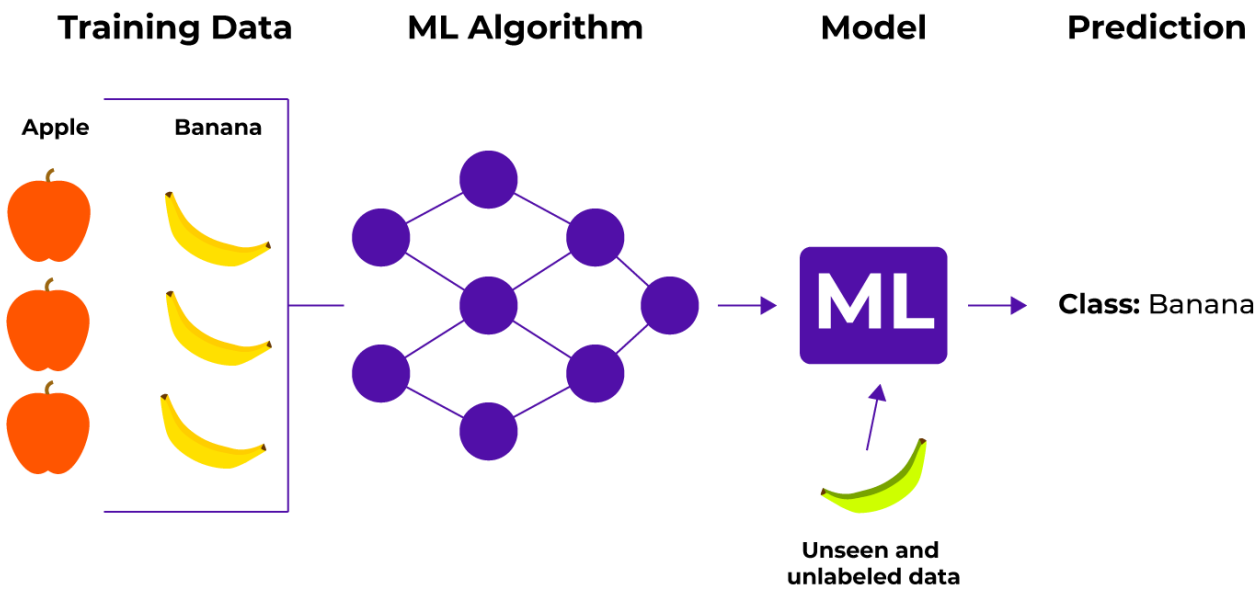
Mikias Berhanu | 2021280115
Assignment III
Supervised and Unsupervised Learning

