Machine Learning for Everyone

What is Machine Learning?

1. Artificial Intelligence (AI)
    - a huge set of tools for making computer behave intelligently
    - machine learning is a subset of AI

2. Machine Learning
    - set of tools for making inferences and predictions from data
    - predict outcome of the future
        - Will it rain tomorrow? Yes (75% probability)
    - infer the cause of the events and behaviours
        - Why does it rain? Temperature, humidity level , time of the year, location
    - infer patterns
        - What are the different types of weather conditions? Rain, sunny, fog, overcast
    - inferences help to do predictions
    - interdisciplinary mix of statistics and computer science

- give computer ability to learn without explicitly programmed (can learn without step by step instructions)
- learn patterns from existing data and apply it to new data
- rely on high-quality data

3. Data Science
    - discovering and communicating insight from the data
    - machine learning is important for data science work especially making predictions

4. Machine learning model
    - statistical representation of real-world process based on data
    - new input → model → outcome

5. Types of machine learning
    1. Reinforcement learning
        - used to decide sequential actions
        - Exp: robot decides its next move in chess game
    2. Supervised learning
        - target: heart disease
        - features are used to help to predict the target

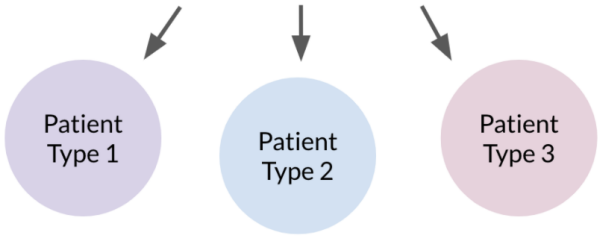
Features

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
50	F	196	0	False	non-anginal pain	98	False
53	F	215	0	True	asymptomatic	110	True
62	M	245	3	False	typical angina	126	True
48	M	190	0	True	non-anginal pain	99	False
70	M	201	0	True	typical angina	105	False

- features (input) → model → predictions (output)
- the training data is label → the target are known

3. Unsupervised learning
    - training data only have features

Age	Sex	Cholesterol	Cigarettes per day	Family history of heart disease	Chest pain type	Blood sugar	Heart disease
55	M	221	5	True	typical angina	118	True
53	F	199	0	True	non-anginal pain	98	True
53	F	215	0	True	asymptomatic	110	True
62	M	245	3	False	typical angina	126	True
...	...	...	...	...	...	...	...



- useful for:
    - anomaly detection
    - clustering: divide the data into groups based on similarity

- the model will find its own patterns

6. Training data
    - existing data to learn from
    - training a model: when a model is being built from training data

7. Machine learning workflow
    - scenario

Our dataset: NYC property sales from 2015-2019

Includes:

- Square feet
- Neighborhood
- Year built
- Sale price
- And more!

Our target: Sale price

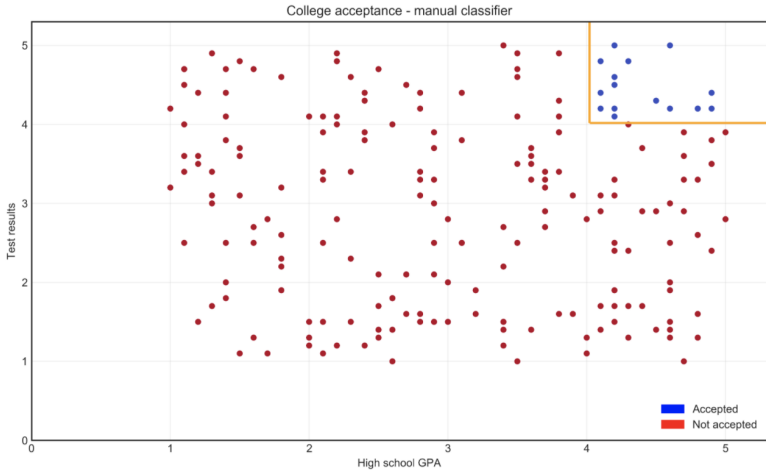
    - 4 steps
        1. Extract features
            - reformatting the dataset
            - need to decide what features to begin with
            - choosing features and manipulating the dataset
        2. Split dataset
            1. train dataset
            2. test dataset
        3. Train model
            - by using train dataset
            - the train dataset is inputted into a chosen

- the train dataset is inputted into a chosen machine learning model

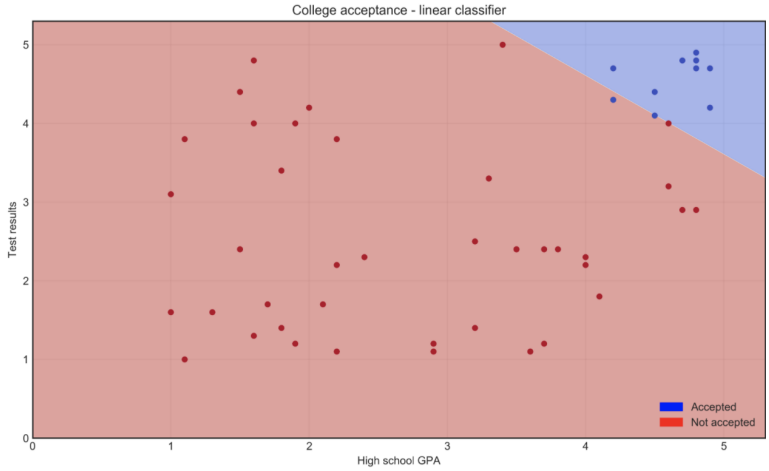
4. Evaluate
    - the train dataset cannot be used again to evaluate because the model already seen the data
    - the test dataset (unseen data) for evaluate in order to determine the accuracy of the model
    - ways to evaluate:
        - What is the average error of the predictions?
        - What percent of apartments did the model accurately predict within a 10% margin?
    - after evaluate the model, we will think of is the performance of the model good enough?
        - yes → the model is ready to use
        - no → the model need to be re-train and tune it
            - tune: change the model's options, add or remove features (take some time)
            - if the performance after tuning did not improve which means the data is no enough

