# Supervised and Unsupervised Learning Fundamentals

## Week 6 – Session 2

BATH SPA UNIVERSITY

## Data Analytics and Machine Learning

# Overview

10:00 – 11:30

LECTURE

11:30 – 12:30

LAB SESSION

1:00 – 2:00

LECTURE

# Lesson Objectives

- To be able to understand the concept of supervised learning
- To be able to understand the concept of unsupervised learning
- To be able to compare and discuss about the differences between supervised and unsupervised learning
- To apply and visualise machine learning algorithms using scikit learn

| Keyword | Description |
|---|---|
| Training dataset | Data used to train a set of machine learning models. |
| Validation dataset | Data used for model selection and validation, e.g., to choose a model which is complex enough to describe the data ‘well’ but not more complex. |
| Supervised Learning | A type of machine learning where the model learns from labelled data, i.e., data where the target variable is known. |
| Testing dataset | Data not used when building the machine learning model but used to evaluate the model’s performance on previously unseen data (generalisation error). |
| Features | The covariates/predictors/inputs/attributes used to train the model. |
| Regression | A type of supervised learning task where the goal is to predict a continuous target variable. |
| Training error | The model’s performance evaluated on the training data (also known as in-sample or resubstitution error). |
| Testing error | The model’s performance evaluated on the testing data (also known as out-of-sample or generalisation error). |
| Classification | This is a type of supervised learning task where the goal is to predict a categorical target variable. |
| Unsupervised Learning | Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyse and cluster unlabelled datasets. |
| Clustering | Clustering or cluster analysis is a machine learning technique which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group." |
| Dimensionality reduction | Dimensionality reduction is a machine learning (ML) or statistical technique of reducing the number of random variables in a problem by obtaining a set of principal variables. |
| Hyperparameter | These are parameters of the learning algorithm itself, not derived from the data, that need to be set before training the model. |
| Neural Network | A set of algorithms modelled after the human brain, designed to recognize patterns. They interpret sensory data through a kind of machine perception, labelling or clustering raw input. |

# Machine Learning

# Machine Learning (ML)

- Facilitating **automation**
  - Improving efficiency in various sectors
- Enabling **predictive analysis**
  - For better decision-making
- Enhancing **user experience**
  - Through personalized recommendations
- ML becoming a pillar of our future civilisation
  - Increasing reliance on ML for technological advancements
    - Object recognition
    - Biomarker discovery in genomics
    - Navigation of autonomous vehicles
    - Fraud detection
    - And more

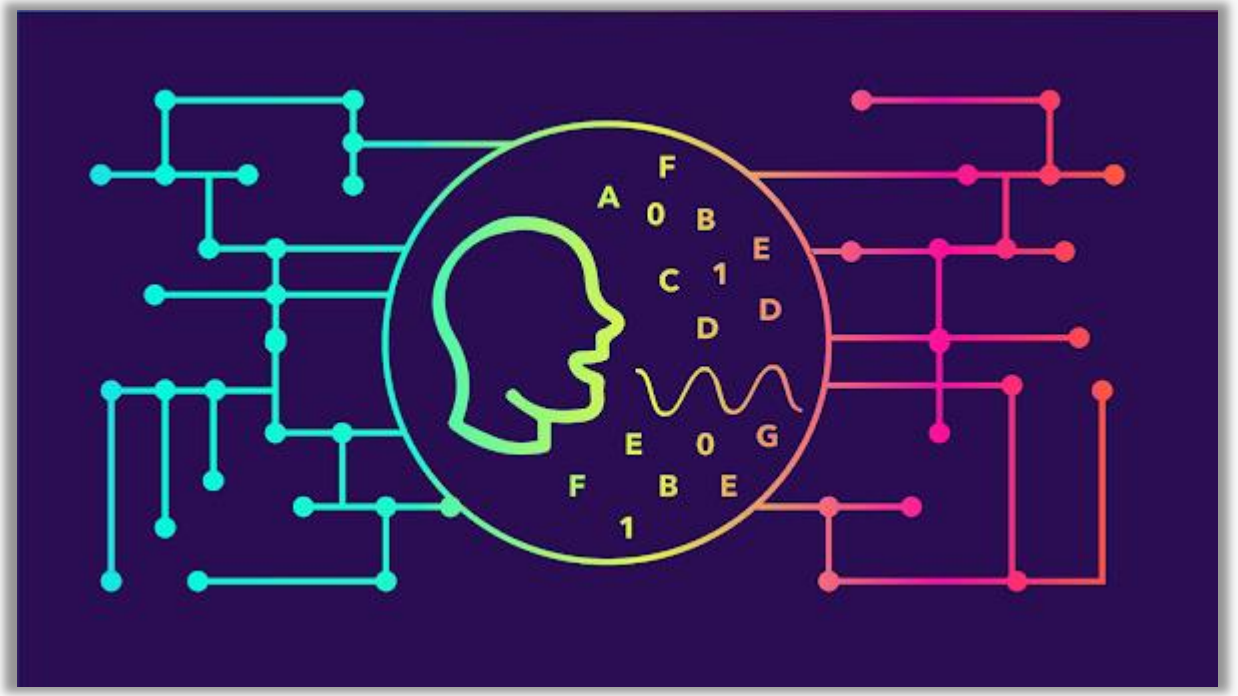
# Machine Learning Revolution

- Potential to revolutionise industries
  - E.g., healthcare, finance, and transportation

- ML's potential solutions in various domains:
  - **Data mining**
    - Extracting useful information
      - From very large datasets

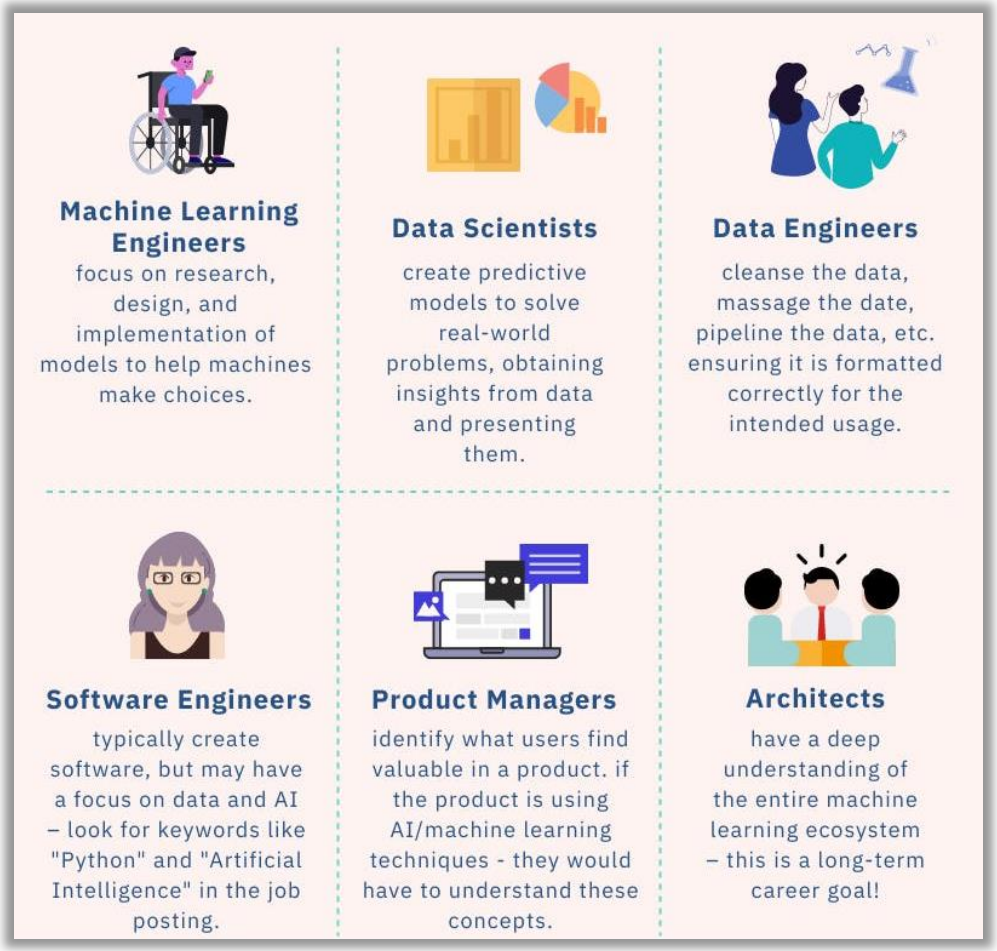
  - **Natural language processing**
    - Understand and respond to human language

  - **Image recognition**
    - Identifying and categorise images
      - Based on their features



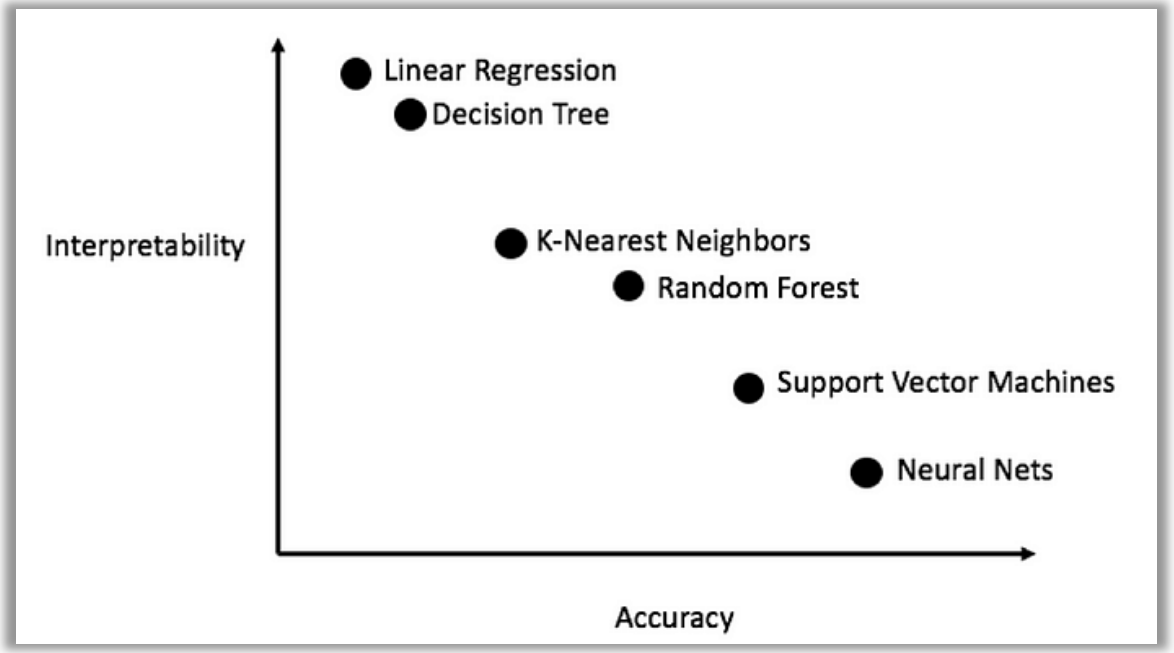
# Machine Learning – Supply and Demand

- Growing need for ML expertise
  - In various industries
  - The demand for expert ML designers is high
  - Increasing number of job opportunities in ML
- The supply of expert ML designers does not yet meet demand:
  - Limited number of professionals with advanced ML skills
- Major reasons
  - ML is tricky to master
    - Requires a good foundation in mathematics and programming
      - Mathematics mainly for ML Engineers
  - Involves understanding complex algorithms and models
    - Step by Step learning process



