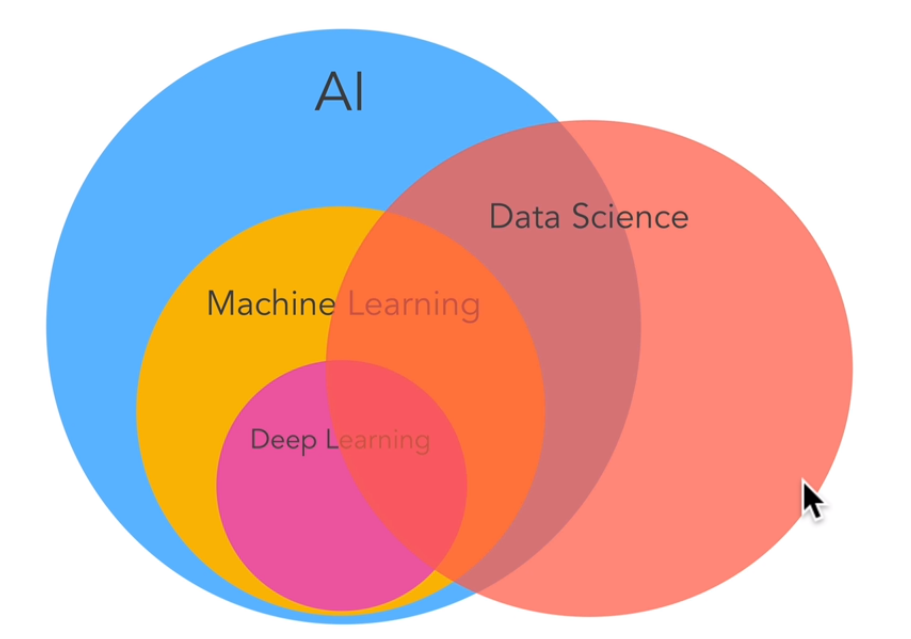
**YES NOTES !**



A.I. Artificial Intelligence:-

Human Intelligence exhibited by a machine.

Narro A.I. : A.I. doing only One task Really well.

General A.I. : A.I. having multiple abilities.

Machine learning:

Science of getting computers to act without being explicitly programmed.

Deep learning or deep neural networks:

Techniques or algorithms for implementing machine learning.

Data science:

It is to analyze data and do something with it. It overlaps with machine learning.

MACHINE LEARING:

Machine learning is using an algorithm or computer program to learn about different patterns in data, and then taking that algorithms(models) and what it's learned to make predictions about the future using similar data.

Data science is running experiments on a set of data with the hopes of finding actionable insights within it.

Data analysis is looking at a set of data and gaining an understanding of it by comparing different examples, different features and making visualizations like graphs.

Machine learning , lets computers learn from data and they make predictions and decisions.It allows computers to do tasks that in the past required humans, and make our lives hopefully easier.



Framework:-

1. **Problem definition:**

There are four Types of Machine Learning Problems:

* **Supervised learning** – Supervised learning, is called supervised because you have data and labels. A machine learning algorithm tries to learn what patterns in the data lead to the labels. If the algorithm guesses a wrong label, it tries to correct itself.

Types:

* Classification: Classification involves predicting if something is one thing or another. Such as if you want it to predict whether or not a patient had heart disease or not ,based on their medical records (Binary classification) or what type of dog breed was in an image (multiple classification).
* Regression: Problems involve trying to predict a number. Do you want to predict a specific number of something? Such as how much a house will sell for? Or how many customers will visit your site next month?
* **Unsupervised learning** -Unsupervised learning has data but no labels. You provided the labels. They weren't there to begin with, but the patterns were. E.g.: customers by purchase history. After inspecting the groups, you provide the labels. There may be a group interested in computer games, another group who prefer console games and another which only buy discounted older games. This is called clustering. Recommendation problems like music Recommendation on previous history comes under unsupervised learning problems.
* **Transfer learning** – Transfer learning is when you take the information an existing machine learning model has learned and adjust it to your own problem. Transfer learning leverages what one machine learning model has learned in another machine learning. E.g. we can take this car model that identifies different cars and use its foundational patterns and apply it to our dog breed problem.
* **Reinforcement learning** - Reinforcement learning involves having a computer program perform some actions within a defined space and rewarding it for doing it well or punishing it for doing not well. E.g. Machine learning model to play chess. The machine learning algorithm’s goal could be to maximize the reward.