Machine Learning Models

1. Supervised learning
    - a labelling machine by taking observations (data)
    - types
        1. Classification
            - assign category
            - 80% of data are used to predict the results and another 20% are plotted manually by the human because only have 2 features

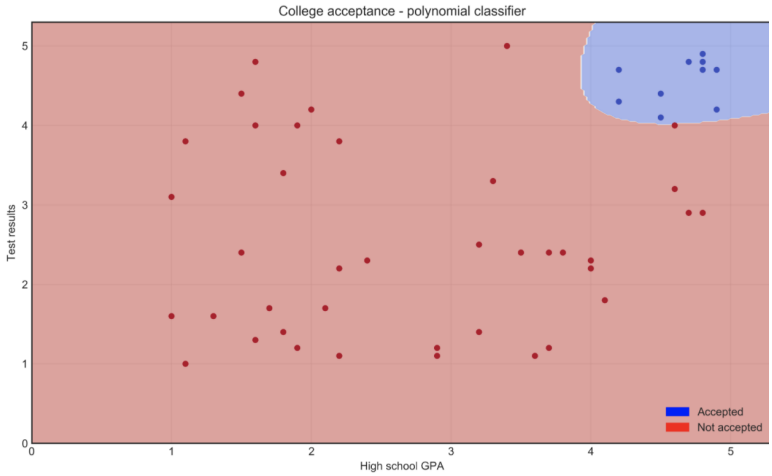


- Support Vector Machine (linear classifier) can be used in order make the data easy to understand



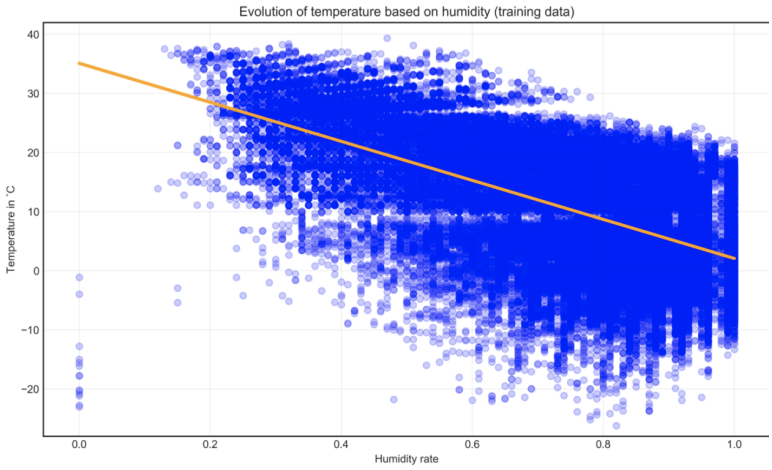
It misclassifies 2 blue points → the applications are wrongly predicted as rejected because it tries to separate the results by using straight

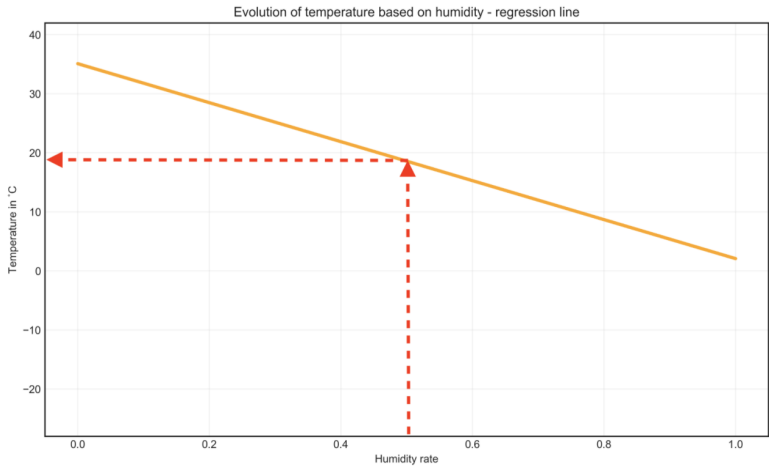
- Support Vector Machine (polynomial classifier) which allow curved lines



