the model is correct when it predicts defaults:

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

- **Recall (or Sensitivity):** Proportion of observed positives that were predicted correctly as positives. In this problem, this is the proportion of actual defaults that the model can correctly identify:

$$\boldsymbol{recall = \frac{TP}{TP + FN}}$$

PERFORMANCE METRICS

- **F1 Score** (or balanced F-score or F-measure): It can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal.
- The formula for the F1 score is:

$$\boldsymbol{f1 = \frac{2 \times (precsion \times recall)}{precision + recall}}$$

These 4 metrics will help you understand which types of mistakes the model is making. So, let's calculate the accuracy, precision, recall and F1 Score for your 3 predictive models.

PERFORMANCE METRICS

		Actual	
		True	False
Predicted	True	33	2
	False	3	12

- **Accuracy** = (TP+TN)/N

  = (33+12)/(33+2+3+12)

  = 0.9

- **Recall (Sensitivity)** = TP/(TP+FN)

  = 33/(33+3)

  = 0.92

- **Precision** = TP/(TP+FP)

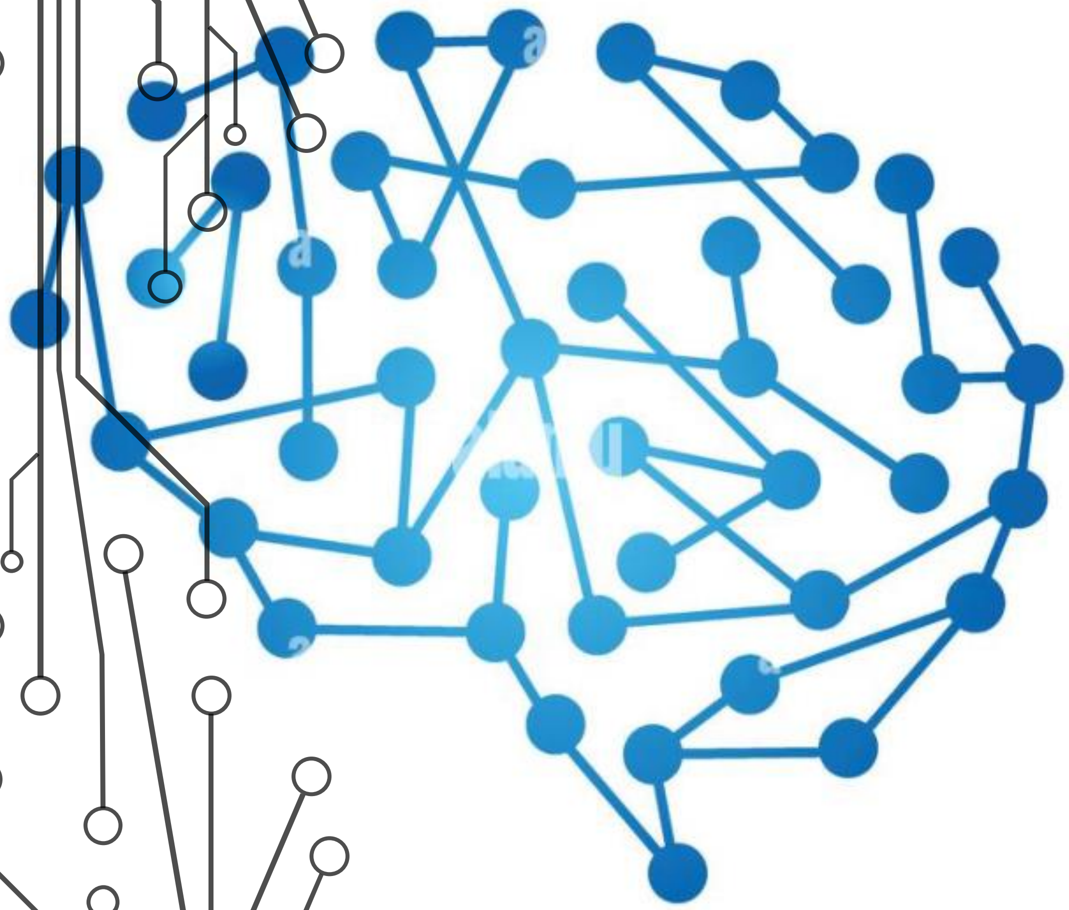
  = 33/(33+2)

  = 0.94

- **F1 Score** = (2*Precision*Recall)/(Precision + Recall)

  = (2*0.94*0.92)/(0.94+0.92)