# Machine Learning – Algorithm Tradeoffs

- **Linear Regression**
  - High interpretability
  - Lower accuracy for complex datasets
- **Decision Trees**
  - Prone to overfitting, reducing accuracy
- **K-Nearest Neighbours (KNN)**
  - Accuracy can be affected by irrelevant features
- **Support Vector Machines (SVMs)**
  - Lower interpretability due to data transformation
  - Can be more accurate for certain datasets
- **Random Forests**
  - Reduced interpretability due to ensemble method
  - Typically, higher accuracy than individual trees or KNN
- **Neural Networks**
  - Least interpretable due to complex architectures
  - Can achieve high accuracy on complex tasks




[Interpreting Machine Learning Models]

# Further Learning – ML Algorithms and Statistics

- **Further learning opportunity: NOT A REQUIREMENT FOR THE COURSE!**

  - Beginner-Intermediate:
    - **Free:**
      - StatQuest:
        - https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw
  - Advanced:
    - **Free:**
      - Probabilistic Machine Learning: Advanced Topics
        - https://probml.github.io/pml-book/book2.html
        - Download ‘Draft pdf of the main book’
    - **Paid:**
      - The Elements of Statistical Learning:
        - https://hastie.su.domains/ElemStatLearn/
      - Machine Learning techniques in Python
        - Applied Machine Learning in Python Course (UMich) | Coursera

# Exploratory Data Analysis

## Machine Learning

BATH SPA UNIVERSITY

Data Analytics and Machine Learning

# Data Preparation

- Importance of cleaning, transforming, and preparing data
- Steps involved in data preparation for supervised learning

- **Categorise the problem Early**
  - Classification
  - Clustering
  - Regression
  - Ranking

- **Check Data Quality**
  - Format Data
    - Consistency
      - 5.93, $5.93, five ninety-three

- **Reduce Data**
  - Attribute Sampling
  - Record Sampling
  - Aggregation

- **Clean Data**
  - Missing Values
    - Substitution
      - Fill with zeroes, replace with mean, etc.
      - Categorical values
        - Could fill with frequent entries

- **Create New Features**
  - From existing ones
  - Opposite of reducing data
  - E.g., Creating a ‘day’ feature from a ‘date’ feature

- **Rescale Data**
  - Data Normalisation
    - Create a 0.0 – 1.0 range for values
      - Categorical data could be set as 0 or 1
      - Value could be set as a large value range 0-100,000
  - Without normalisation
    - Price could have a bigger weight
      - Than 0 – 1 categories

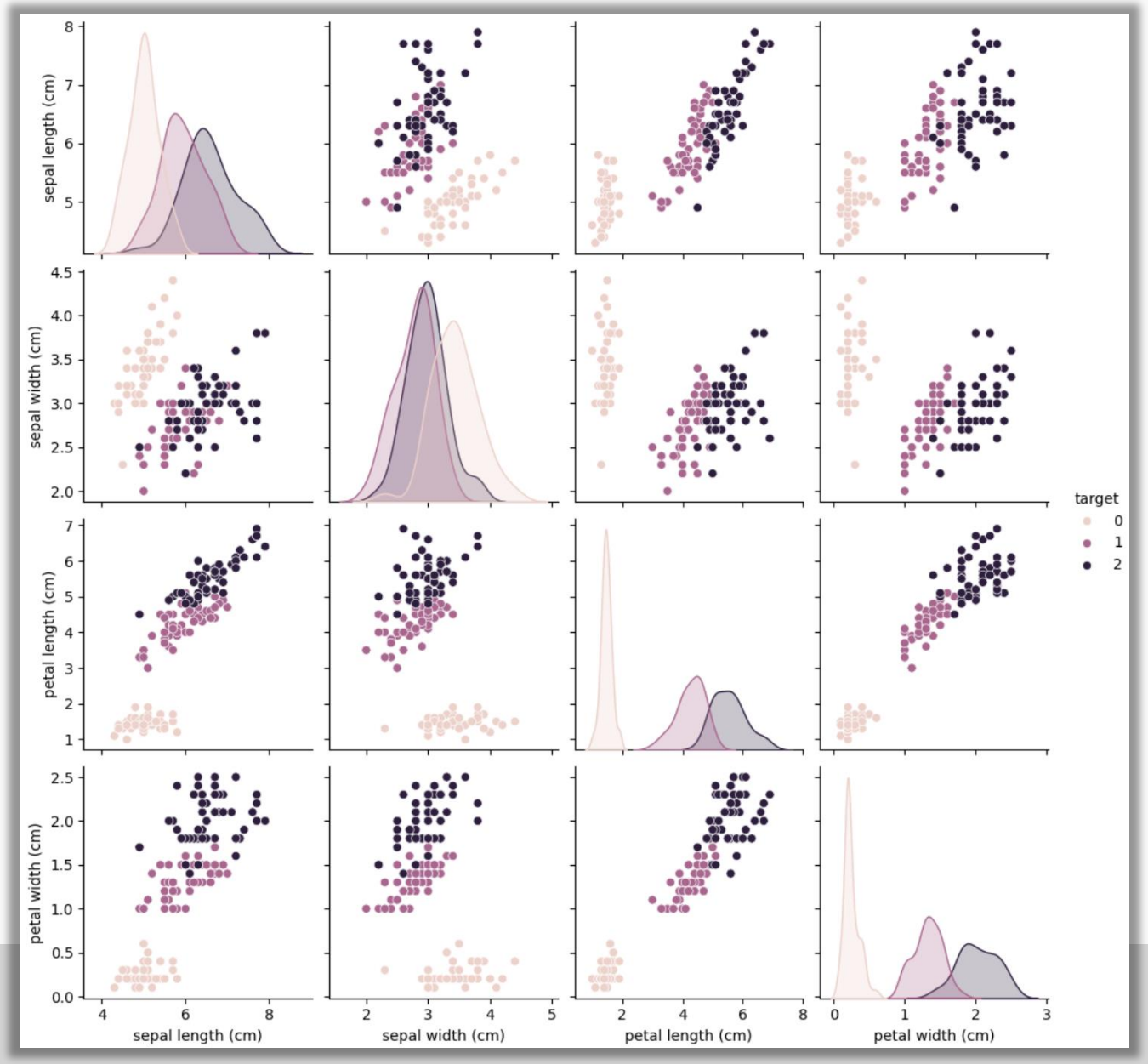
# Iris Dataset Demo

- Exploring the Iris dataset
  - https://www.kaggle.com/datasets/uciml/iris

- Type and shape
- Feature names
- Target names
- Array types

- Converting to a Pandas DataFrame

- Exploratory Data Analysis
  - With Visualisation



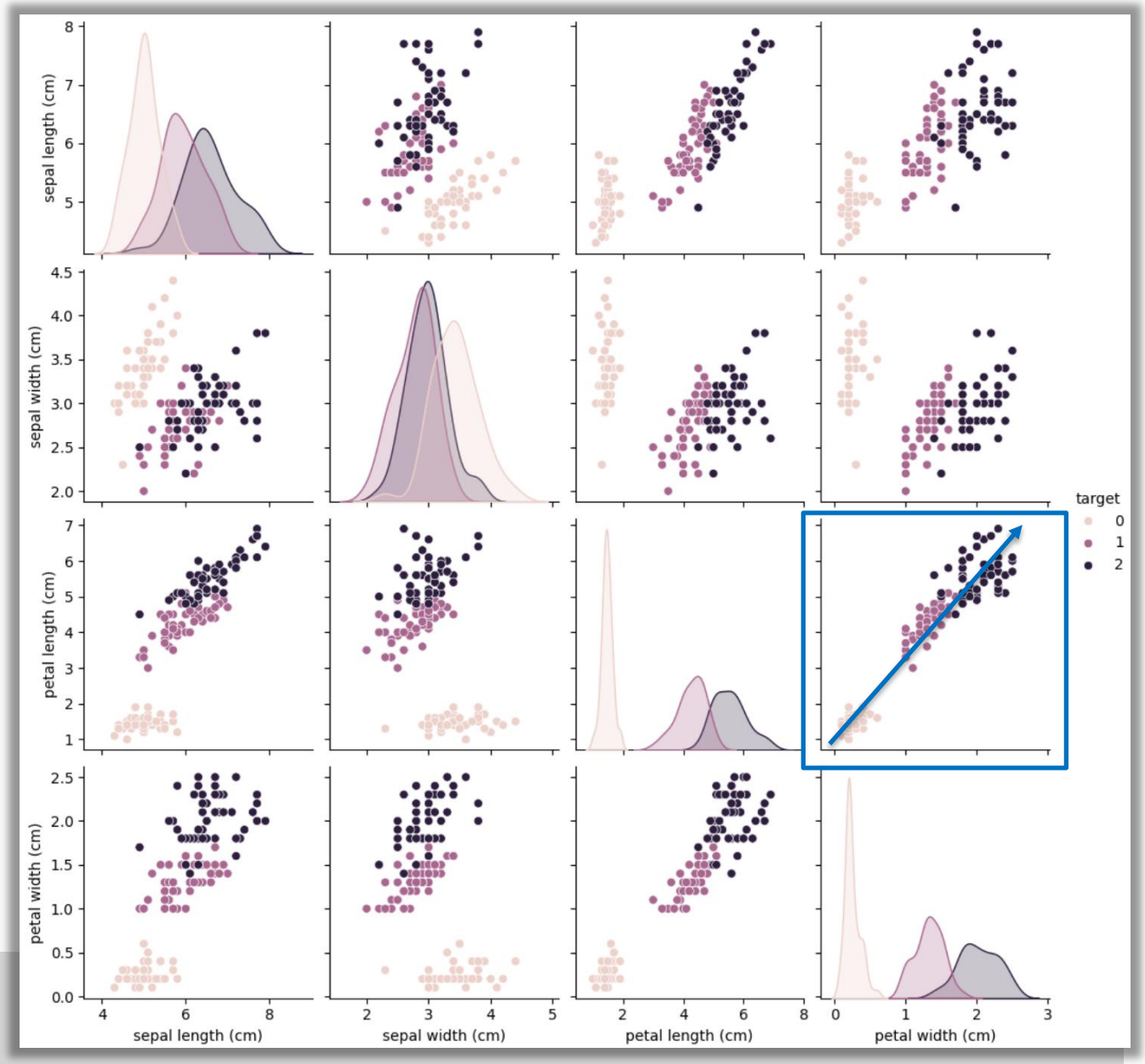
# Iris Dataset - Pairplot

- **Pairplot/Scatterplot Matrix**
  - Grid of scatter plots
  - Each feature plotted against every other feature