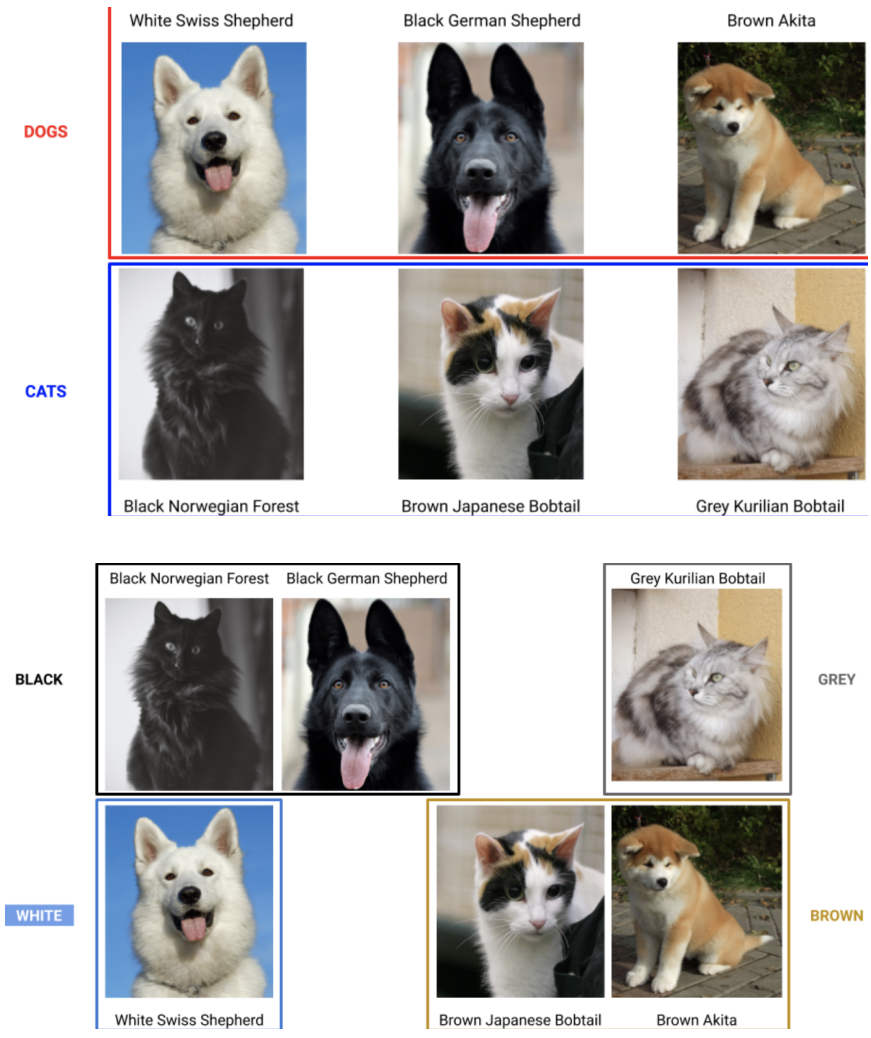
The results were predicted correctly

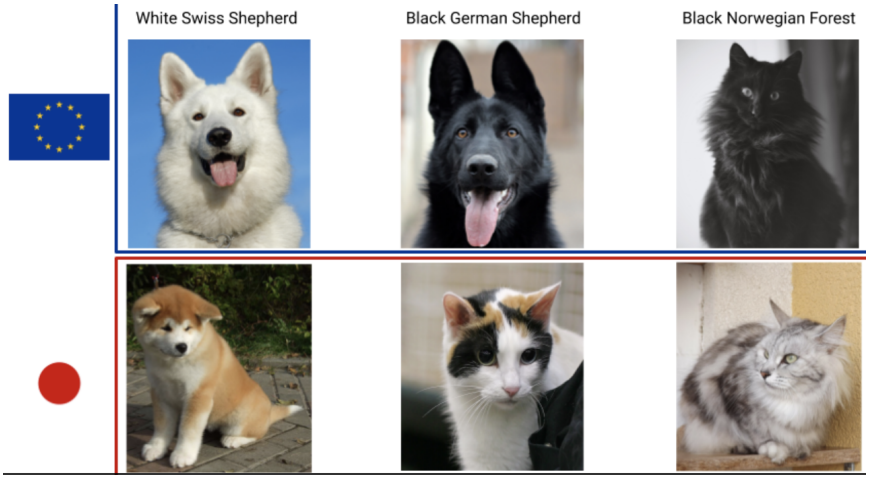
2. Regression
    - assign a continuous variable (the variable can take any value)
    - able to predict a value



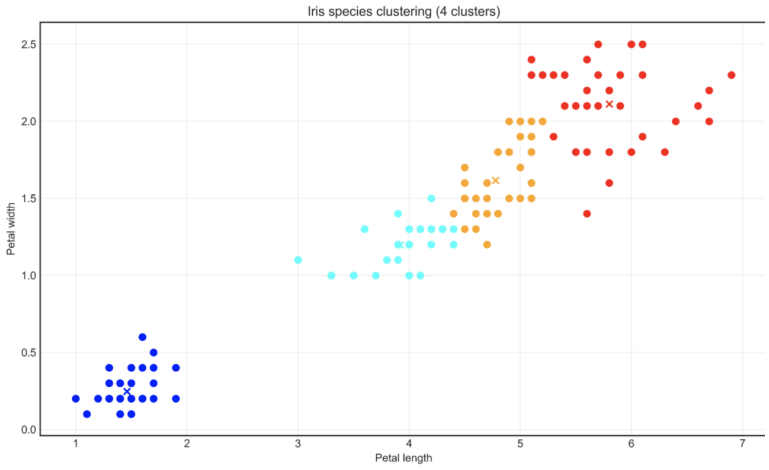


2. Unsupervised learning
    - similar with supervised learning but it do not have a target column
    - no guidance
    - looks at whole dataset
    - try to find patterns from the dataset
    - types
        1. Clustering
            - identify groups in dataset
            - find similarity
            - Exp: group the dog and cats in species cluster, colour cluster, origin cluster

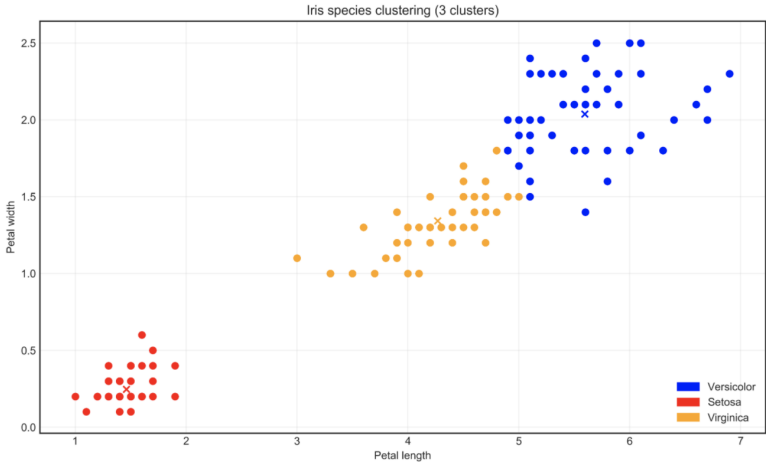




- it would tell the user why it classify like these, the user need to understand by themselves
- Example model:
    1. K-Means
        - need to specify the number of clusters in advance that would like to identify