1. **Data:**

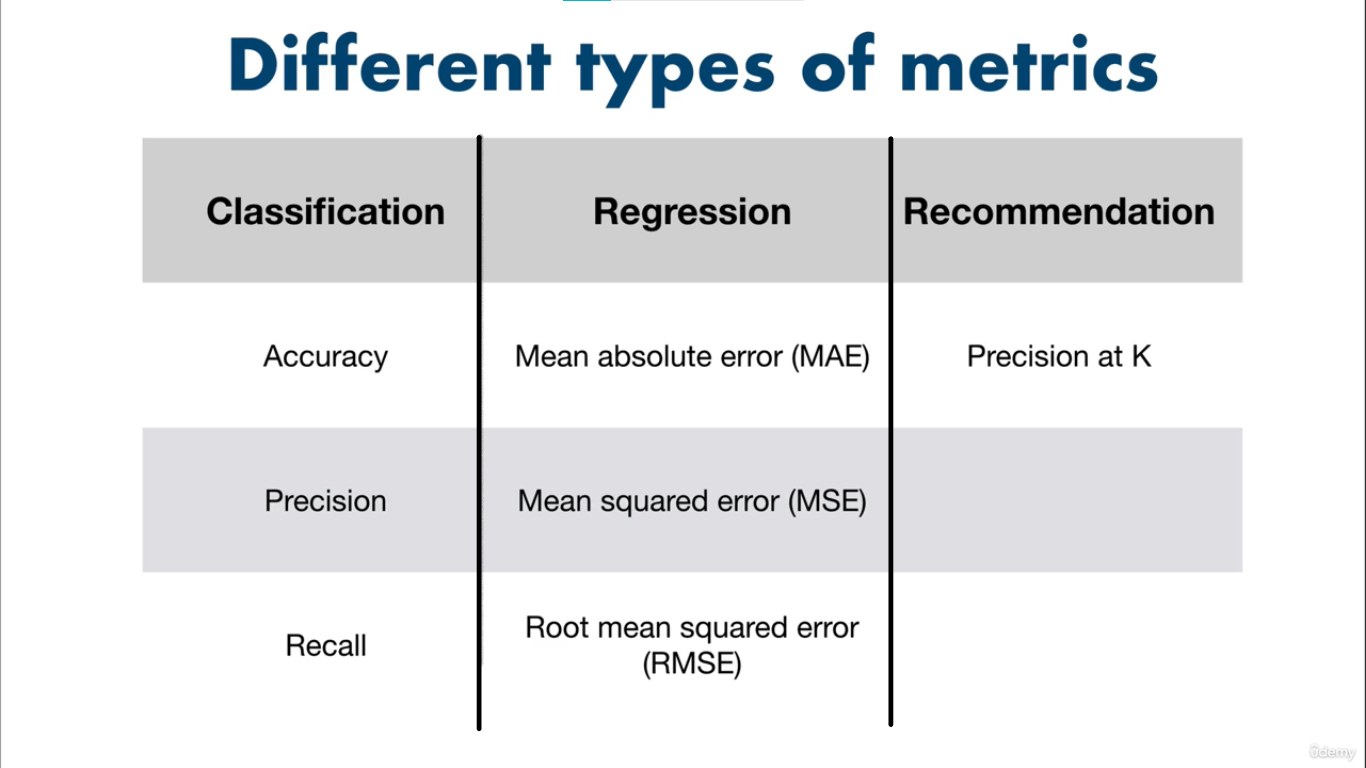
* **Structured data** — Think a table of rows and columns, an Excel spreadsheet of customer transactions, a database of patient records. Columns can be numerical, such as average heart rate, categorical, such as sex, or ordinal, such as chest pain intensity.
* **Unstructured data** — Anything not immediately able to be put into row and column format, images, audio files, natural language text.
* **Static data** — Existing data which is unlikely to change. Your company’s customer purchase history is a good example.
* **Streaming data** — Data which is constantly updated, older records may be changed, newer records are constantly being added.

There can be overlaps.

1. **Evaluation:**

An evaluation metric is a measure of how well a machine learning algorithm predicts the future.

The question you want to answer is What defines success for us?