- **Reading a Pairplot**
- **Diagonal**
  - Shows distribution of a single variable
  - Usually a form of histogram

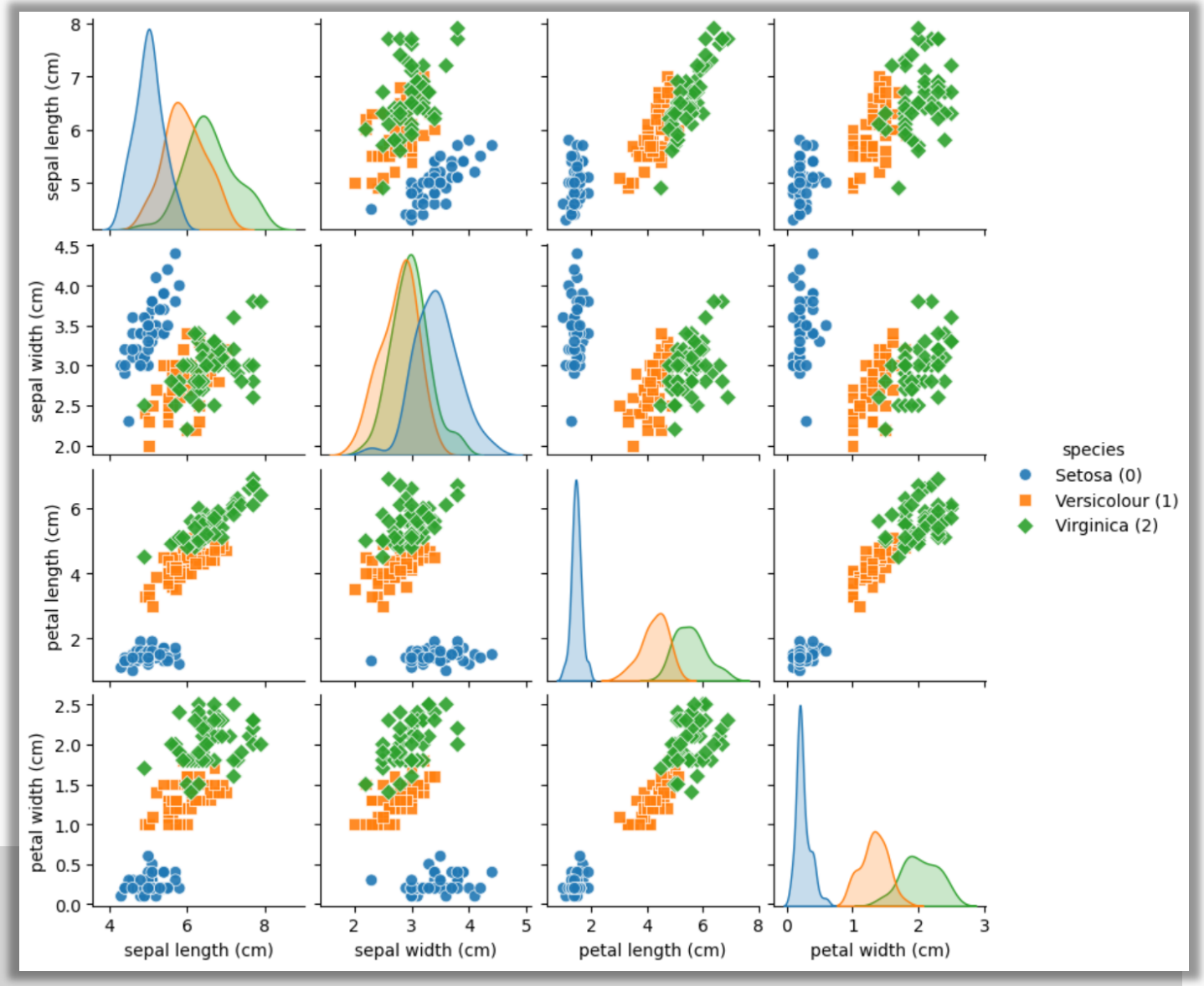
- **Off-diagonal**
  - Shows scatter plot between two variables
  - Rising from left to right
    - Suggests positive correlation
    - Petal Length and Petal Width

- **Insights from Pairplot**
  - Understand relationships between variables
  - Identify clusters and outliers
  - Discover trends and patterns



# Iris Dataset - Customising

- **Legend**
  - Descriptors can be customised
- **Color/Hue**
  - Based on one of the variables
  - Helps distinguish different classes
- **Marker shape and size**
  - Can represent different categories or properties



# Introduction to Supervised Learning

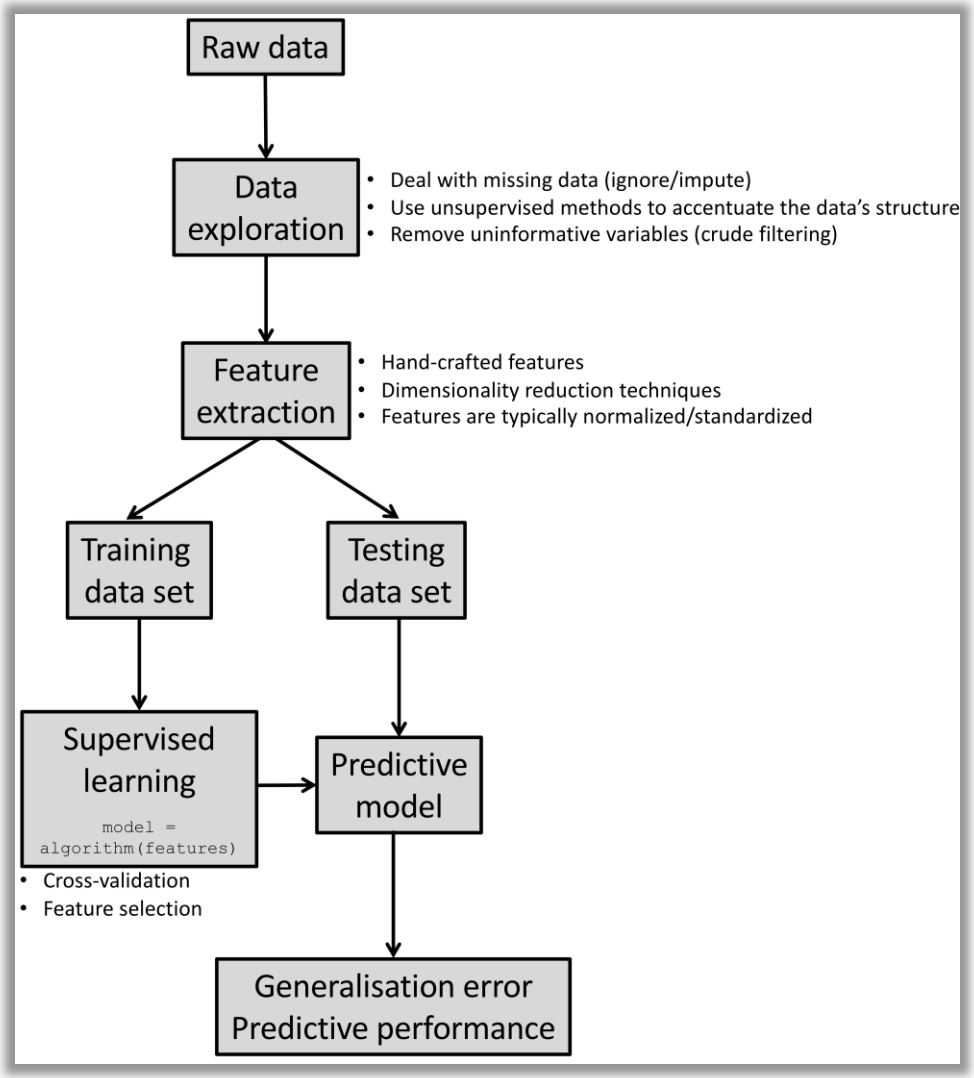
# Supervised Learning

- Like traditional statistical models
  - Generalised linear models

- **Discovers relationships**
  - Between
    - An outcome
    - A set of explanatory variables

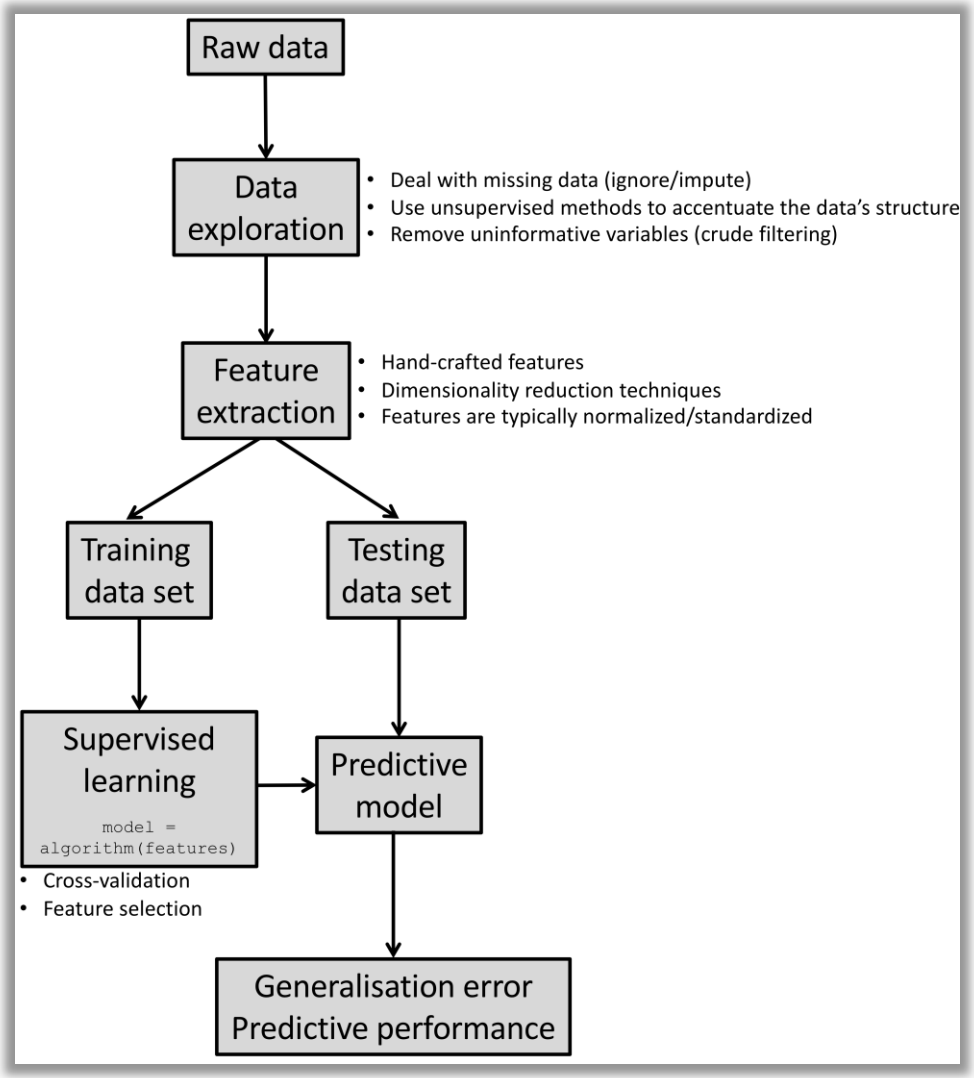
- **Training Data**
  - Model learns the mapping
    - Predictive model
    - Between
      - A set of features
      - An outcome using training data

