EXPLORE AI ACADEMY

Advanced Classification Kick-Off

2304PTDS

November 2023

1 / **Objectives**

2 / **Recap: What is Machine Learning?**

3 / **Recap: Regression?**

4 / **What is Classification?**

5 / **What will we be seeing again?**

6 / **Natural Language Processing**

7 / **Sprint overview**

8 / **Predict overview**

Objectives

By the end of this session you should be able to:

- Recall key machine learning concepts
- Explain the concept of classification
- Distinguish between a regression problem and a classification problem

1 / **Objectives**

2 / **Recap: What is Machine Learning?**

3 / **Recap: Regression?**

4 / **What is Classification?**

5 / **What will we be seeing again?**

6 / **Natural Language Processing**

7 / **Sprint overview**

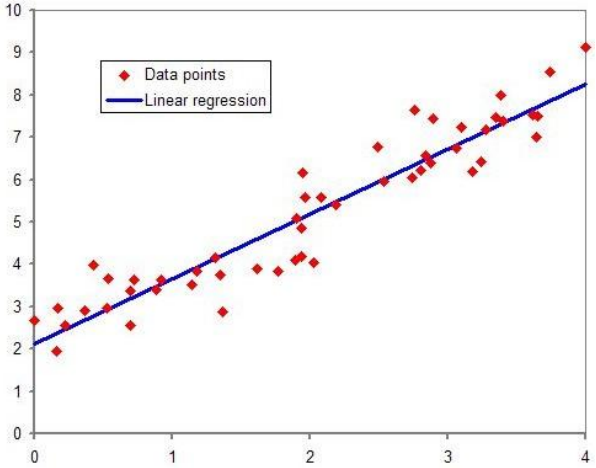
8 / **Predict overview**

What is Machine Learning?

Umbrella term for finding patterns amidst noise.

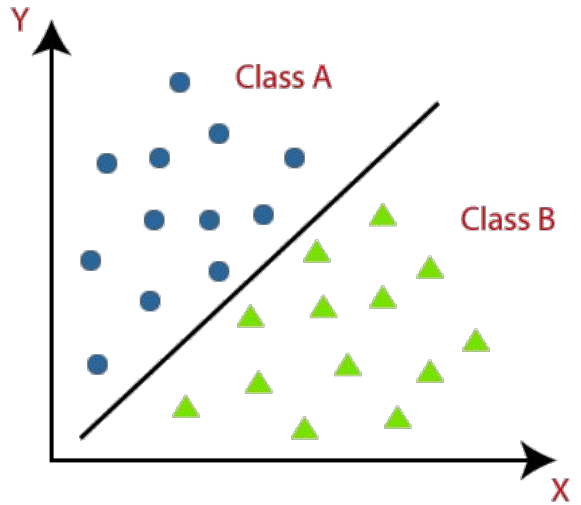
Underlying assumption:

- A process has an underlying pattern that relates some aspect of the process to another
- In the data there may be deviations from this pattern - called **noise**