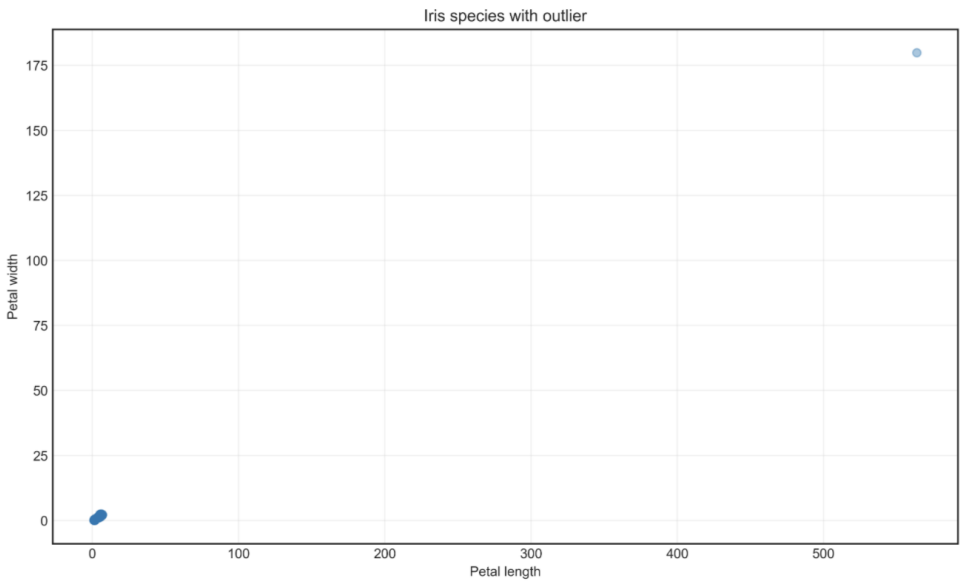
Specify the cluster have 4 means



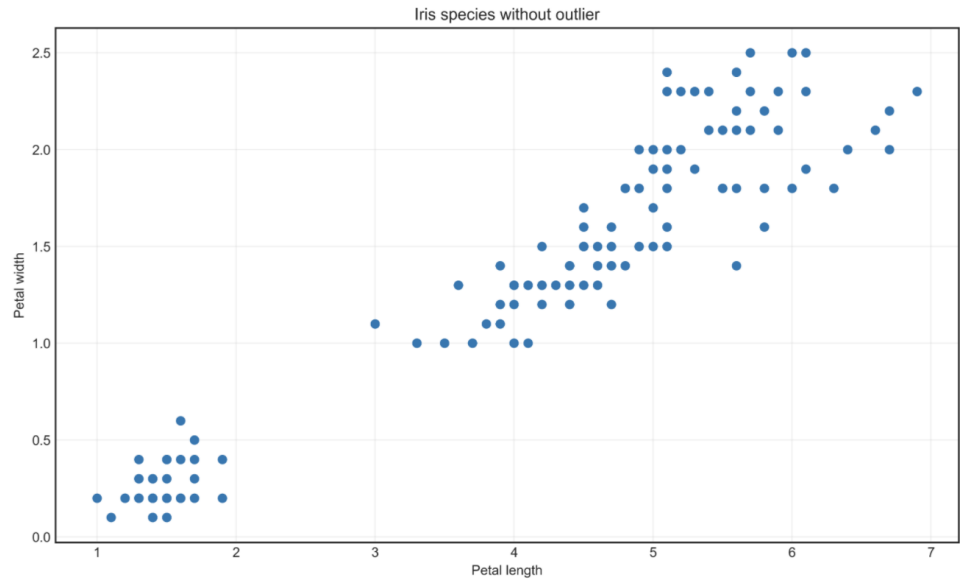
Specify the cluster have 3 means

2. DBSCAN (density-based spatial clustering of applications with noise)
    - not require to specify the number of cluster in advanced
    - need to define what constitute a cluster (Exp: minimum number of observations in one cluster)

2. Anomaly detection
    - detecting outliners (observations that strongly differ from the rest)
    - Exp:



The light blue point is an outliner (error) so that it is removed as shown as the figure below

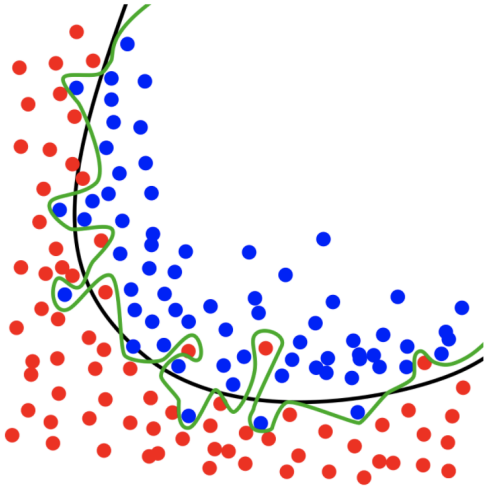


- anomaly detection use cases
    1. discover the device that fail faster or last longer
    2. discover fraudsters that manage trick the systems
    3. discover patients that resist a fatal disease