- **Types of Outcomes**
  - Continuous Outcome
    - Regression Model
  - Categorical Variable
    - Classification Model



# Supervised Learning

- Model has both a known input and output used for training
- Knows the output during the training process
- Trains the model to reduce the error in prediction
- Two major types of supervised learning methods
  - Classification
  - Regression



# Supervised Learning

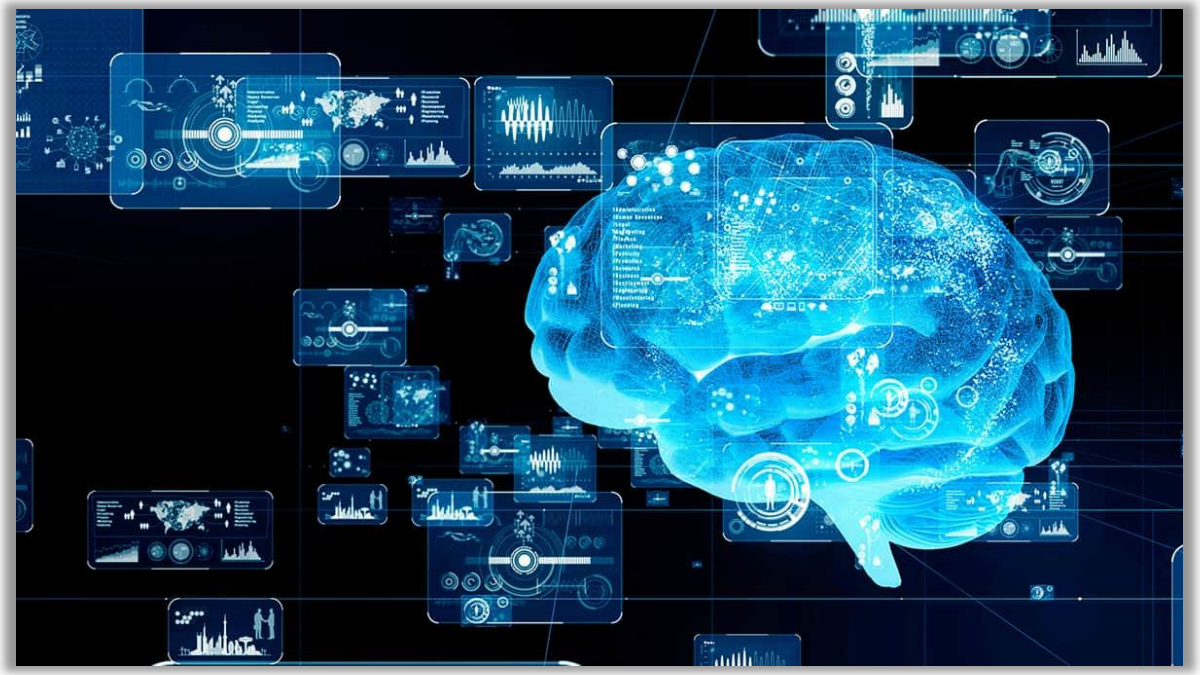
- Outcome measurement Y
  - Dependent variable, response, target
- Vector of p predictor measurements X
  - Inputs, regressors, covariates, features, independent variables
- In the regression problem
  - Y is quantitative
    - e.g., price, blood pressure
- In the classification problem
  - Y takes values in a finite, unordered set
    - e.g., survived/died, digit 0-9, cancer class of tissue sample
- Training data (x1, y1), . . . ,(xN , yN )
  - These are observations of these measurements
    - Examples, instances

**Objective of Supervised Learning**

- Accurately predict unseen test cases
- Understand which inputs affect the outcome
  - And how
- Assess the quality of our predictions
  - And inferences

# Applications of Supervised Learning

- **Medical Imaging**
  - Identifying a tumour as benign or cancerous
- **Gene Expression**
  - Determining a patient's phenotype based on their gene expression 'signature'
- **Computer Vision**
  - Detecting and tracking a moving object
- **Biogeography**
  - Predicting land cover usage using remote sensing imagery
- **Speech Recognition**
  - Translating audio signals into written text
- **Biometric Authentication**
  - Identifying a person using their fingerprint
- **Epidemiology**
  - Predicting the likelihood of an individual developing a particular disease
    - Given a number of risk factors



# Supervised – Teacher/Student Example

**1. Teacher's Role - Training the Model**

- Teacher guides a student through the learning process
- Provides student with problems
  - Input data
- And their solutions
  - Labels
- E.g., Teacher provides several math problems
  - Along with their solutions

**2. Student's Role - The Model**

- Student learns from these examples
  - Trying to understand the relationship between the problems and their solutions
- ML Comparison
  - A model learning to map input data to labels

**3. Testing the Student - Making Predictions**

- Test what the student has learned
- Teacher provides new problems without solutions
- Student applies what they've learned to these new problems
  - Tries to solve them
- ML Comparison
  - A trained machine learning model makes predictions on unseen data

**4. Grading the Student - Evaluating the Model**

- Teacher grades the student's solutions
- Provides feedback on what was right or wrong
- ML Comparison
  - Evaluating a machine learning model's predictions
    - Against true labels
    - Adjusting the model parameters accordingly

**5. Student learns from each iteration**

- Making fewer mistakes in more iterations

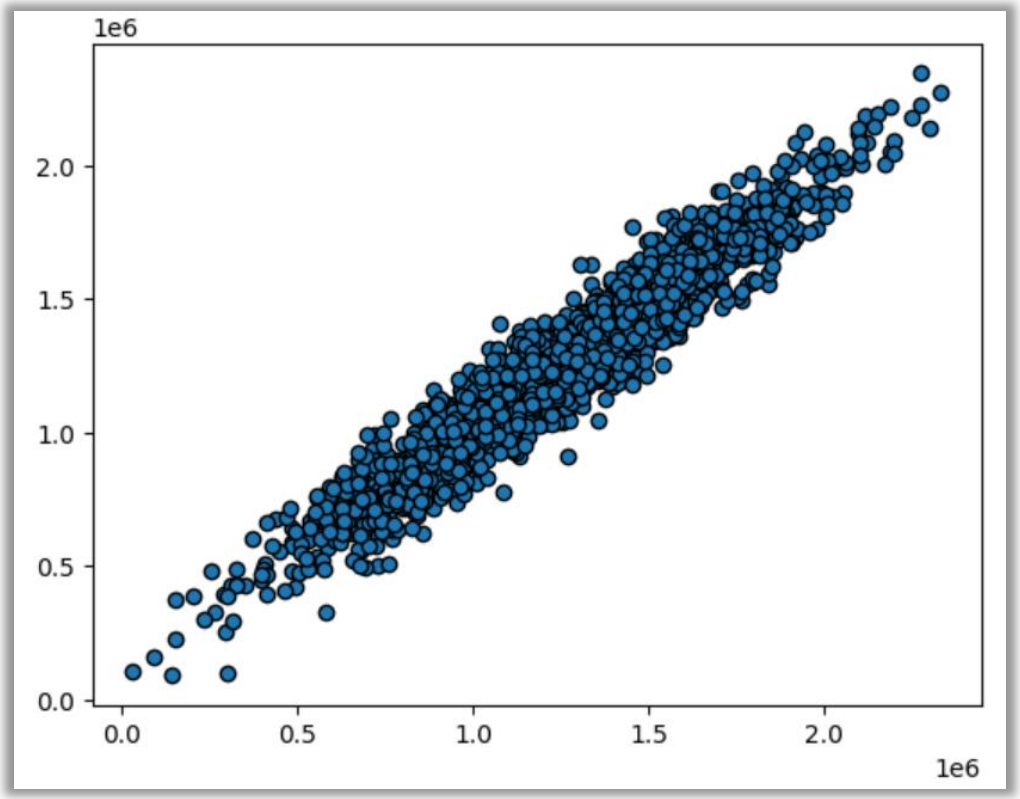
# Supervised Learning Algorithms
## Regression

BATH SPA UNIVERSITY

Data Analytics and Machine Learning

# Linear Regression

- Supervised learning method used for predicting
  - A continuous outcome variable (Y)
  - Based on one or more predictor variables (X)
- Assumes a linear relationship between the predictor variables and the outcome
- Coefficients in the model represent the change in the outcome
  - Associated with a one-unit change in the predictors
- Model fitting involves minimising the sum of the squared differences
  - Between the observed and predicted values of the outcome
- Performance is evaluated using measures
  - Mean Absolute Error, Mean Squared Error and Root Mean Squared Error (RMSE)
- Assumptions include linearity, independence, and normality of residuals




Left to right diagonal rise suggests good prediction model results

BATH SPA UNIVERSITY

More Information: Lecture Notebook

# Logistic Regression

- A statistical model used for binary classification problems
- Probability Estimation
  - Estimates the probability of an event occurrence
    - Based on one or more predictor variables
- Sigmoid Function
  - Uses logistic sigmoid function
    - To return a probability value between 0 and 1
- Binary Outcome
  - Output is binary
    - Predicts one of two possible outcomes
- Coefficient Estimation
  - Coefficients of the logistic regression algorithm
    - Must be estimated from your training data
- Applications
  - Healthcare, finance, social sciences
  - For risk assessment and prediction