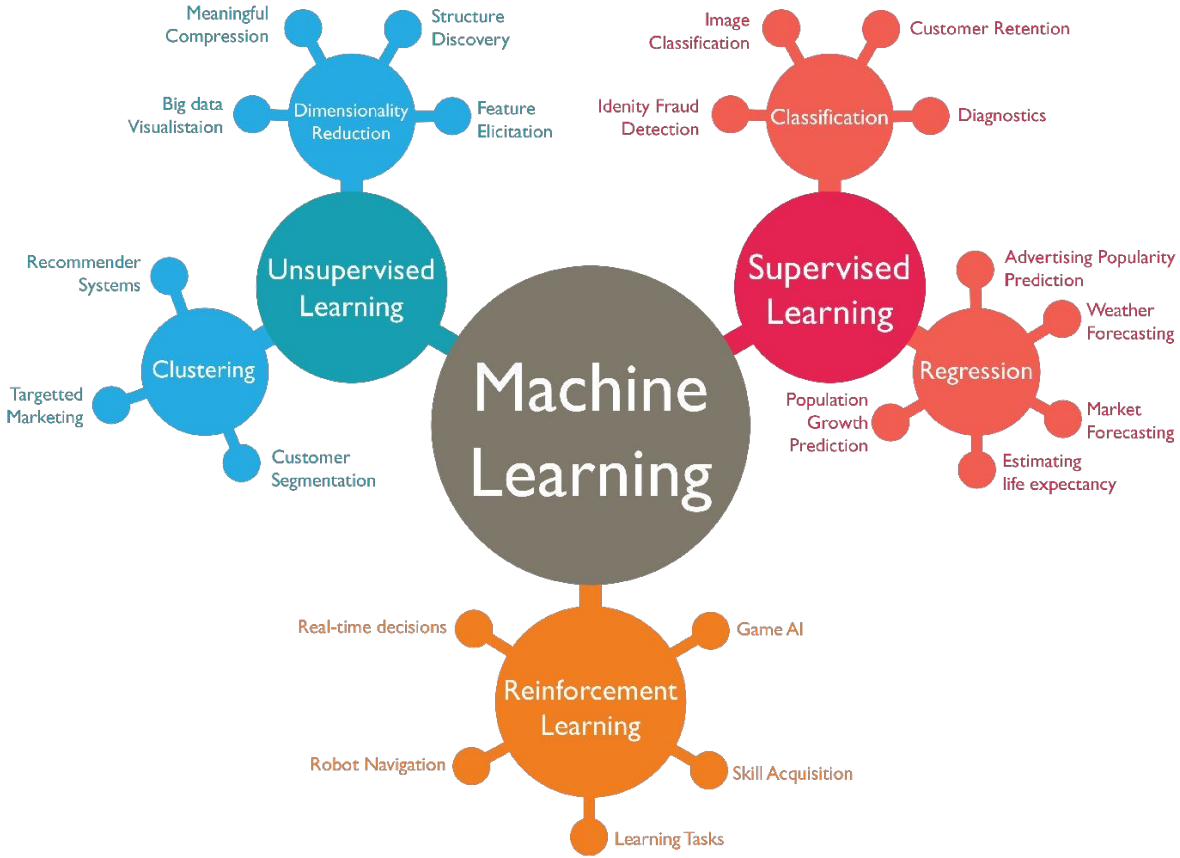


Example:

- Information on insurance clients
- Age, smoking status, drinking status, income - **input variables**
- Want to find a relationship between these factors and individual's risk category (low/medium/high)
- Risk category - **output variable**



The goal of machine learning is to find this pattern.