3. Association
    - finding relationships between observations/ finding events that happen together

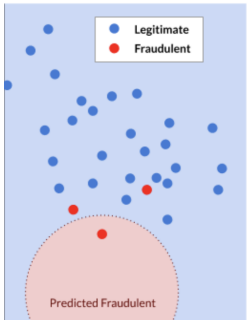
- often used for market basket analysis

3. Evaluating performance (What aspects to evaluate performance?)
    1. Supervised learning
        1. Classification
            1. Overfitting
                - performs great on training data but poorly on testing data
                - the model memorised training data and cannot generalise learnings to new data
                - Exp: The green line is overfit



It classify the data perfectly → perform great on the dataset but poorly on the unseen data. The black line make more errors on the specific dataset but generalise better

            2. Accuracy
                - accuracy = correctly classify observations / all observations
                - not always a best metric because of fraud example



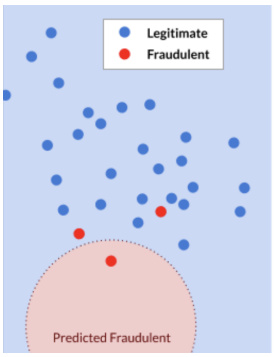
Accuracy of this model:

$$\frac{28\ correctly\ classified}{30\ total\ points} = 93.33\%$$

- Misses majority of fraudulent transactions
- Need a better metric

It missed all the fraudulent so we need a better metric which is confusion matrix

- confusion matrix



		Actual values	
		Fraudulent	*Not Fraudulent*
Predicted	*Fraudulent*	1 true positives	0 false positives
	Not Fraudulent	2 false negatives	27 true negatives

The red points outside the red area is false negative (something is totally true but the result is false). The red point inside the red area is true positive. There are no false positive which means legitimate points that are incorrectly predicted as fraudulent (no blue points in red area). There are 27 true negative which means the legitimate points are correctly predicted as not fraudulent.

- Why need to fill out the matrix above?
  - want the better metric than accuracy for the scenario
  - **Sensitivity** = true positives / true positives + false negatives (used to value true posotives)

		Actual values	
		Fraudulent	Not Fraudulent
Predicted	Fraudulent	1 true positives	0 false positives
	Not Fraudulent	2 false negatives	27 true negatives

How many fraudulent transactions did we classify correctly?

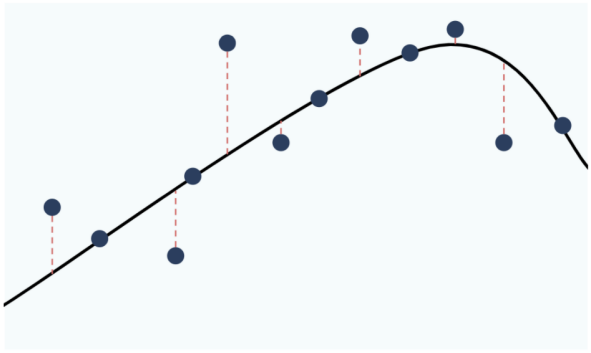
$$Sensitivity = \frac{true\ positives}{true\ positives + false\ negatives} = 1/3 = 33.33\%$$

- Rather mark legitimate transactions as suspicious than authorize fraudulent transactions

Sensitivity = 33.33% which means it is a bad score for the results. Therefore, the model needs improvement

- **Specificity** = true negatives / true negatives + false positives (used to value true negatives)
  - useful metric for spam filter

2. Regression



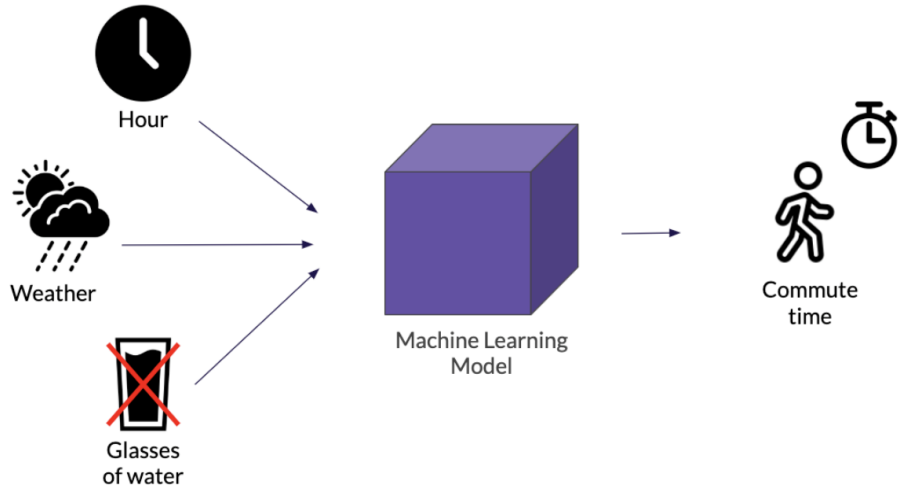
- Error = distance between point (actual value) and line (predicted value)
- Many ways calculate this. e.g, root mean square error

A general idea to evaluate the performance

2. Unsupervised learning
   - do not have predictive outputs so that do not have the values to compare with
   - will be evaluated by human