  = 0.93



MODEL BUILDING

PART 4 – UNDERSTANDING KNN

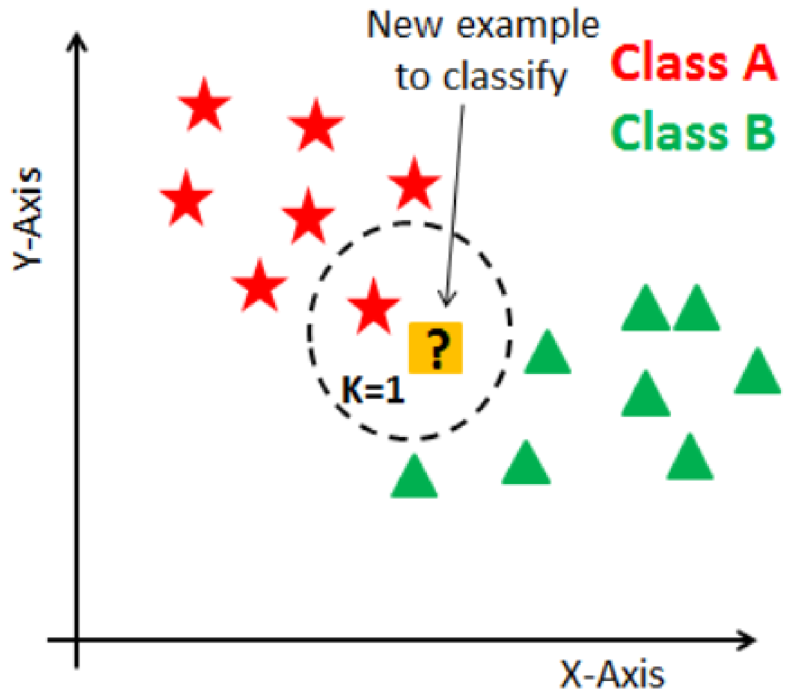
K-NEAREST NEIGHBOURS (KNN)

- KNN (*k*-nearest neighbours) method can be used for both regression and classification
- It belongs to the class of *non-parametric** models, hence the predictions are not based on the calculation of any parameters.
- KNN is also known as a *lazy algorithm*. It means it does not need any training data points for model generation.
- Despite its simplicity, it frequently produces good results.

* The term *non-parametric* means there is no assumption for underlying data distribution. In other words, the model structure is determined from the dataset.

KNN – HOW THE ALGORITHM WORKS?

- For a fixed number, K, which is the number of neighbours, and a given observation whose target value we want to predict, do the following:
  - Find the *K* data points that are closest in their feature values to the given data point
  - Calculate the average target value for those *K* data points
  - That calculated average is the prediction for the given data point



BINARY VS MULTICLASS CLASSIFICATIONS

- **Binary classification** is when there are only two possible outcomes like a person is infected with COVID-19 (outcome =1) or is not infected with COVID-19 (outcome=0).
- In **multi-class classification**, there are multiple outcomes like the person may have the flu or an allergy, or cold or COVID-19.

KNN CLASSIFIER

1. Importing KNN method from scikit-learn framework

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_train)
y_pred_knn = model.predict(X_train)
```

2. Calling the KNN method with the number of neigbours = 3

3. From the training dataset and labels provided in the parameters, the KNN method train the model to make prediction.

4. The predict method is used to classify incoming data point

- *X_train* matrix contains all the features of the training set
- *y_train* matrix contains the label from the training set

SAVING A PREDICTIVE ANALYTICS MODEL