# K-Nearest Neighbours (KNN)

- K-NN algorithm compares a new data entry with existing data entries

- Assigns the new data to a class
  - Based on closeness to neighbours

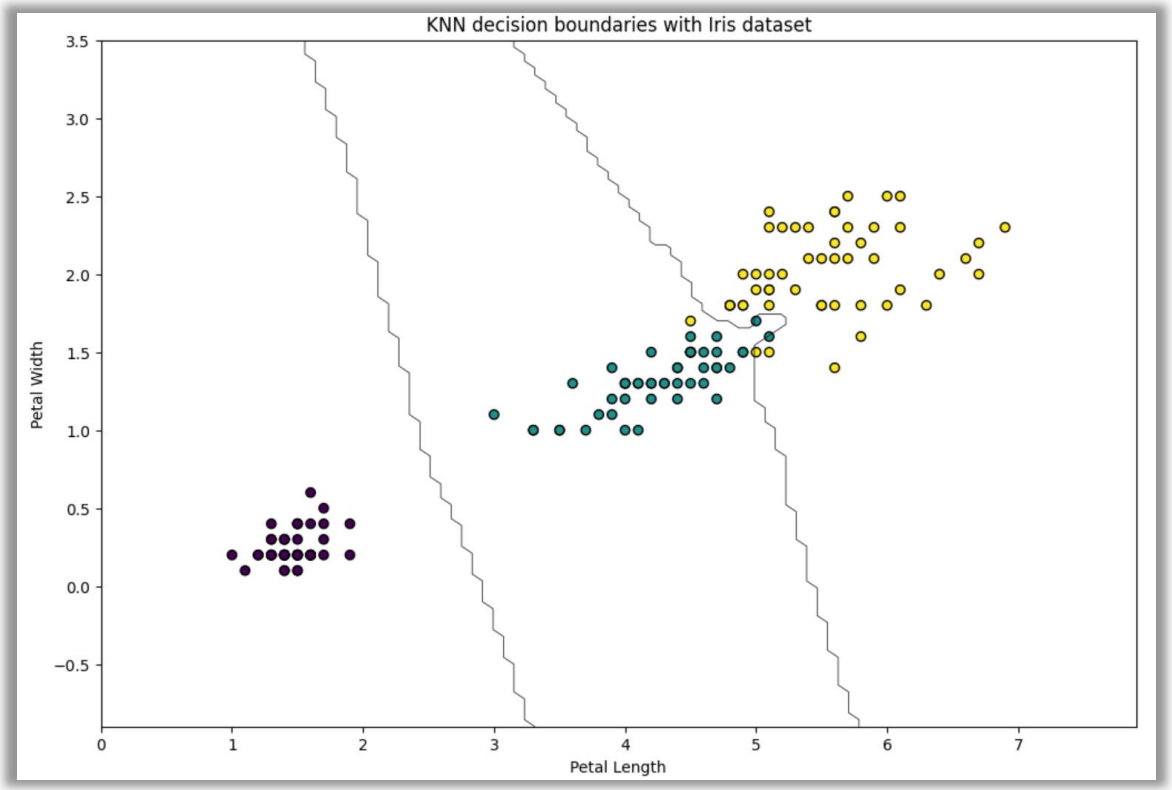
**Steps:**

1. Assign a value to K
   - K is the range of neighbours considered

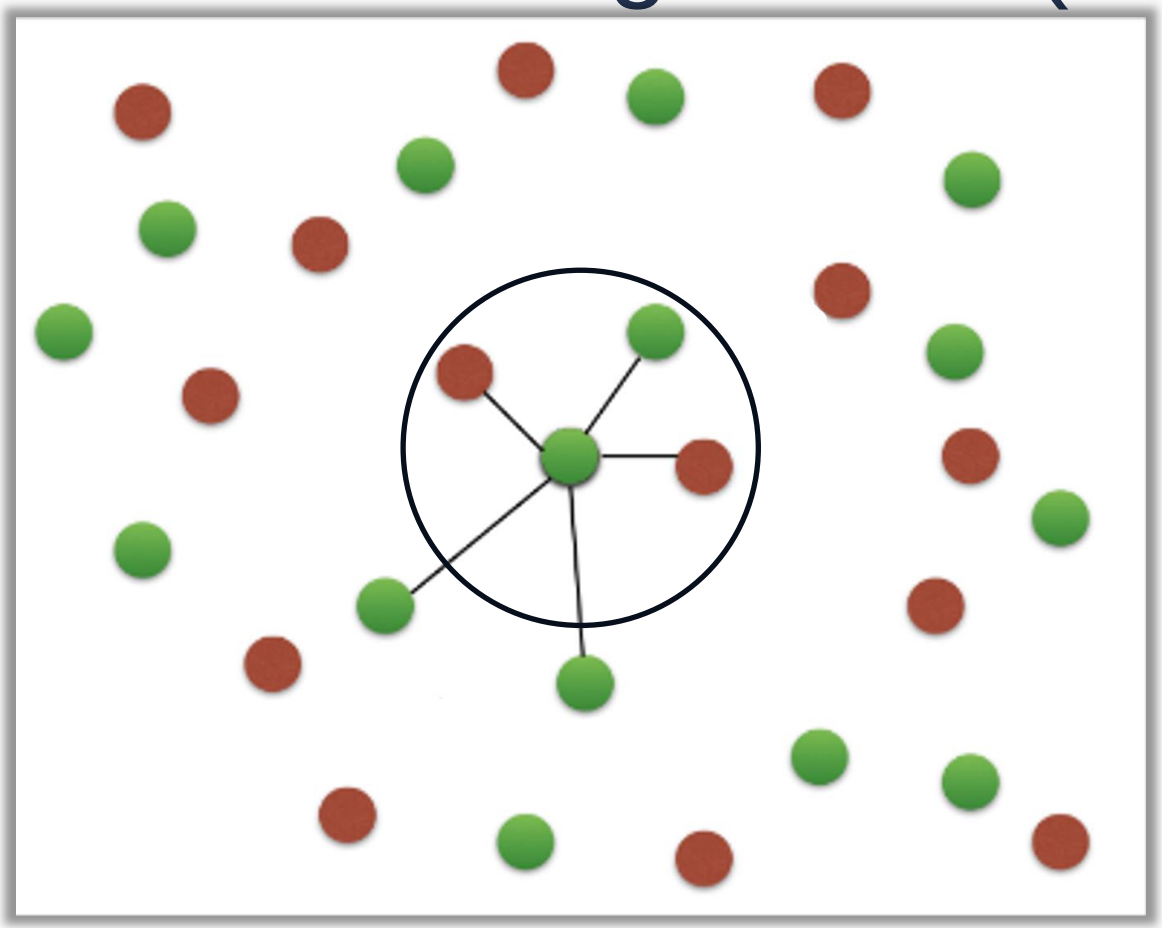
2. Calculate the distance between the new data and all other data.
   - Arrange distances in ascending order

3. Identify the K nearest neighbours to the new entry
   1. Based on the calculated distances

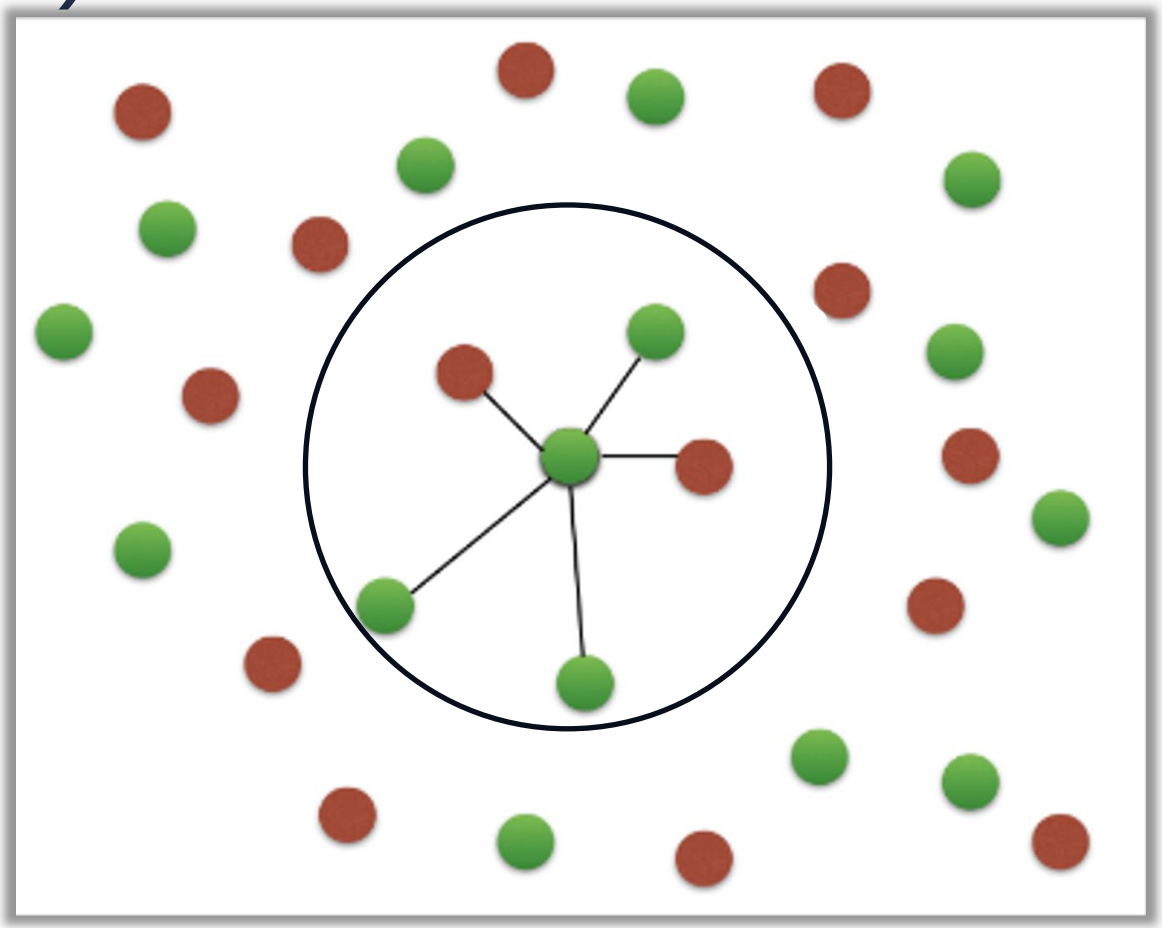
4. Assign the new data to the majority class
   - Among the nearest neighbours
   - The new entry is categorized based on its closest neighbours



# K-Nearest Neighbours (KNN)



K = 3
Classification: Red



K = 5
**What will the classification be?**

More Information: https://www.freecodecamp.org/news/k-nearest-neighbors-algorithm-classifiers-and-model-example/