1 / **Objectives**

2 / **Recap: What is Machine Learning?**

3 / **Recap: Regression?**

4 / **What is Classification?**

5 / **What will we be seeing again?**

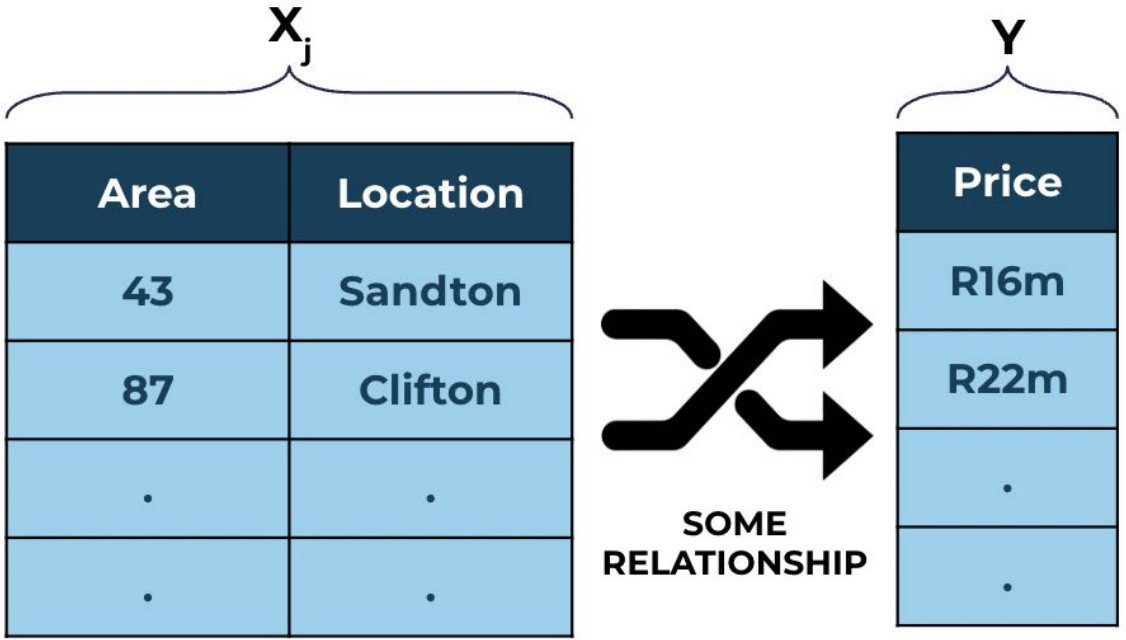
6 / **Natural Language Processing**

7 / **Sprint overview**

8 / **Predict overview**

What is Regression?

Predicting a **number** from predictor variables



1 / **Objectives**

2 / **Recap: What is Machine Learning?**

3 / **Recap: Regression?**

4 / **What is Classification?**

5 / **What will we be seeing again?**

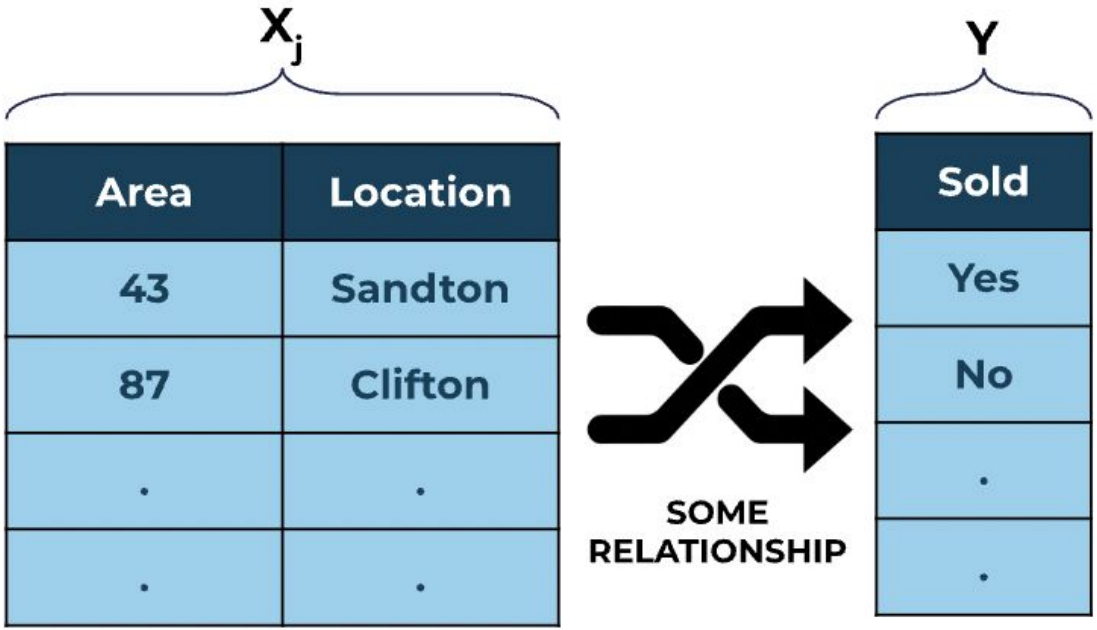
6 / **Natural Language Processing**

7 / **Sprint overview**

8 / **Predict overview**

What is Classification?

