Any machine learning task can be broken down into a series of steps:

1. Selecting the Machine Learning approach: Before starting any steps, the machine learning problem needs to be expressed. What do we want to find out? Do we want to classify our data, to predict new values, to group our data based on some criteria? After we decide what type of
machine learning task we would like to perform, we select our model.

2. Collecting data: Data can be written on paper, recorded text files and spreadsheets or stored in an SQL database. Data need to be gathered in an electronic format suitable for analysis.

3. Exploring and preparing the data: The quality of any machine learning project is based on the quality of the data it uses. It is suggested that 80-% of the effort in ML is devoted to data preparation. This step requires a great deal of human intervention.

4. Training the model on the data: The specific machine learning task will inform the selection of an appropriate algorithm. We then “feed” the data to the model during this phase and we will get a learner. A learner is a ML algorithm that has been trained on some data and adjusted to fit the data as best as possible.

5. Evaluating model performance: Because each learner results in a biased solution, it is important to evaluate how well the algorithm learned from its experience. Depending on the model used, we might be able to evaluate the accuracy of the learner using a test dataset.

6. Improving model performance: If better performance is needed, it becomes necessary to utilize more advanced strategies to improve the performance of the model, or switch to a different model, supplement with additional data and perform additional preparation work on the data (step 3).

After these steps have been completed, if the model appears to be performing satisfactorily, it can be deployed for its intended task. The model can be used to provide scores for predictions, projected values for novel data, generate useful insight for marketing or research.
It is important to keep track of the successes and failures of the deployed model in order to provide additional data to train the next generation of your model.

Choosing a machine learning algorithm:
The choice of machine learning algorithm is dependent on the data we have available and the proposed task at hand. It is important to be thinking about this process while you are gathering, exploring and cleaning your data.

(No free lunch)

Input data:
All machine learning algorithms require training data. Training data takes the form of examples and features.

An **example** is a single exemplary instance of the underlying concept we want learned. It is one set of data describing the atomic unit of interest for the analysis. If we are building a learning algorithm to identify spam email, the examples would be data from many electronic messages. To detect cancerous tumors, the examples might comprise biopsies from a number of patients.

A **feature** is a characteristic or attribute of an example, which might be useful for learning the desired concept. In the previous examples, attributes in the spam detection dataset might be words used in the email messages or domain names. For the cancer dataset, it might be genomic data from the biopsied cells, or measured characteristics of the patient such as weight, height, age, blood pressure.
Features come in various forms. If a feature represents a characteristic measured in numbers, it is called **numeric**, for example height and weight are numeric. Alternatively, if a feature measures an attribute that is represented by a set of categories, the feature is called **categorical** or **nominal**, for example gender is nominal. If these categories have a given order to them, then the feature is called **ordinal**, for example shirt sizes (XS, S, M, L, XL,) or a measurement of customer satisfaction on a scale from 1 to 5. It is important to consider what the features represent because the type and number of features in your dataset will assist with determining an appropriate machine learning algorithm for your task.

Types of machine learning algorithms:

Machine learning algorithms can be divided into two main groups: **supervised learners** that are used to build predictive models and **unsupervised learners** that are used to build descriptive models. Which type is needed depends on the learning task you hope to accomplish.

A predictive model is used for tasks that evolve the prediction of one value using other values in the dataset. The learning algorithm attempts to discover and model the relationship among the **target** feature (the feature being predicted) and the other features. Predictive models do not necessarily imply forecasting future events.

Because predictive models are given clear instruction of what they need to learn, in terms of positive instances and negative instances and how they are intended to learn it, this process is known as **supervised learning**. The supervision does not refer to human
involvement, but rather that the target values provide a supervisory role, which guides the learner in what they need to learn. Specifically, given a set of data, the algorithm attempts to optimize a function (the model) to find a combination of feature values that will result in the target output.

Often, the supervised machine learning of predicting which category an example belongs to is known as **classification**.

The target feature to be predicted is a categorical feature known as the **class** and is divided into categories called **levels**. A class can have multiple levels.

Supervised learners can also be used to predict numeric values such as income, laboratory values, and test scores... etc. To predict such numeric values linear regressions models are usually used. Although regression models are not the only type of numeric models, they are by far the most widely used.

A **descriptive model** is used for tasks that would benefit from the insight gained by summarizing data in new and hopefully interesting ways. As opposed to predictive models that predict a target of interest; in a descriptive model, no single feature is more important than another. Because there is no target to learn, the process of training a descriptive model is called **unsupervised learning**. These methods are regularly used for data mining. The descriptive modeling task called **pattern discovery** is used to discover frequent association with data. Pattern discovery is often used for **market basket analysis** on transactional purchase data. Here the goal is to identify items frequently purchased together.

The descriptive modeling task of dividing a dataset into homogeneous groups is called **clustering**. This is used for segmentation analysis that identifies groups of individuals with similar purchasing, donating, reading or demographic information so that advertising campaigns can be tailored.