# Importance of Scaling Data

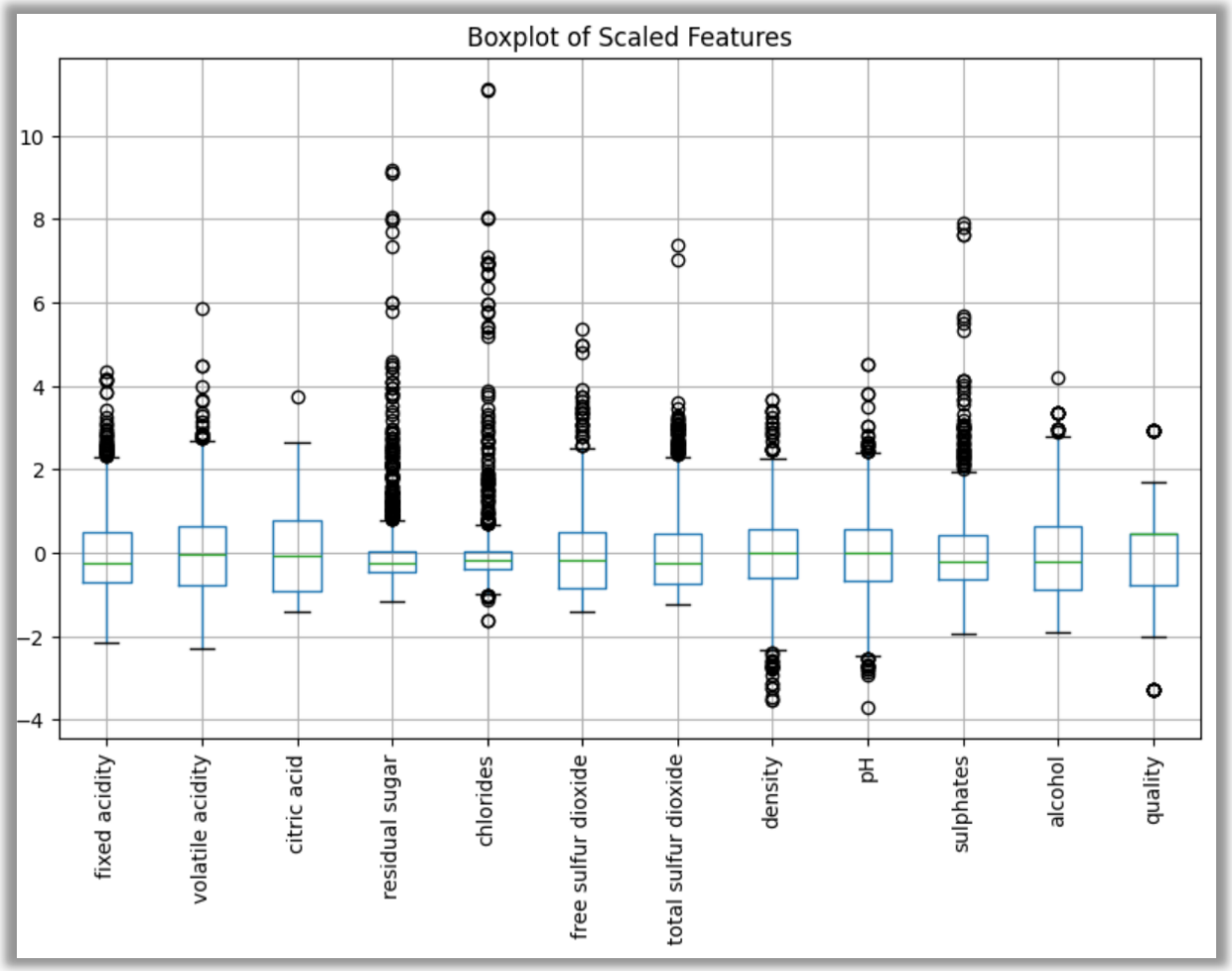
- Features in datasets can have varying ranges

- Example
- Red wine quality dataset
  - Density
    - varies from 0.99 to 1
  - total sulfur dioxide
    - from 6 to 289

- Machine learning models use distance to inform them

- Features on larger scales can unduly influence your model
  - such as K-nearest neighbours

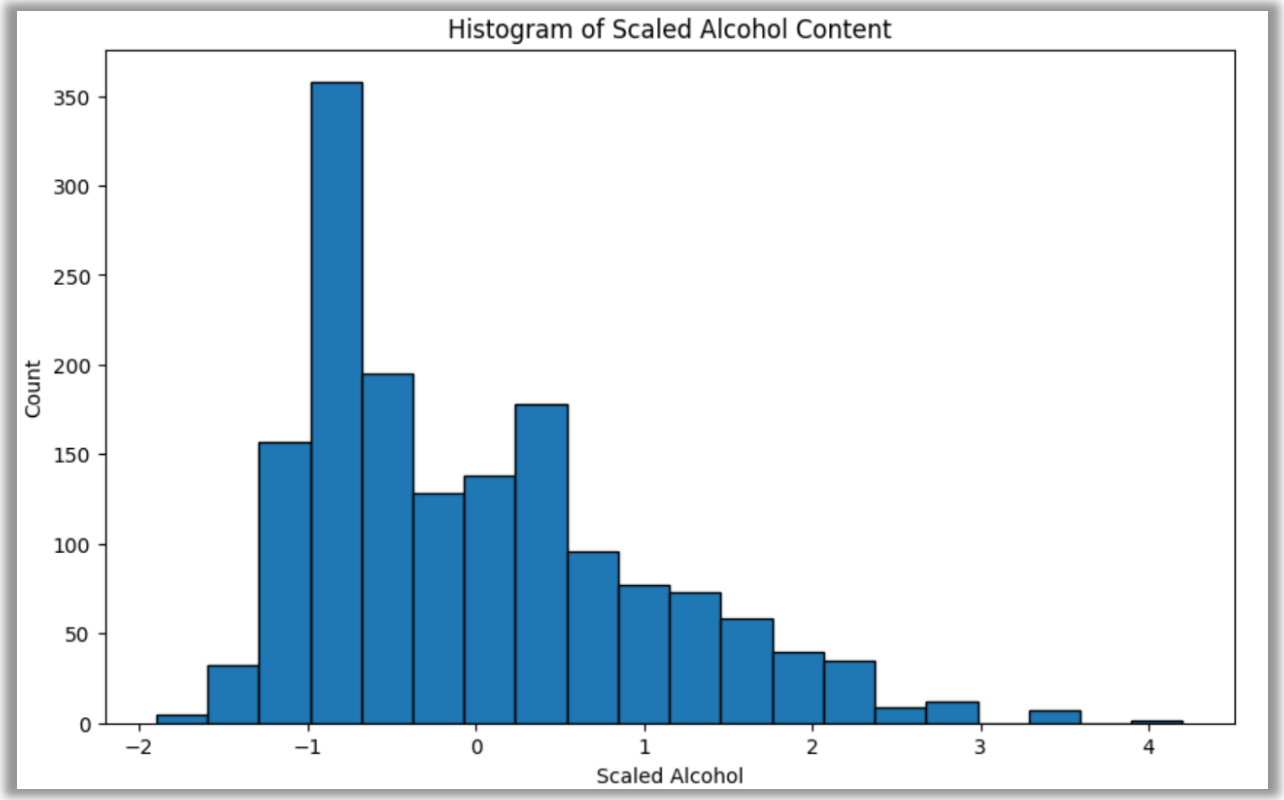
- Image Example: A histogram showing the distribution of the scaled ‘alcohol’ feature




More Information: Lecture Notebook

# Methods of Normalising Data

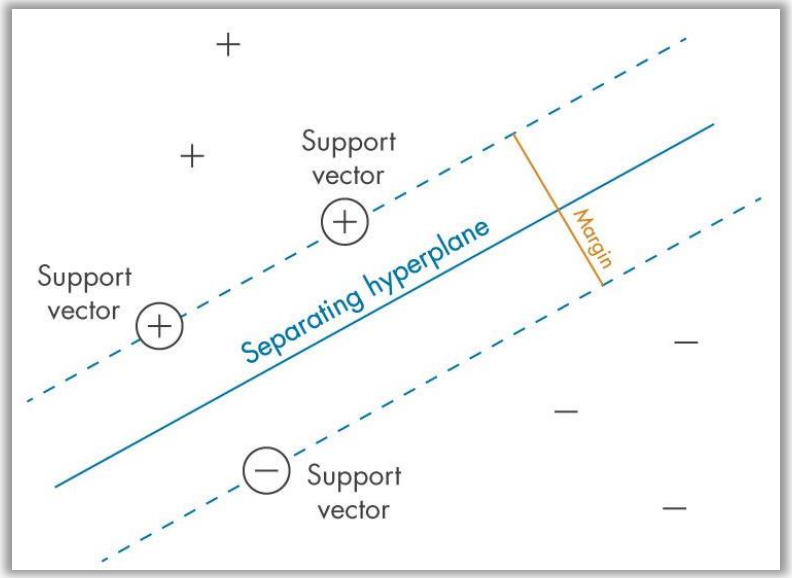
- Standardisation
  - Subtract the mean and divide by the variance
    - Features are centred around zero
      - Have variance one
- Min-Max Scaling
  - Subtract the minimum and divide by the range of the data
    - Normalised dataset has minimum zero
      - And maximum one
- Scaling to a range
  - Normalise so that data ranges from -1 to 1
- Verification
  - Check the mean and standard deviation
    - Both the original and scaled data
      - In their columns

