Building blocks templates, general research

1. What machine learning building blocks do we need?

On top of all, we need sample datasets to make learning and testing easy.

Does each pipeline (consisting of building blocks) run on a single container (Docker/Kubernetes) or each individual block run by itself on a container?

# What are the most common machine learning algorithms called?

* **Linear regression**: This is a simple algorithm used for predicting a continuous outcome variable (e.g. a real number) based on one or more predictor variables. It assumes that the relationship between the predictor and outcome variables is linear, and it finds the line that best fits the data according to a measure of error.
* **Decision trees**: This is a supervised learning algorithm that can be used for classification or regression. It works by creating a tree-like structure, where each internal node represents a decision based on the value of a predictor variable, each branch represents the possible outcomes of that decision, and each leaf node represents a prediction. Decision trees are intuitive and easy to interpret, but they can be prone to overfitting.
* **Support vector machines (SVMs)**: This is a supervised learning algorithm that can be used for classification or regression. It works by finding the hyperplane in a high-dimensional space that best separates the data points into different classes. SVMs are effective in high-dimensional spaces, but they can be computationally expensive and may not scale well to large datasets.
* **K-nearest neighbors (KNN)**: This is a simple, non-parametric supervised learning algorithm that can be used for classification or regression. It works by finding the K data points in the training set that are closest to the new data point, and then using those data points to make a prediction. KNN is easy to implement and can be effective for small datasets, but it can be computationally expensive and may not scale well to large datasets.
* **Neural networks**: This is a supervised learning algorithm that is inspired by the structure and function of the brain. It works by creating a network of interconnected nodes, where each node represents a unit of computation and each connection represents a weight that can be adjusted to improve the model's performance. Neural networks can be highly effective for complex, non-linear problems, but they require a large amount of data and can be computationally expensive.
* **Deep learning**: This is a subset of machine learning that uses neural networks with many layers (hence the term "deep") to learn complex patterns in data. Deep learning algorithms can automatically learn features from raw data, which makes them particularly well-suited to tasks such as image and speech recognition. However, they require even larger amounts of data and computation than other neural network algorithms.
* **Random forests**: This is an ensemble learning method that can be used for classification or regression. It works by training multiple decision trees on different subsets of the data, and then combining their predictions to make a more accurate and stable prediction. Random forests can be effective for a wide range of problems, but they can be slow to train and may not be the best choice for real-time predictions.
* **Clustering**: This is an unsupervised learning algorithm that is used to group data points into clusters based on their similarity. Clustering algorithms can be used to discover hidden patterns in data, and they are often used for exploratory data analysis. Some popular clustering algorithms include k-means, hierarchical clustering, and density-based clustering.

It is generally not recommended to run multiple applications in a single Docker container. Each container should have a single, specific purpose, and running multiple applications in a single container can make it difficult to manage and update the applications. It can also make it more difficult to troubleshoot problems or isolate issues with a specific application.

Instead of running multiple applications in a single container, it is better to use multiple containers, each with a single application. This will allow you to manage and update each application separately, and it will make it easier to scale and distribute your applications across multiple host machines.

**Therefore:  
Each ML Block should run on a separate container since each algorithm have different dependencies.**

## More ML Options

Sample datasets

## Config building blocks

Each type of algorithm or Dataset can have different configuration that depends on the requirements of the algorithm.

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

Split Data example

# Arrows (Ways of connecting ML Blocks)

Arrows should only be pointing at different ways of connecting the blocks.

Some blocks containing some algorithm, can have multiple ways of connecting to other blocks (containing other algorithms)