Supervised Learning and Unsupervised Learning

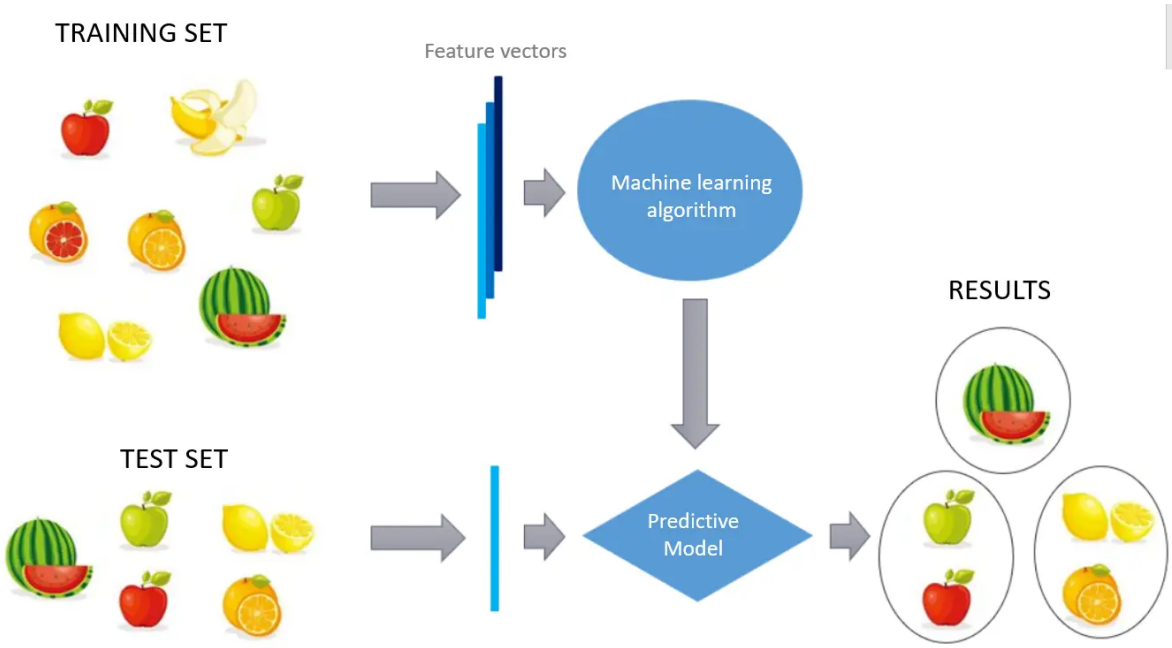
Machine Learning can be classified into four main categories: Supervised Learning, Semi-Supervised Learning, Unsupervised Learning and Reinforcement Learning. The sort of classification is mainly based on the amount of supervision the models get from humans. Supervised Learning is one of the most common methods of training Machine Learning models where the models are trained with labeled data, meaning the models are given the answer to some subset of features of predictors. All the models do is match the features to their respected labels. Supervised Learning can be divided into two sub categories based on the task we want to perform: Regression and Classification. Regression is when we use labeled data and try to predict some continuous numeric values. For example: predicting the price of a car given the car properties like engine power, model number, body strength ... Another example can be trying to predict the price of a house given the number of rooms available, the area of the house ... The second type of Supervised Learning is Classification whereby we try to classify certain objects into their respective class or label given their properties. Classification can be classified into two as binary classification and multi-class classification. Binary classification is when we only have two classes and the given object can be a part of only one class. For example, classifying the images of cats and dogs, given a number of labeled images as cats and dogs we wanna be able to predict if a new image which is not part of the training set is cat or dog. This kind of classification is called binary classification since we only have two classes cats and dogs which can be numerically represented as 0 and 1 thus the name binary classification. Multi-classification on the other hand, has more than two classes and given properties of an object we predict which one of the classes it belongs to. As an example, given a picture of an animal, predict which class it belongs to: lion, tiger, zebra ... Some of the popular algorithms for tackling regression problems are Linear Regression, Polynomial Regression, using Random Forest Trees for regression, Decision trees etc... And

some popular algorithms for handling classification tasks: Logistic Regression, Decision Trees, Random Forest Classifier, Supper Vector Machines etc...



Supervised Learning for Classification Task source

Unlike Supervised Learning, Unsupervised learning learns on data which is not labeled and the models try to extract some sort of pattern from the data and make predictions. These models are very useful when we want to cluster a big chunk of unlabeled data into a cluster or subgroups, when we want to find some hidden patterns from our data ... Some of the well known Unsupervised learning models which can be used for clustering are: Hierarchical Clustering, K-means clustering and K-means clustering is further subdivided into two as: Agglomerative Clustering and Dendrogram. Compared to Supervised learning, getting high accuracy might be very challenging due the fact that the training data is unknown and has no labels, the models have to extract the patterns on their own.



Unsupervised Learning for Clustering task source

Getting back to Supervised Learning one of the most common use cases is Regression where we try to predict continuous numeric values given a set of features of predictors with their label or true values. Regression can be subdivided into three groups: Multiple Regression when we use more than one or two features to make our prediction, Univariate Regression when we predict a single value and Multivariate Regression when we try to predict multiple values. Whenever we work with ML models there should always be a performance measuring comparison which will help us to know whether our model is performing well or not. And for Regression models, commonly used performance measures are Root Mean Squared Error(RMSE) and Mean Absolute Error(MAE) or Mean Absolute Deviation. These two functions are used as a cost function and they measure distance between the vectors of predicted values and vectors of actual values. Both of these functions measure vector distance RMSE measures Euclidean Norm whereas MAE measures Manhattan Norm. The big difference between two cost functions is that RMSE is more affected by outliers in the data whereas MAE is not that much affected.

A short summary about supervised and unsupervised learning is given below

Supervised Learning	Unsupervised Learning
Uses labeled data for training, validation and testing the model	Uses unlabeled data for training validation and testing the model
The data and any existing patterns are analyzed offline	The data and any existing patterns inside the dataset are analyzed on the fly.
In the case of classification tasks the number of classes are well known before training and testing the model.	There won't be any prior knowledge about the number of classes, classes are extracted after the learning process.
The results are accurate	The results are somehow average
Computationally expensive and need more processing power	Does not require a large amount of computing resources.