```
import pickle as pk

model_filename= "./model/knn.mdl"
with open(model_filename, "wb") as file:
    pk.dump(knn, file)
print("Model Saved")
```

1. Importing pickle method to serialize and de-serialize a Python object structure

2. Open a binary file and write the pickled representation of the AI model into it

LAB DEMONSTRATION

LAB 17-1

USING A SIMPLE KNN MODEL

RECALL FROM L01: WHERE CAN AI BE APPLIED?

Let's zoom in on the problem statement

……….

The challenge with existing appliances from current consumer electronic companies lies in the fact that they are closed-systems

……….

To ascertain

(i) availability or

(ii) estimated time before availability, ………

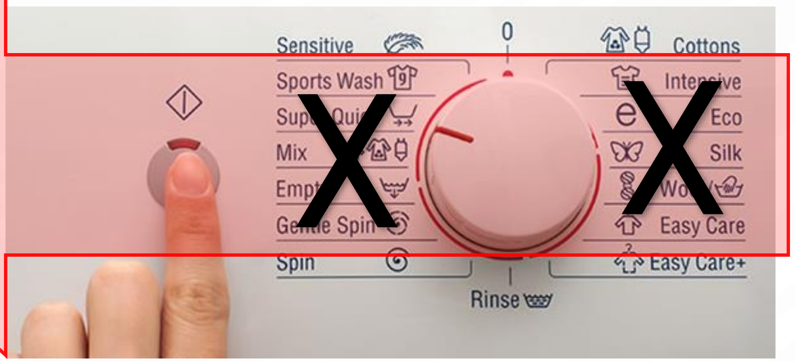
(iii) update at what stage of the washing or drying programme that the appliance is in is desirable.

……….

Closed-systems

