PRIOR KNOWLEDGE

Supervised learning:

Supervised machine learning requires labelled input and output data during the training phase of the machine learning life cycle. This training data is often labelled by a data scientist in the preparation phase, before being used to train and test the model. Once the model has learned the relationship between the input and output data, it can be used to classify new and unseen datasets and predict outcomes.

The reason it is called supervised machine learning is because at least part of this approach requires human oversight. The vast majority of available data is unlabelled, raw data. Human interaction is generally required to accurately label data ready for supervised learning. Naturally, this can be a resource intensive process, as large arrays of accurately labelled training data is needed.

Unsupervised learning:

Unsupervised machine learning is the training of models on raw and unlabelled training data. It is often used to identify patterns and trends in raw datasets, or to cluster similar data into a specific number of groups. It's also often an approach used in the early exploratory phase to better understand the datasets.

Unsupervised machine learning is mainly used to:

- Cluster datasets on similarities between features or segment data
- Understand relationship between different data point such as automated music recommendations
- Perform initial data analysis

The main differences of supervised vs unsupervised learning include:

- The need for labelled data in supervised machine learning.
- The problem the model is deployed to solve. Supervised machine learning is generally used to classify data or make predictions, whereas

unsupervised learning is generally used to understand relationships within datasets.

- Supervised machine learning is much more resource-intensive because of the need for labelled data.
- In unsupervised machine learning it can be more difficult to reach adequate levels of exhplainability because of less human oversight.