**HOME LEARNING TASK 6**

**PART A**

Select one or more choices from the list of common Machine Learning Algorithms, do some investigations and write me a short summary. I am looking for the following:

* Is it Supervised/Unsupervised/Reinforcement learning?
* What does the algorithm do?
* In which situations will it be most useful?
* (Optional) Can you find any examples of where this algorithm has been used.

**Linear Regression**

* Linear Regression is a machine learning algorithm based on supervised learning.
* It performs a regression task. Regression models a target prediction value based on independent variables.
* It is mostly used for finding out the relationship between variables and forecasting.
* Healthcare - used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

Logistic Regression

* Logistic regression is a supervised learning algorithm because it uses true labels for training.
* It is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost function between 0 and 1.
* It useful when the response variable is binary but the explanatory variables are continuous. This would be the case if one were predicting whether or not an customer is a good credit risk, using information on their income, years of employment, age, education, and other continuous variables.
* Some of the examples of classification problems are Email spam or not spam, Online transactions Fraud or not Fraud, Tumour Malignant or Benign.

**Decision Tree**

* Decision tree is a supervised learning algorithm
* The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).
* A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes).
* Classification of a new groups of species.

**SVM (Support Vector Machine)**

* Support vector machine is a supervised machine learning algorithm,.
* In the SVM algorithm each data item is plotted as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).
* Analysing data and recognising patterns in close related datasets.
* Distinguishing between closely related gene expression of proteins when characterising and developing a new atlas of a biological niche.

**Naive Bayes**

* Naïve Bayes is a semi-supervised machine learning algorithm.
* The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. Thus, making them to be considered as naive. You can derive probability models by using Bayes' theorem (credited to Thomas Bayes).
* Used in a wide variety of classification tasks. Typical applications include filtering spam, classifying documents, sentiment prediction etc…

**KNN (K- Nearest Neighbours)**

* K-Nearest Neighbour is a supervised machine learning algorithm.
* The algorithm is simple and easy-to-implement and is used to solve both classification and regression problems.
* Applications that require high accuracy but that do not require a human-readable model. The quality of the predictions depends on the distance measure.
* Hand computation.

**K-Means**

* K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
* The method of [vector quantization](https://en.wikipedia.org/wiki/Vector_quantization), originally from [signal processing](https://en.wikipedia.org/wiki/Signal_processing), that aims to [partition](https://en.wikipedia.org/wiki/Partition_of_a_set) *n* observations into *k* clusters in which each observation belongs to the [cluster](https://en.wikipedia.org/wiki/Cluster_(statistics)) with the nearest [mean](https://en.wikipedia.org/wiki/Mean) (cluster centres or cluster [centroid](https://en.wikipedia.org/wiki/Centroid)), serving as a prototype of the cluster.
* Used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.
* Tourism in a specific location year-round or whilst analysing trends on destinations that the airline keeps an eye on.

**Random Forest**

* Random forest is a supervised machine learning algorithm.
* Used widely in classification and regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. It performs better results for classification problems.
* Used by interface platforms to classify whether an email is spam or non-spam.
* Using the random forest approach to capture genetic effects in the presence of population structure.