4. Improve performance
   - several options
     1. Dimensionality reduction

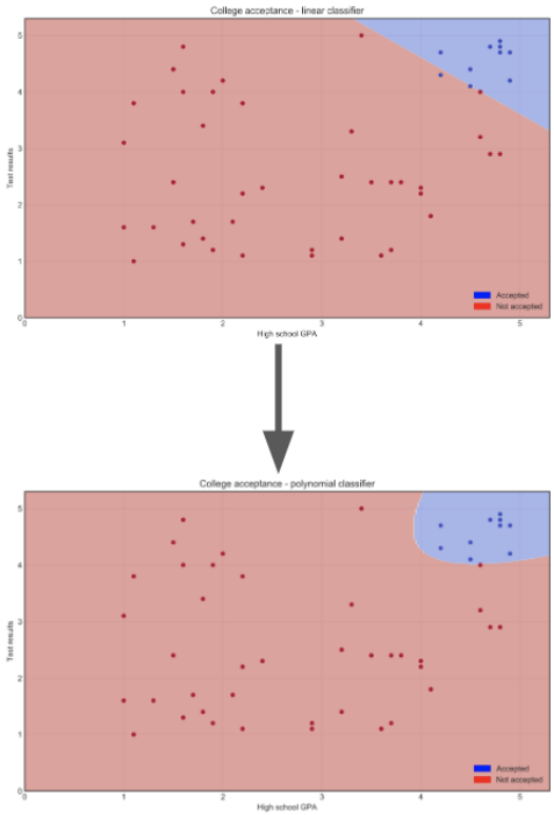
- reduce number of features
- when making prediction, do not means that the more features, the more accurate the predictions because:
  - **irrelevant:** some of the features are irrelevant



When predict how long is needed to go to the office, number glasses of water is become irrelevant

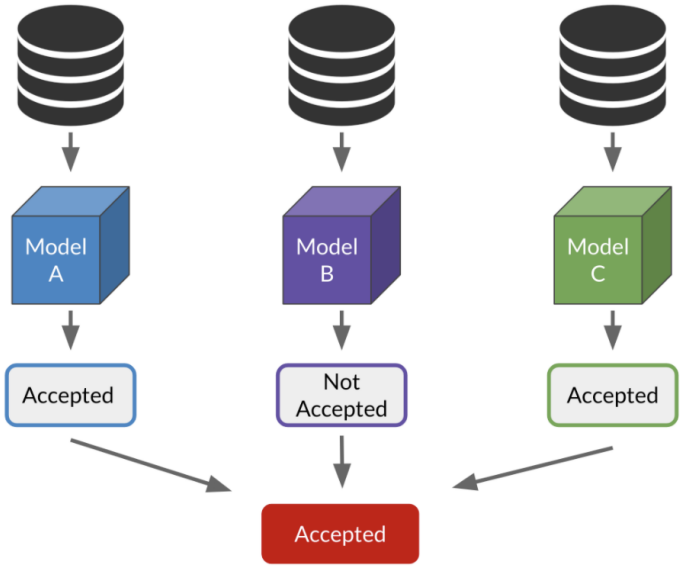
  - **correlation:** some features carry similar information
    - can only keep one feature (Exp: height and shoe size → keep height because tall people need large shoe size)
    - collapse multiple features into one underlying feature (Exp: height and weight → BMI (Body Mass Index))

2. Hyperparameter tuning
   - parameter whose value is used to control the learning process
   - there are many hyperparameters can be tuned in SVM model which will impact the model's performance
   - Exp: SVM algorithm hyperparameter

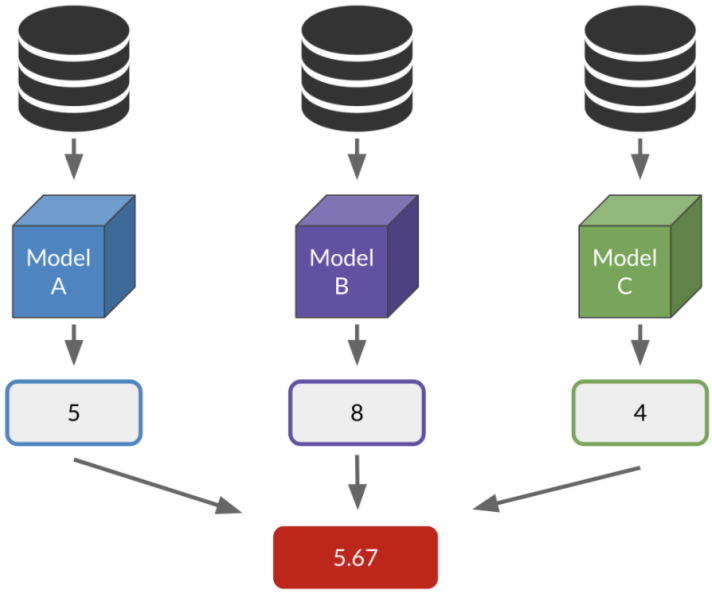


kernel: linear → poly

3. Ensemble methods
    - combine several models to produce one optimal model



Classification: Model A and C said they are accepted but B is not accepted. Then the observation is accepted because it was the most common prediction amongst all models.