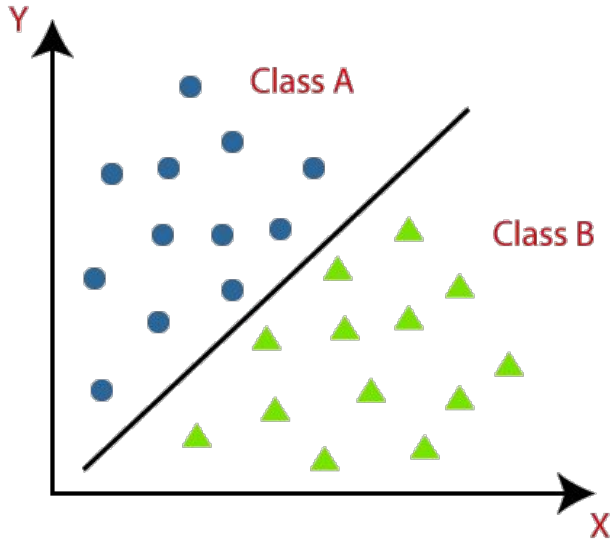
Predicting a **category/discrete class** from predictor variables



What is Classification?

Predicting a **category/discrete class** from predictor variables

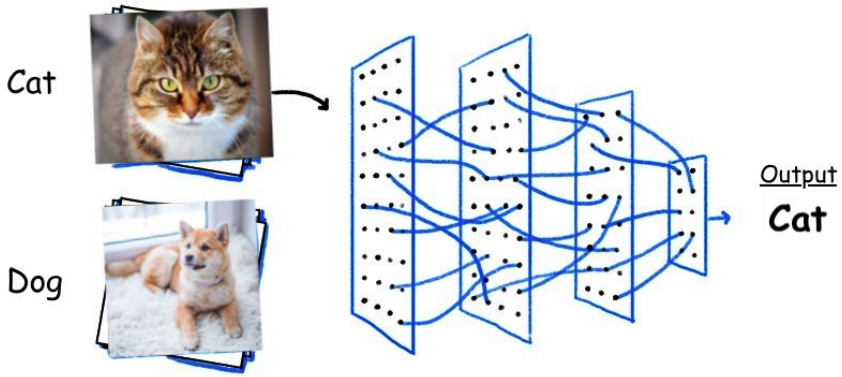
- Supervised Machine Learning task
- Predict categorical or discrete class labels
- Data instances are assigned to predefined classes based on shared features or qualities
- Outcome variable ⟶ class!
- Model is trained on these features in order to be able to predict the class or category the data point belongs to
- Once trained, your model should be able to make classifications on unseen data
- Classification can aid in domains that involve decision making, and providing insights into data and patterns



Classification use cases

Where is it used? Why is it important?

- Healthcare
- Finance
- Customer Service
- Marketing
- Natural Language Processing
- Image and Object recognition
- Social Media analysis
- Environmental Sciences
- Manufacturing and Quality Control
- Security and Intrusion Detection



Types of classification

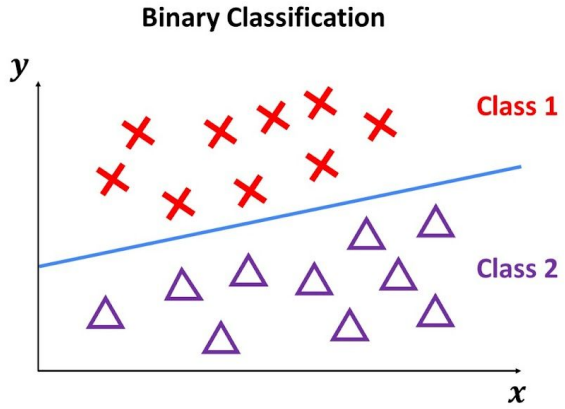
How do we classify?