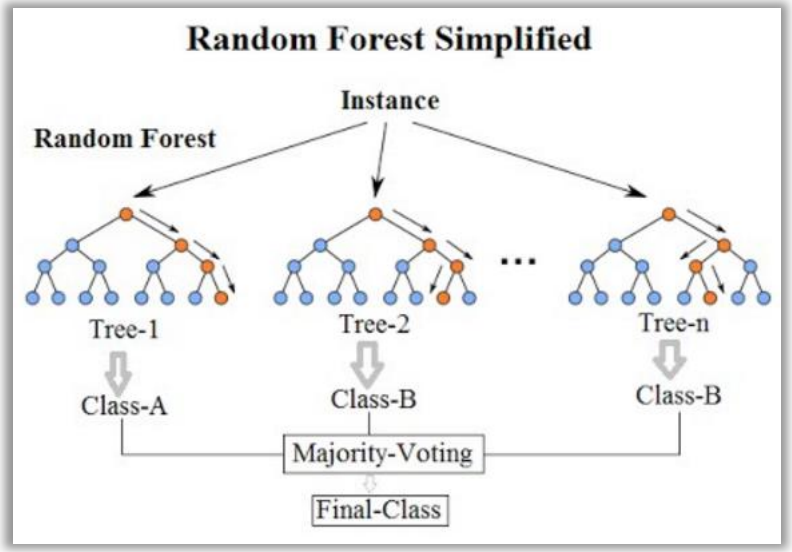


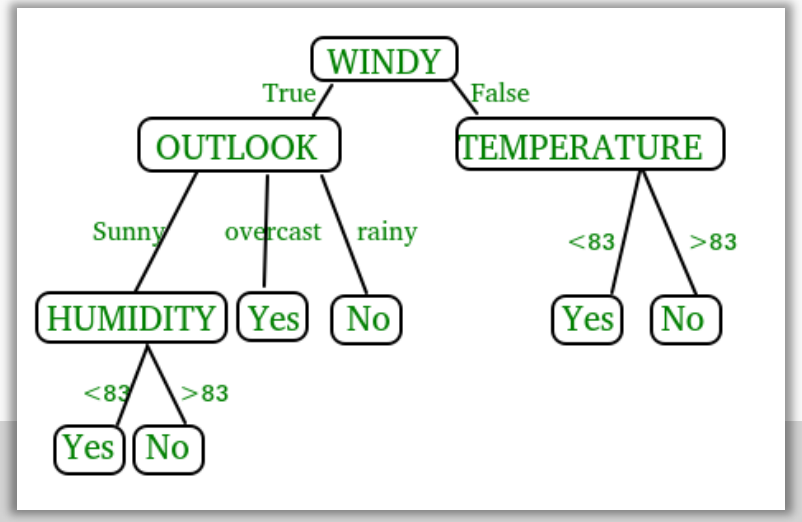

More Information: Lecture Notebook

# Other Classification Techniques

- SVM
- Decision Trees
- Random Forests
- Naïve Bayes




More Information: Lecture Notebook

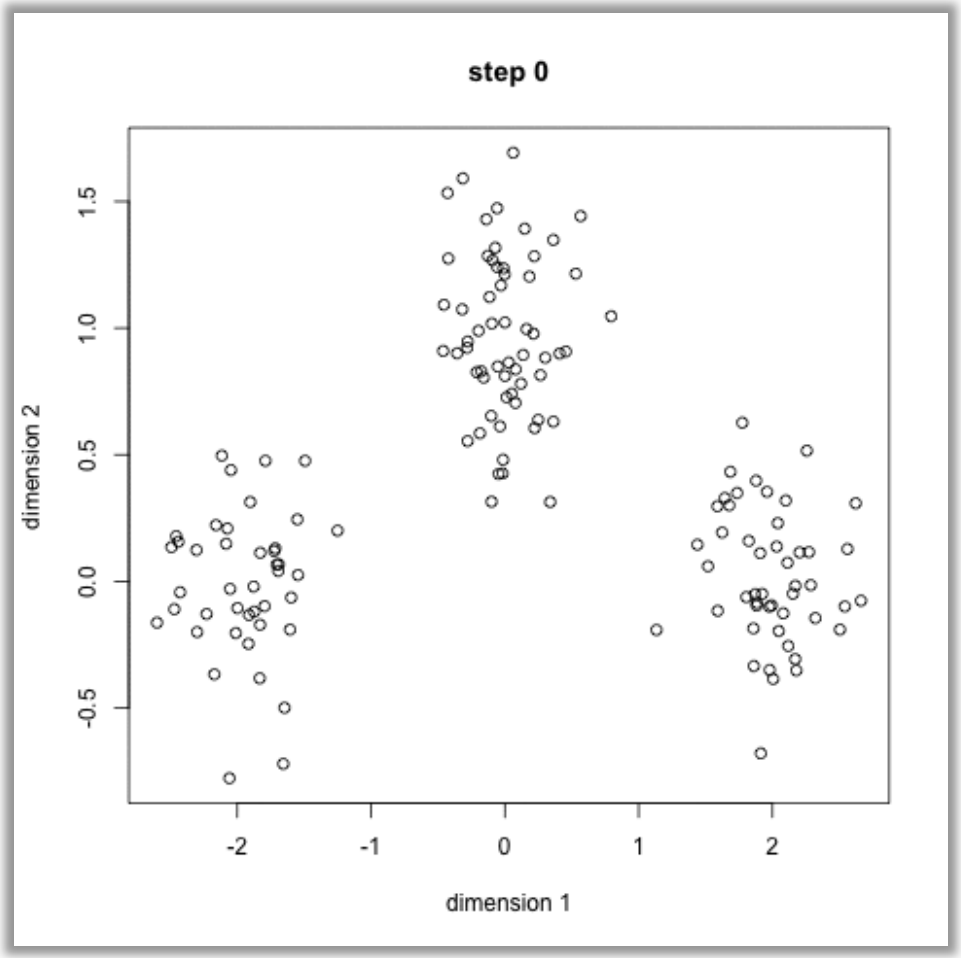
# Introduction to Unsupervised Learning

# Unsupervised Learning

- Machine learning techniques for discovering patterns in data

- **Clustering**
  - Finding natural clusters of customers based on purchase histories

- **Dimension Reduction**
  - Searching for patterns and correlations among purchases
  - Using these patterns to express data in a compressed form



# Supervised vs Unsupervised Learning

- **Supervised Learning**
  - Defined with labels
  - Pattern discovery is guided for predicting the label
  - Primarily used for
    - Classification
    - Regression

- **Unsupervised Learning**
  - Defined without labels
  - Pure pattern discovery
  - Unguided by a prediction task
  - Primarily used for:
    - Clustering
      - Unlabelled data
    - Dimension reduction

# Clustering

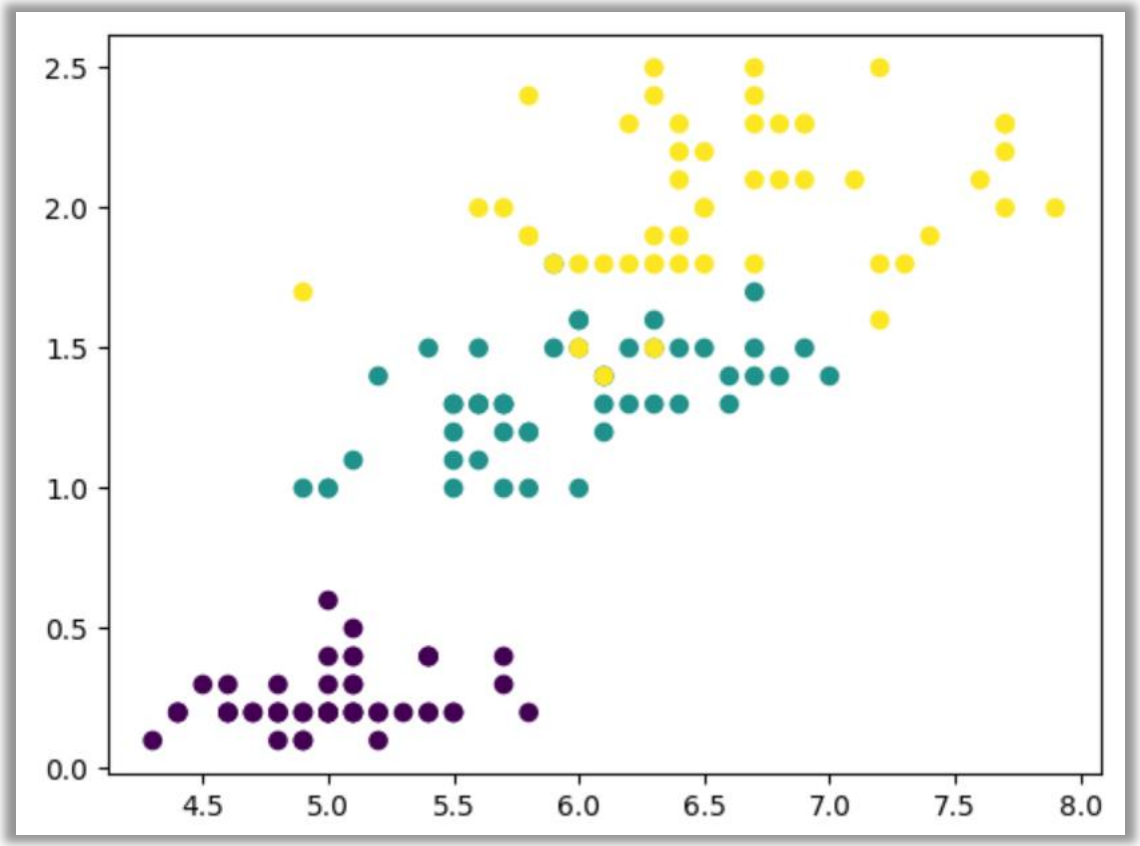
- Technique used in data mining and machine learning
- Groups similar objects into clusters
- Finds a specified number of clusters in the samples

- **K-means Clustering**
- Widely used method for cluster analysis

- Aims to partition a set of objects into K clusters

- Minimises the sum of the squared distances between the objects
  - And their assigned cluster mean

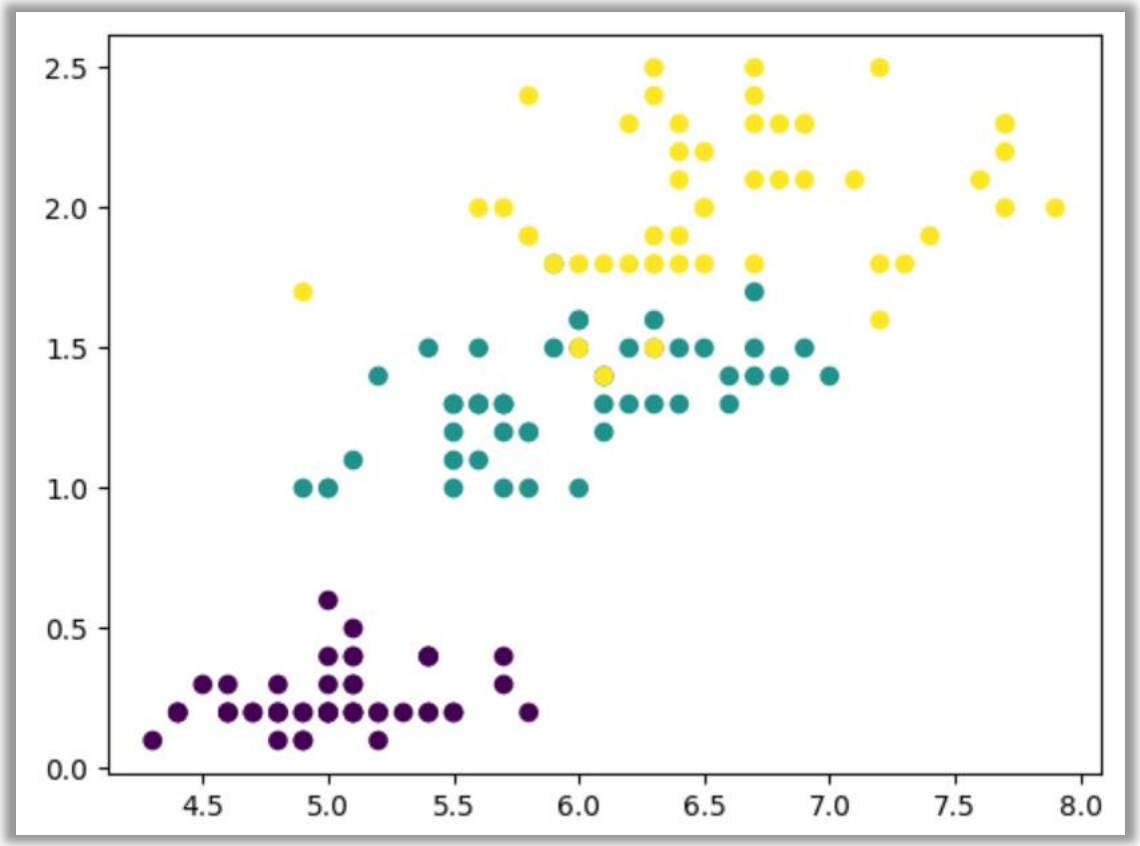
- Implemented in a library
  - scikit-learn or “sklearn”



# K-Means Clustering

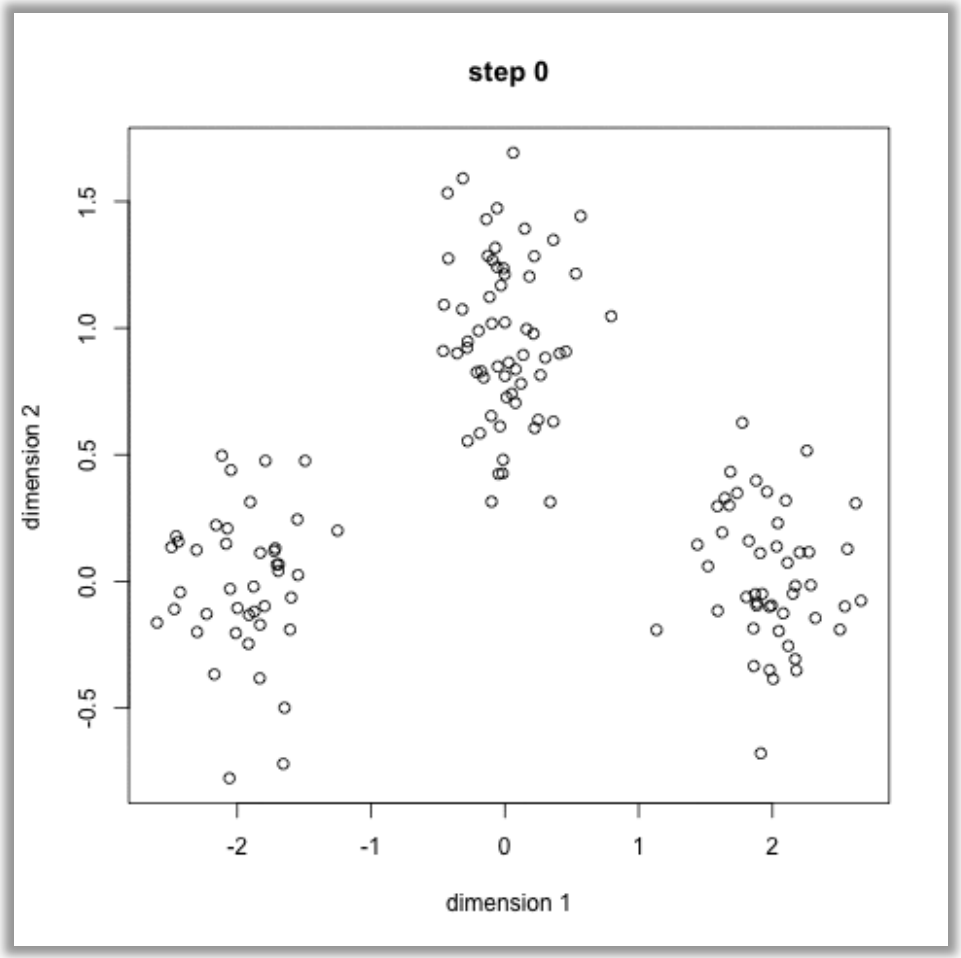
- **Similarity within Clusters**
  - Data points in a cluster
    - Should be similar to each other
    - Helps in targeted marketing
    - Impacts business strategy
- **Dissimilarity between Clusters**
  - Data points from different clusters
    - Should be as different as possible
  - Ensures meaningful clusters

- **K-Means Algorithm**
  - Uses an iterative approach
  - Minimises the sum of squared distances between data points
    - And their assigned cluster centroid



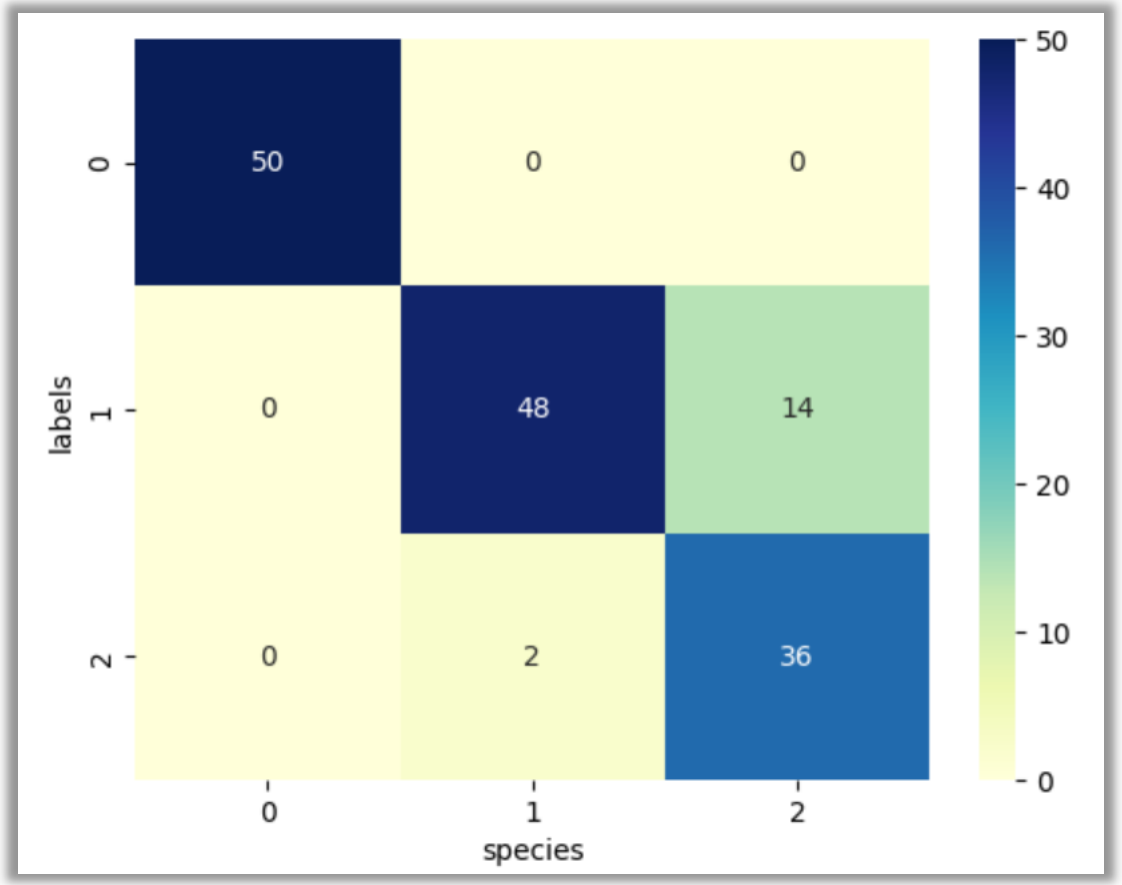
# Applications of Clustering

- **Customer Segmentation**
  - Various domains
    - Banking, telecom, e-commerce, sports, advertising, sales, etc
- **Document Clustering**
  - Groups similar documents together
- **Image Segmentation**
  - Groups similar pixels in the image together
  - Applied to create clusters
    - Having similar pixels in the same group



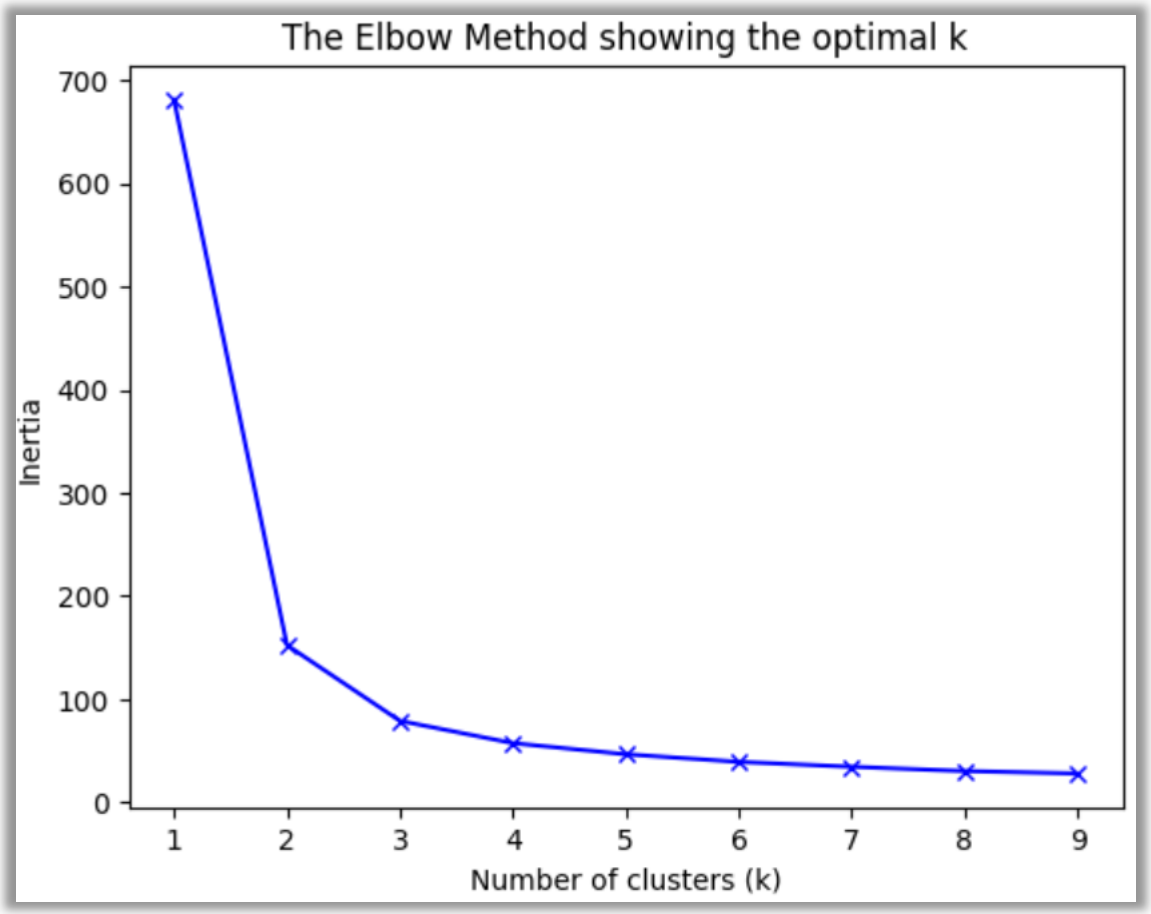
# Evaluating Clusters

- **Direct Approach**
  - Compare clusters with iris species
- **Quality Measure**
  - Measure quality of clustering
    - Without pre-grouped species
  - Used to make an informed choice
    - About the number of clusters
- **Cross-tabulation with Pandas**
  - Construct a table showing number of samples
    - For each cluster label/species combination
  - Provides insights into which samples are in which cluster



# Measuring Cluster Quality

- **Inertia**
  - Measures how far samples are from their centroids
  - Lower inertia values are better
    - Indicates tight clusters
- **Number of Clusters**
  - Trade-off between low inertia (tight clusters)
    - And not having too many clusters
  - Choose an elbow in the inertia plot
    - Inertia begins to decrease more slowly



# Exploratory Task: Machine Learning

1. **Create a Jupyter Notebook**
   - Create a file named 06-2 Task Workbook.ipynb

2. **Choose a dataset**
   - Suggestions: https://archive.ics.uci.edu/ml/datasets.php https://www.kaggle.com/

3. **Apply Machine learning algorithms**
   - This will include:
     - Obtaining a Dataset
     - Exploring the data
     - Preparing and Cleaning
     - Shaping the data to match your chosen algorithm
     - **Must contain at least 4 visualisations**
     - **Must contain explanation comments through the code**
     - **Must contain a conclusion on findings**

# Session Review

- Able to understand the concept of supervised learning
- Able to understand the concept of unsupervised learning
- Able to compare and discuss about the differences between supervised and unsupervised learning
- Can apply and visualise machine learning algorithms using scikit learn