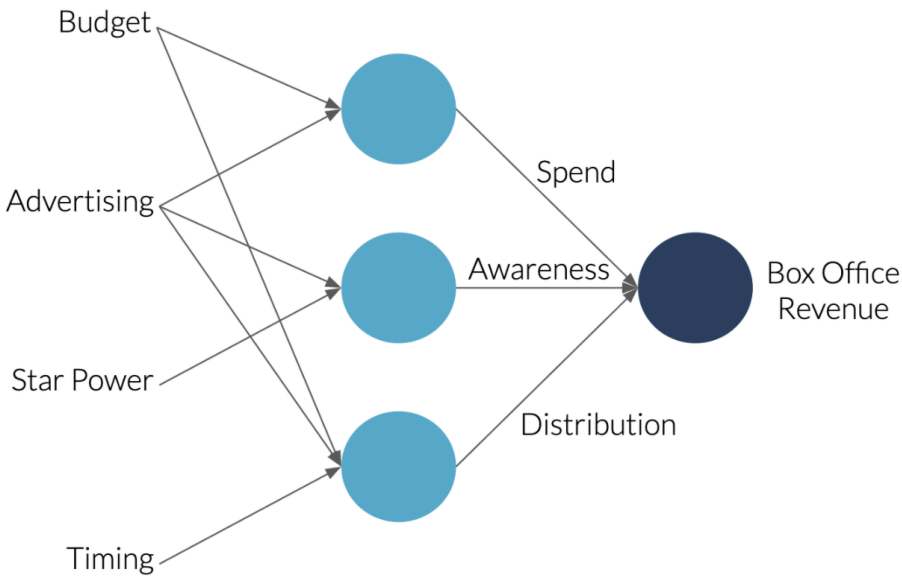


Regression: calculate the average

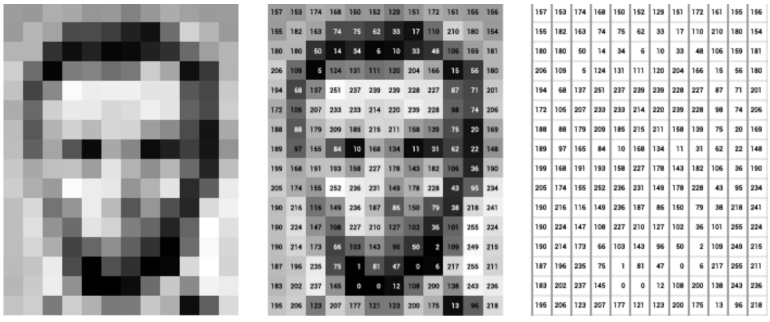
Deep Learning

1. Deep learning
    - uses an algorithm (neural networks)
    - special area of machine learning that can solve more problems but it requires more data than the traditional machine learning
    - it is best to used when the inputs are image or text
    - deep learning = neural network with many neurons → can solve complex problems
    - when to use deep learning?
        - when there are a lot of data (if only have smaller dataset, can just use traditional machine learning)
        - access to processing power
        - lack of domain knowledge for understanding the features → because the neural network will help to figure out
        - complex problems (computer vision and natural language processing)
        - automatic feature extraction

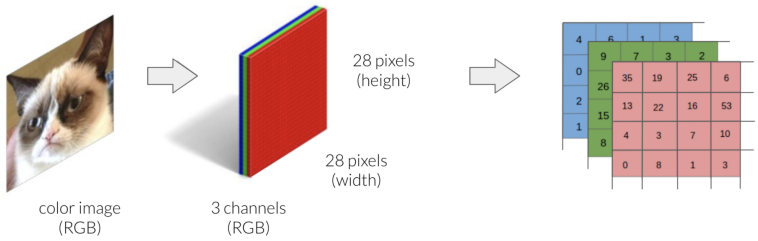
2. Neural network
    - basic unit of neural network is node (neuron)
    - map relationship between different combinations of variables to the desired output



3. Computer vision
    - help computers to see and understand the content of the digital images
    - can also generate a new things based on the dataset given (Deep fake → generate a face that did not exist)
    - Image data
        - image is made up of pixels (contains information about colour (stored in RGB system) and intensity(0 -255))

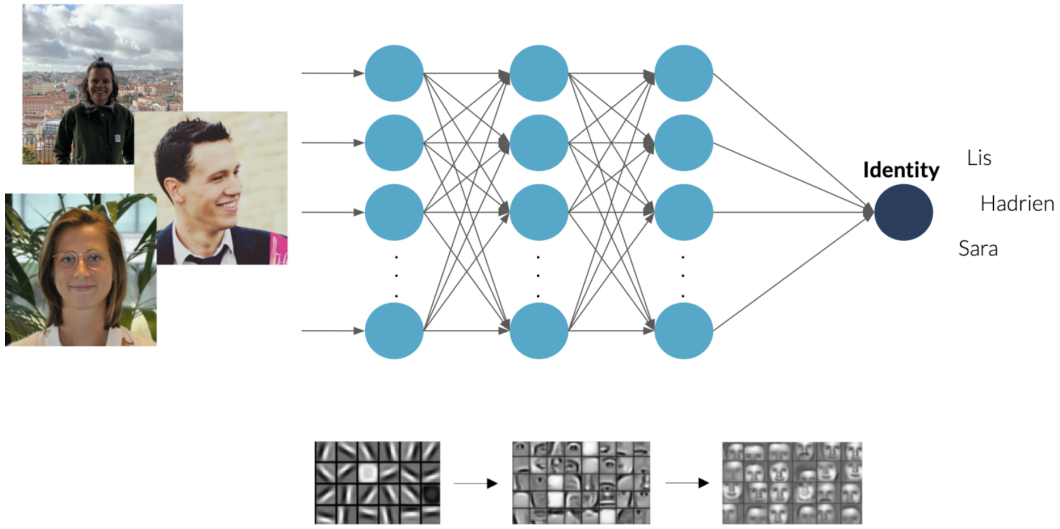


Grayscale image



Colour image

- Exp:



- applications of computer vision
  - facial recognition
  - self-driving vehicles
  - automatic detection of tumors in CT scans

4. Natural Language Processing
   - the ability of the computer to understand the meaning of the human language
   - techniques to identify the text data
     1. bag of words
        - to count the number of times important words appear in each piece of text

"U2 is a great band"

Word	Count
U2	1
Queen	0
is	1
a	1
great	1
band	1

"Queen is a great band"

Word	Count
U2	0
Queen	1
is	1
a	1
great	1
band	1

- types:
  1. bag of words: n-grams (able to capture more information)
     - to count the sequence of words