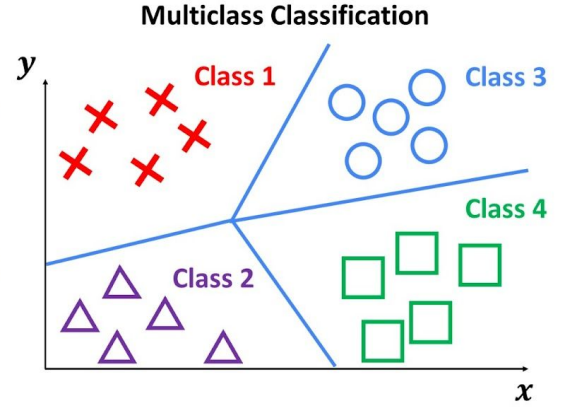
Soft vs Hard predictions:

- Soft - data points have predicted probabilities of being in each class
- Hard - data points are predicted to be in one class and only one class

Binary vs Multiclass classification:

- Binary - data points are in one of two classes
- Multi-class - data points are in one of multiple classes



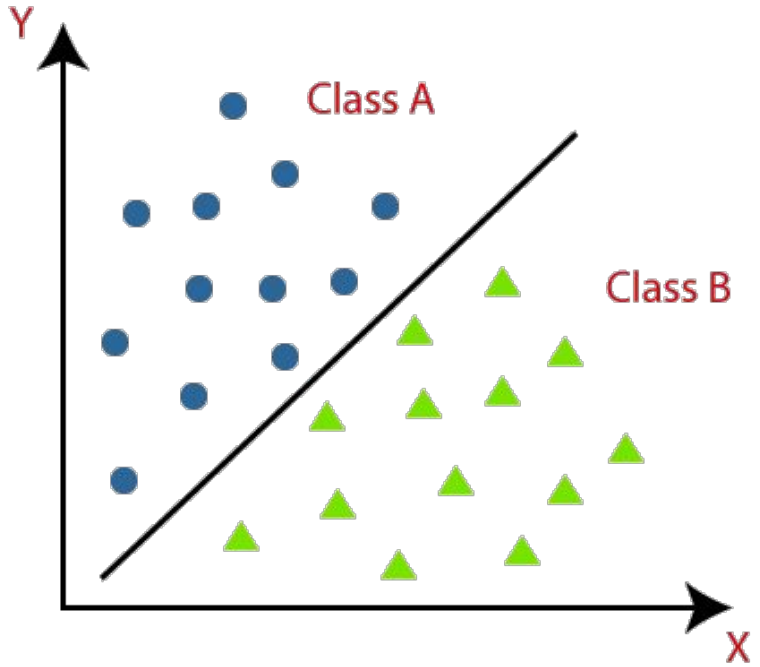


Classification algorithms

How do we classify?

Some popular algorithms:

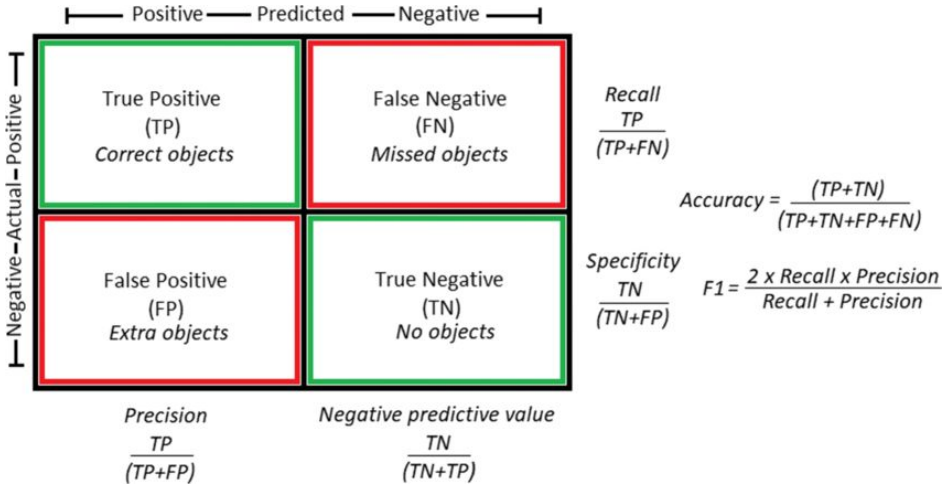
- Logistic Regression
- Decision Tree Classifiers
- Random Forest Classifiers
- Support Vector Machines
- Naive Bayes
- *and more...*