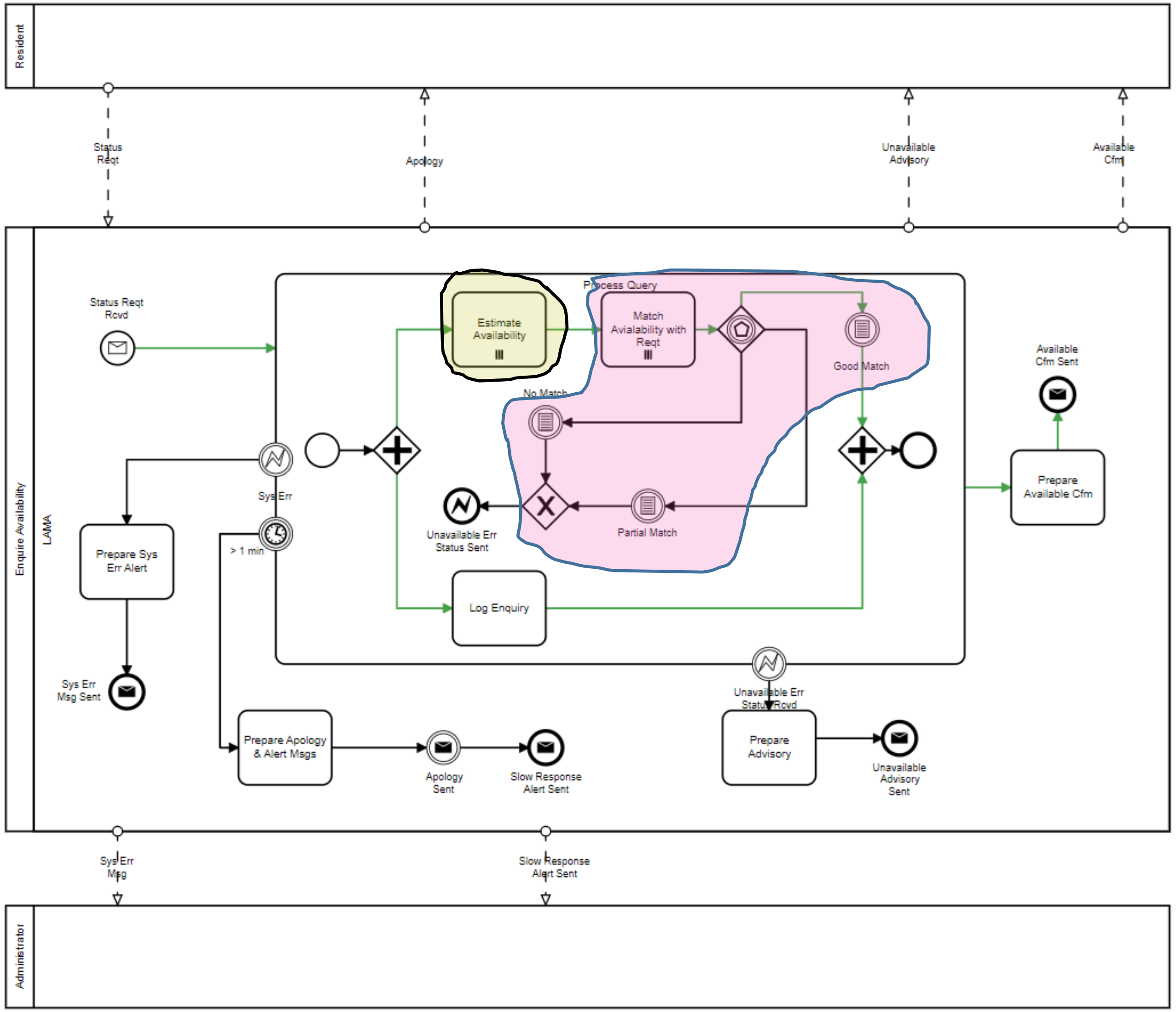
NO explicit data is sent or received by the washing machine to provide any information of status

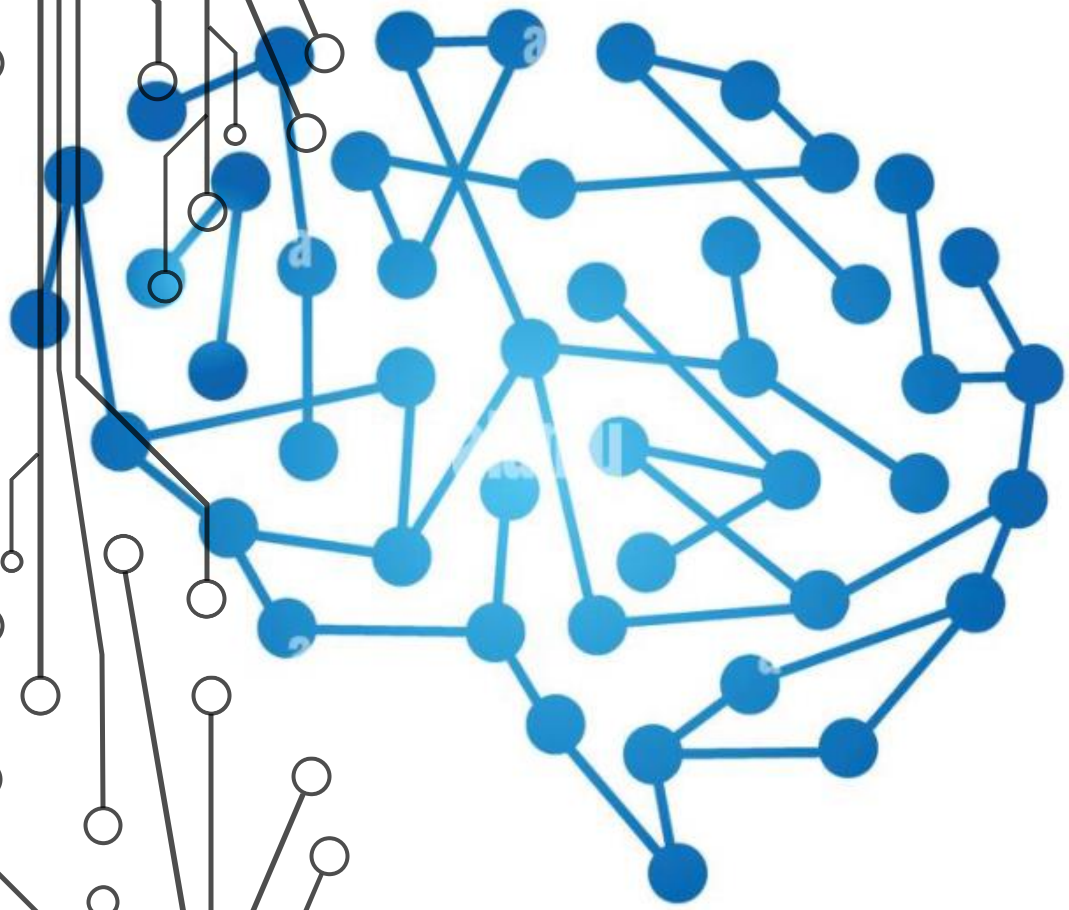
Status must be deduced based on indirect data, e.g. power consumption



ENQUIRE AVAILABILITY, APPLICATION OF AI

An AI solution can be created to estimate when (date/time) the machine(s) will be available and determine how closely the requests can be met





PROBLEM SOLVING

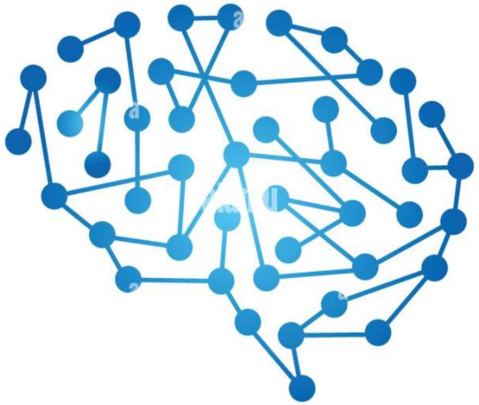
APPLYING KNN

PRACTICAL

LAB 17-2

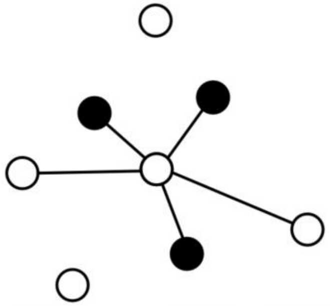
BUILDING A KNN MODEL TO SOLVE LAUNDROMAT PROBLEM

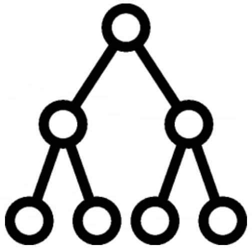
UNSCHEDULED ASYNCHRONOUS E-LEARNING

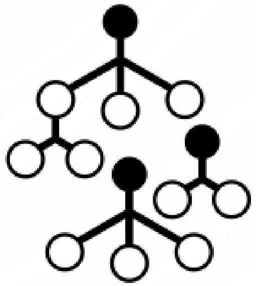


Watch the following videos before Lesson 18:

- Decision Tree - https://www.youtube.com/watch?v=7VeUPuFGJHk







60 mins