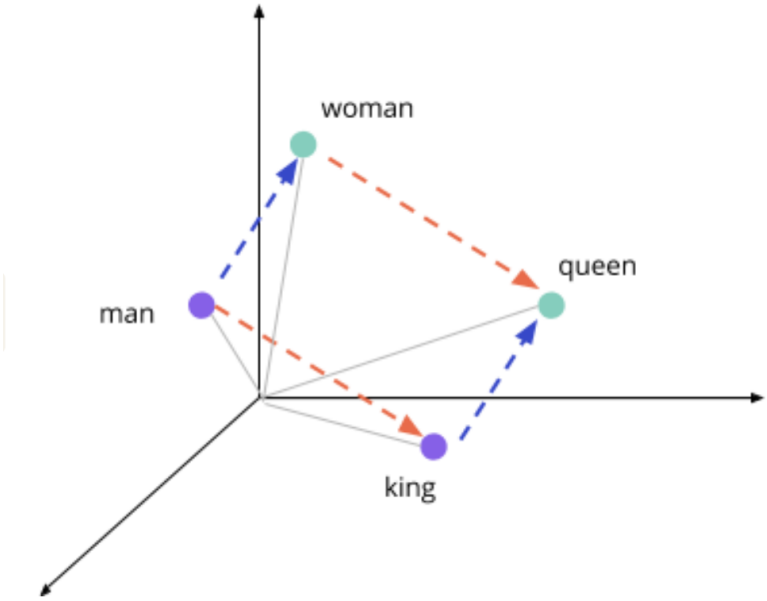
"That book is not great"

Word	Count
That	1
book	1
is	1
not	1
great	1

2-gram (bi-gram)

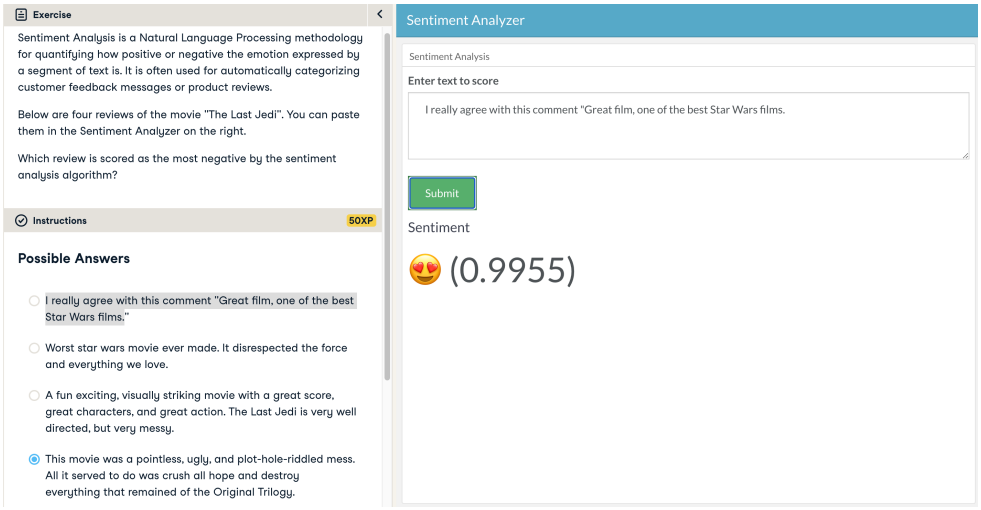
Word	Count
That book	1
book is	1
is not	1
not great	1

- limitations:
  - word counts do not help us to consider synonyms
    - Exp: blue → sky-blue, aqua, cerulean
  - want to group as a single feature
    - can be solved by word embeddings (create features that group similar words)
    - features have mathematical meaning:
      - Exp: king - man + woman = queen

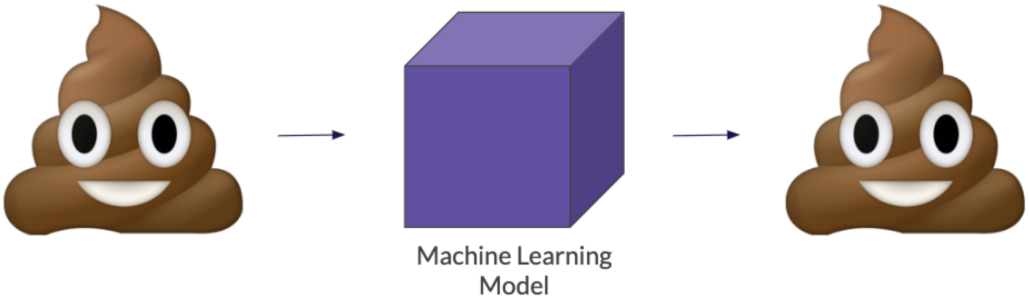


2. language translation
    - translate input language to different language

- applications of NLP
    - language translation
    - chatbot
    - personal assistants
    - sentiment analysis (used to quantify how positive or negative the emotion expressed by a segment of text)
        - Exp:



5. Limits of machine learning
    1. data quality



output quality depends on the inout quality

- data quality assurance
    - high-quality data requires:
        - data analysis (data characteristics, distribution, source and relevance)
        - review of outliers, exceptions or anything is suspicious
        - domain expertise from expert to explain unexpected data patterns
        - documentation

    2. explainability
        - normally machine learning is described as a black box
        - but sometimes the transparent box is needed in order to increase trust, clarity and understanding
        - the deep learning can make accurate decision but the human do not know how it can make the specific prediction

When we need black box or explainable AI?

1. Black box
    - recognise the letters
2. Explainable AI

- why the nurse want to quit from the hospital in a month

Black box

- Deep learning
- Better for "What?"
- Highly accurate predictions

Explainable AI

- Traditional machine learning
- Better for "Why?"
- Understandable by humans