Evaluation metrics

How do we know how well our model is doing?

- **Accuracy**: overall correctly classified instances
- **Precision**: proportion of correctly predicted positives out of all predicted positives
- **Recall**: proportion of correctly predicted positives out of all actual positives
- **F1 Score**: Harmonic mean of precision and recall - balanced measure of model performance



1 / **Objectives**

2 / **Recap: What is Machine Learning?**

3 / **Recap: Regression?**

4 / **What is Classification?**

5 / **What will we be seeing again?**

6 / **Natural Language Processing**

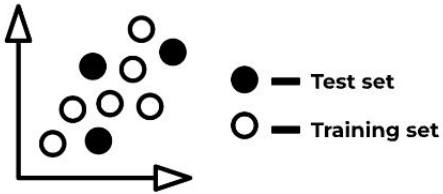
7 / **Sprint overview**

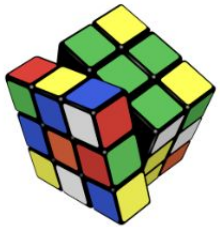
8 / **Predict overview**

The Data Science Process

How do we solve a data science problem?

Preprocessing → **Model Building** → **Model Evaluation** → **Deployment**





- Data Cleaning
  - Impute
  - Normalise/Standardise
  - Label/Dummy encode
- Train-Test split / Kfold

- Model Selection
- Model Training
- Hyperparameter Tuning

- Evaluate Model on Test set
- Report Performance metrics

- Hosting and Versioning
- Dashboards
- Containerization
  - Docker
  - Kubernetes

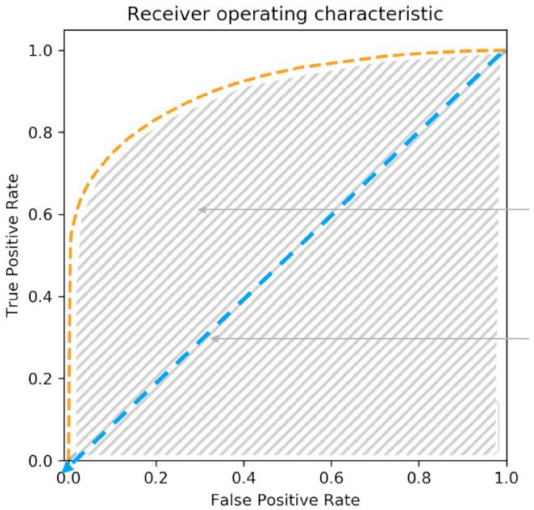
Processing and splitting our data

Remember: trash in = trash out

- **Features:** variables or predictors we use as inputs to our model
- **Labels:** our output - the class or category the data belongs to
- **Training Data:** the labelled data we use to train our model
    - Splitting our data into a ‘train’ and ‘test’/‘validation’ portion, so we can train our data, and then evaluate how well it is performing
    - This can help us to tune our model and select best hyperparameters
- **Test Data:** the unlabelled data we use to see how well our model can perform
- **Exploring data:** we need to actually *understand* our data
- **Processing data:** we need to know how to get our data and our features into an optimal format for our modelling process!

Modelling and evaluation

- Create and train our models
- Evaluate their performance on a test/validation set
- Tuning models - improving and optimising performance
- Selecting our best performing model(s)
- Make predictions on truly unseen data!



1 / **Objectives**

2 / **Recap: What is Machine Learning?**

3 / **Recap: Regression?**

4 / **What is Classification?**

5 / **What will we be seeing again?**

6 / **Natural Language Processing**

7 / **Sprint overview**

8 / **Predict overview**

Natural Language Processing (NLP)

What is it?

- Human written or spoken language is largely unstructured
- NLP enables computers to understand and interpret human language
- Uses algorithms and techniques to process and analyse natural language data
- NLP is crucial for certain elements of human-computer interaction and for extracting insights from text data
- Applications in machine translation, chatbots, sentiment analysis and more

Natural Language Processing

Where is it used?

- **Spam filters**
  - Scan the text of each email.
  - Attempt to gain context or understanding.
  - Determine whether spam or not.
- **Algorithmic Trading**
  - Read and digest masses of news and articles relevant to stocks.
  - Combined with ML, determines buy/hold/sell positions.
- **Answering questions**
  - Major use-case: have search engines understand what we mean.
  - Bonus: respond in the same language, tone, etc.
  - Used widely in Siri, Google Assistant, Alexa, etc.
- **Summarising information**
  - Far too much info out there for us to process wholly.
  - Using NLP we can parse large document volumes.
  - Attempt to understand meaning and generate summaries.

1 / **Objectives**

2 / **Recap: What is Machine Learning?**

3 / **Recap: Regression?**

4 / **What is Classification?**

5 / **What will we be seeing again?**

6 / **Natural Language Processing**

7 / **Sprint overview**

8 / **Predict overview**

Advanced Classification - Learning Journey

1 Overview

- Familiarise yourself with what you'll cover over the next few weeks.
- Look at your project instructions, download the data and look at the problem statement.

2 Natural Language Processing

- How Machines Understand Language
- NLP Practical Test [Code Challenge]
- NLP Theory Test [MCQ]

3 Logistic Regression & Classification Metrics

- Intro to Binary Classification Using Logistic Regression
- Logistic Regression [Code Challenge]
- Binary Classification Metrics

4 Model Improvements

- Improving Classification Models
- Improving a Logistic Regression Model
- Dealing with Imbalanced Data
- Introduction to Multiclass Classification

5 Tree-based Methods & Support Vector Machines

- Tree-Based Methods for Classification
- Support Vector Machines

6 Ensemble Methods, Tuning & Validation

- KNNs and Naive Bayes
- Hyperparameters and Model Validation
- Hyperparameter Tuning [Code Challenge]

7 Neural Networks & Comparing Models

- Neural Network Classifier
- Build All the Classifiers
- Image Classification [Code Challenge MCQ]

8 Predict Project

- Advanced Classification Predict

Learning activities include videos, interactive tools, knowledge tests and curated external training.

Collect "BELT POINTS" along the way to level-up!

EXPLORE | DIGITAL SKILLS

Important note

- There will be a break over the Christmas/New Year’s period
- Academy will close and course will pause (but you will still be able to access Athena over this time)
- Course will resume in January and we will pick up where we left off
- **Dates and more details will be shared soon**

1 / **Objectives**

2 / **Recap: What is Machine Learning?**

3 / **Recap: Regression?**

4 / **What is Classification?**

5 / **What will we be seeing again?**

6 / **Natural Language Processing**

7 / **Sprint overview**

8 / **Predict overview**

Predict Project

Build and deploy Classification models and to participate in a Kaggle challenge.

You

You are tasked with building a classification model(s) to identify users' sentiments towards climate change based on their novel tweet data

NLP + Classification Models

Supervised machine learning techniques covered throughout this sprint will be used to build a model to classify your data.

Python

You are free to use any relevant classification method(s).

Learn

The purpose of this predict is to guide you through the typical steps of a real-world data science projects from initial EDA, to model development and deployment and finally to communication of results.

Questions?