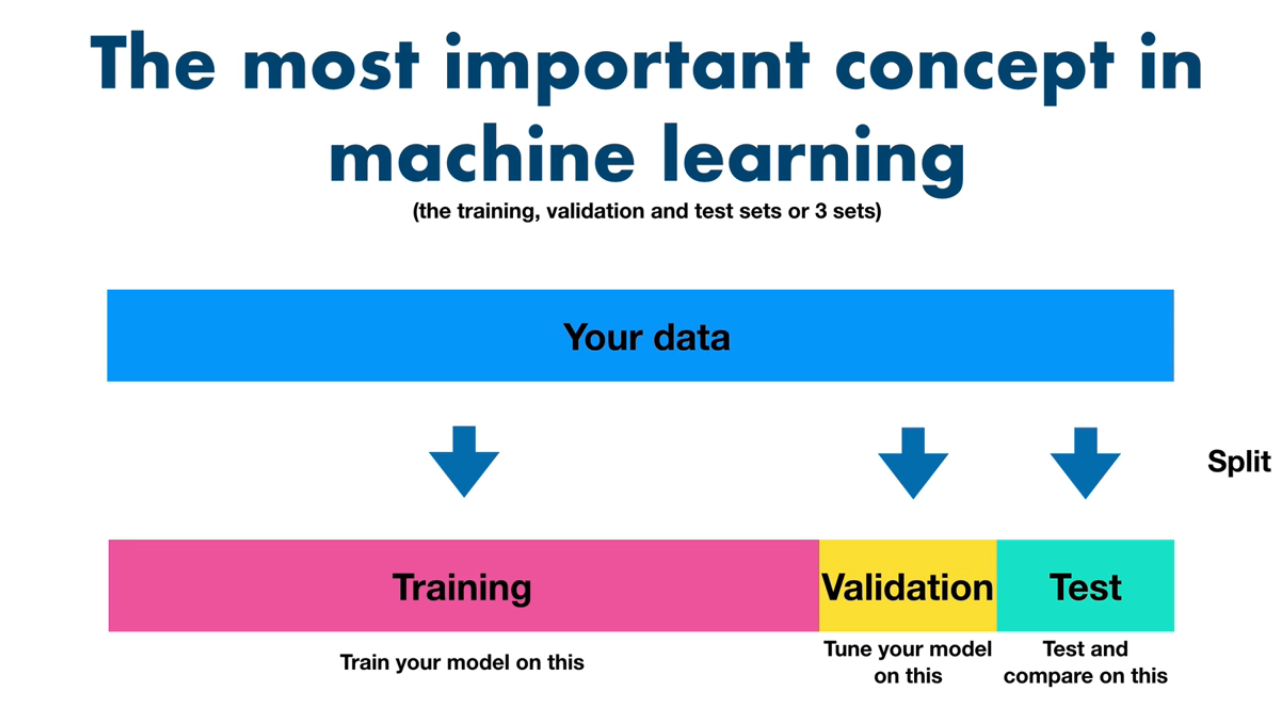
1. **Features:**

Types:

* **Categorical features** — One or the others. For example, in our heart disease problem, the sex of the patient. Or for an online store, whether someone has made a purchase or not.
* **Continuous or Numerical features** — A numerical value such as average heart rate or the number of times logged in.
* **Derived features** — Features you create from the data. Often referred to as feature engineering. Feature engineering is how a subject matter expert takes their knowledge and encodes it into the data.

The process of ensuring all samples have similar information is called feature coverage. In an ideal data set you've got complete feature coverage. Means all Rows of columns must be filled.

1. **Modelling:**



In machine learning, since you want to be using machine learning models to gain insights on some data to predict the future, it's important to test how well they would go and do in the real world.

To do this, you split your data into three different sets ; a **Training set** to train your model , a **Validation set** to churn your model on and a **Test set** to test and compare your different models. It is also called as 3 sets.

**Generalization**: Ability for a machine learning model to perform well

on data it hasn't seen before because of what it's learned on another data set.

A diagram of a model

Description automatically generated

Modelling is done in 3 parts ;

1. Choosing and training a model:

A diagram of a model

Description automatically generated

A white background with black text

Description automatically generated

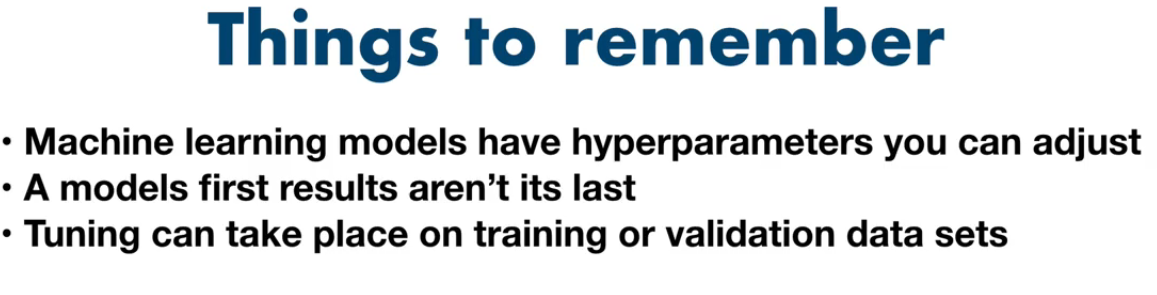
1. Turning a model:

Tuning a model involves changing hyperparameters such as learning rate or optimizer. Or model-specific architecture factors such as number of trees for random forests and number of and type of layers for neural networks.

These used to be something a practitioner would have to tune by hand but are increasingly becoming automated.

A diagram of a model

Description automatically generated



1. Model comparison:

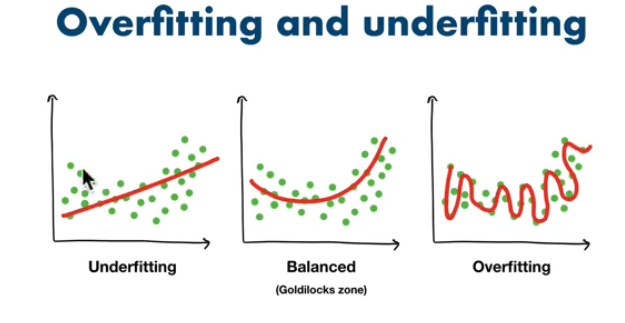
A computer testing a model

Description automatically generated with medium confidence

if the training set performance is dramatically higher than the test set, also known as unbefitting.

And if the test set performance is higher than the training set performance also known as overfitting.

Overfitting and unbefitting of both examples of a model not being able to generalize well, which is what we don't want.



Main reasons for unbefitting and overfitting can happen, are: data leakage and data mismatch.

Data leakage happens when some of your test data leaks into your training data, and this often results in overfitting or a model doing better on the test set than on the training data set.

Data mismatch happens when the data you're testing on is different to the data you're training on,

such as having different features in the training data to the test data.

A diagram of overfitting and overfitting

Description automatically generated

A close-up of a sign

Description automatically generated

* **Training data set** — Use this set for model training, 70–80% of your data is the standard.
* **Validation/development data set** — Use this set for model hyperparameter tuning and experimentation evaluation, 10–15% of your data is the standard.
* **Test data set** — Use this set for model testing and comparison, 10–15% of your data is the standard.

These amounts can fluctuate slightly, depending on your problem and the data you have.

Poor performance on training data means the model hasn’t learned properly and is **underfitting**. Try a different model, improve the existing one through hyperparameter or collect more data.

Great performance on the training data but poor performance on test data means your model doesn’t generalize well. Your model may be **overfitting** the training data. Try using a simpler model or making sure your the test data is of the same style your model is training on.

Another form of **overfitting** can come in the form of better performance on test data than training data. This may mean your testing data is leaking into your training data (incorrect data splits) or you've spent too much time optimizing your model for the test set data. Ensure your training and test datasets are kept separate at all times and avoid optimizing a models performance on the test set (use the training and validation sets for model improvement).

Poor performance once deployed (in the real world) means there’s a difference in what you trained and tested your model on and what is actually happening. Ensure the data you're using during experimentation matches up with the data you're using in production.

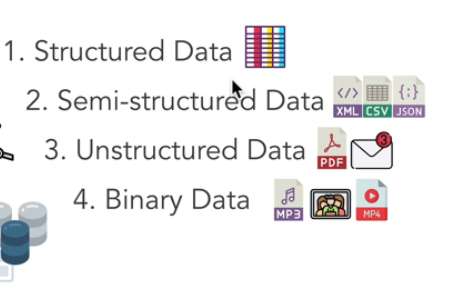
1. **Experimentation:**

The Questions we ask here are, what have we tried ?, what else we can try?.

This step involves all the other steps. Because machine learning is a highly iterative process, you’ll want to make sure your experiments are actionable.

DATA ENGINEERING:

* TYPES OF DATA:



* Profiles:

