Machine Learning Workflow

Affine
Data ▸ Insights ▸ Transformation

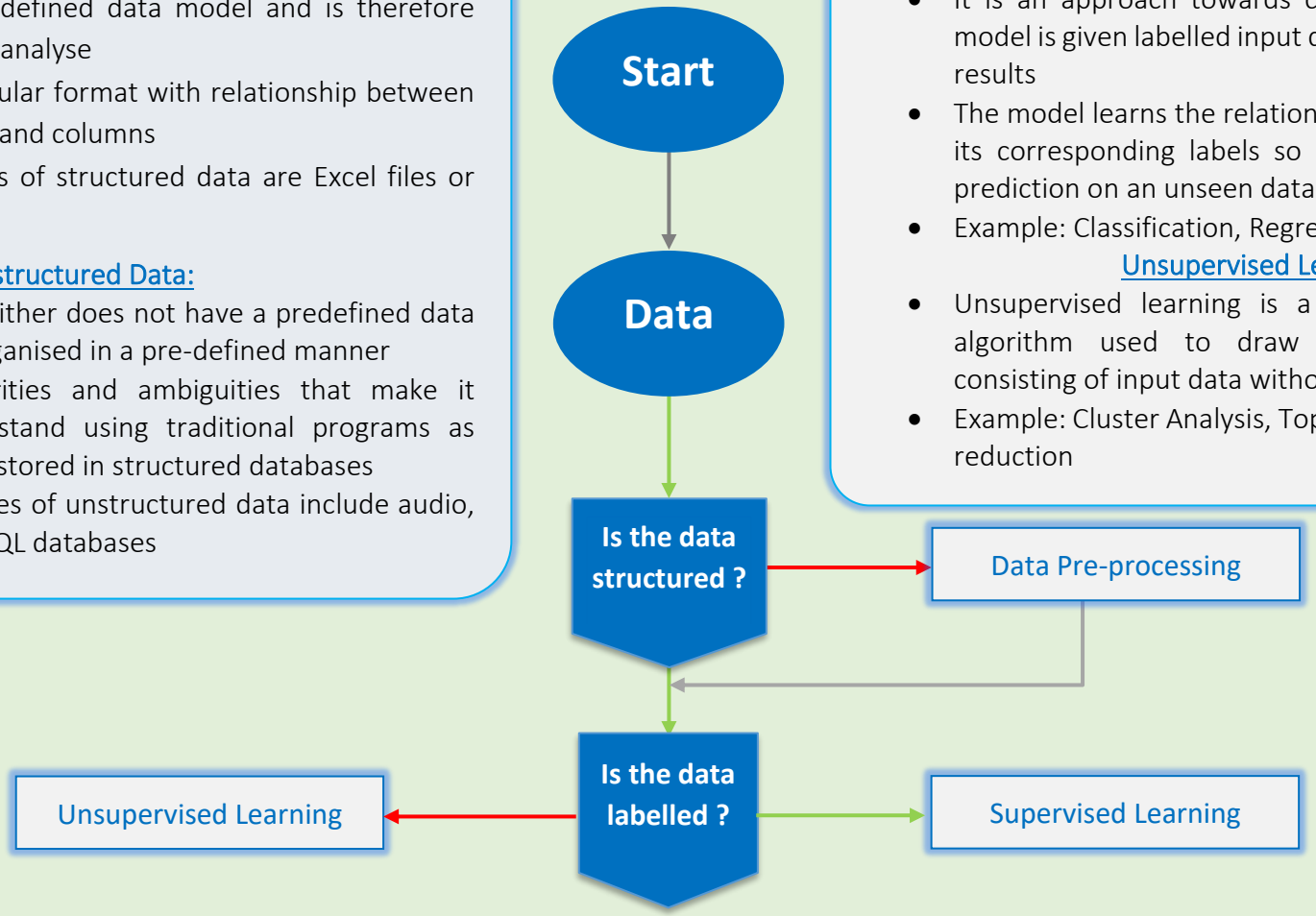
DATA → **EDA:** Data exploration, Outlier treatment, Data visualization → **Pre-processing:** Missing value treatment, Data cleaning, Data transformation → **Pre-modeling:** Sampling, feature importance, dimensionality reduction → **Modeling** → **Post modeling:** Hyperparameter tuning, model evaluation, cross validation

Structured Data:

- Adheres to a pre-defined data model and is therefore straightforward to analyse
- Conforms to a tabular format with relationship between the different rows and columns
- Common examples of structured data are Excel files or SQL databases

Unstructured Data:

- Information that either does not have a predefined data model or is not organised in a pre-defined manner
- Contains irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in structured databases
- COMMON examples of unstructured data include audio, video files or No-SQL databases



Supervised Learning:

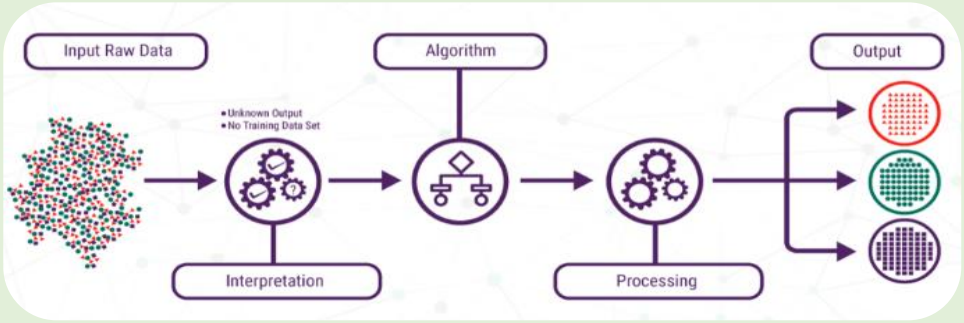
- It is an approach towards creating a model where the model is given labelled input data and the expected output results
- The model learns the relation between the input data and its corresponding labels so that the model gives some prediction on an unseen data
- Example: Classification, Regression, Sentiment analysis

Unsupervised Learning:

- Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses
- Example: Cluster Analysis, Topic Modelling, Dimensionality reduction

Algorithm Selection

Unsupervised Learning



- ✓ Algorithm selection can be done as per modelling requirement from the flow chart
- ✓ For the common algorithms, their characteristics mention under one type of problem will hold good for other problems as well
- ✓ If (no. of features > no. of samples) please get more data before proceeding
- ✓ An extension of OPTICS called OPTICS-OF (OF for Outlier Factor) is used for outlier detection.

* For data too large, the limit beyond which it can be said that the data at hand can be considered large or not completely depends on the problem at hand, type of dataset and other such specific factors. However, to give an example, a dataset containing 100k datapoints can be considered large.

References:

1. https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/
2. https://scikit-learn.org/stable/tutorial/machine_learning_map/
3. https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet

Mini-batch – Mini Batch K-means**, CF-Tree –** Clustering feature tree, **GMM** – Gaussian Mixed Model, **BIRCH** - Balanced iterative reducing and clustering using hierarchies, **K** = Number of clusters, **Good cluster:** Similarities within clusters and dissimilarities between clusters.

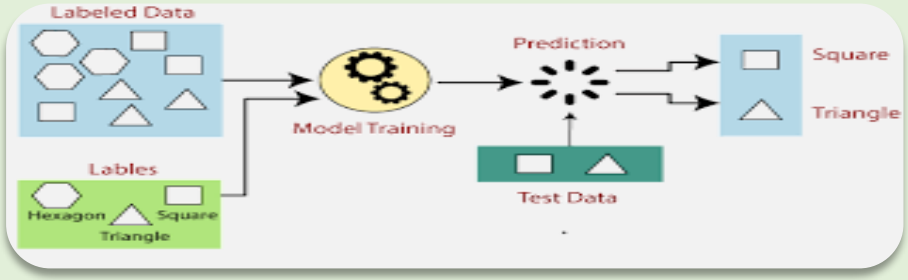
** **Bottom Up** considers local pattern or neighbor points without initially taking into account the global distribution of data while **Top down** takes into consideration the global distribution of data.

Algorithm Selection

Supervised Learning

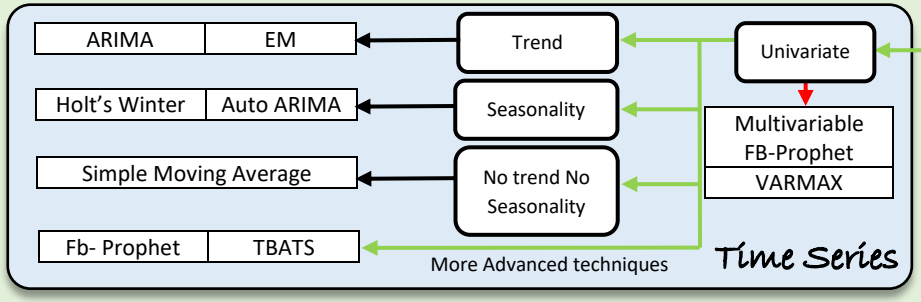
Affine
Data ▸ Insights ▸ Transformation

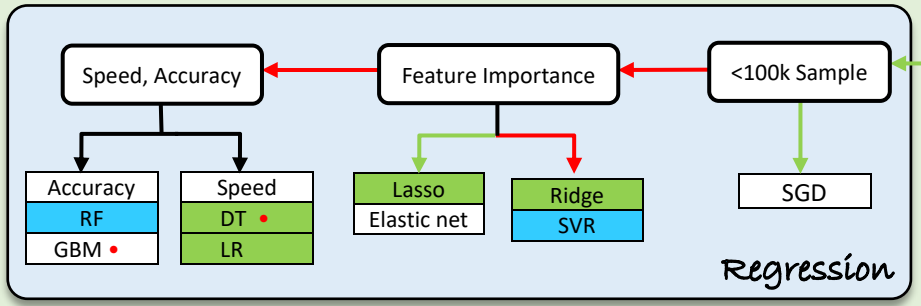
No	Good Interpretability
Yes	Poor Interpretability
• Large Memory Footprint	

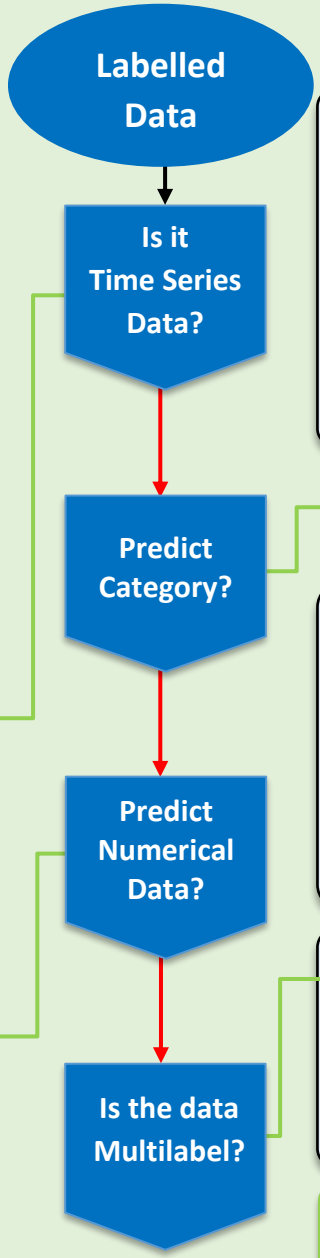


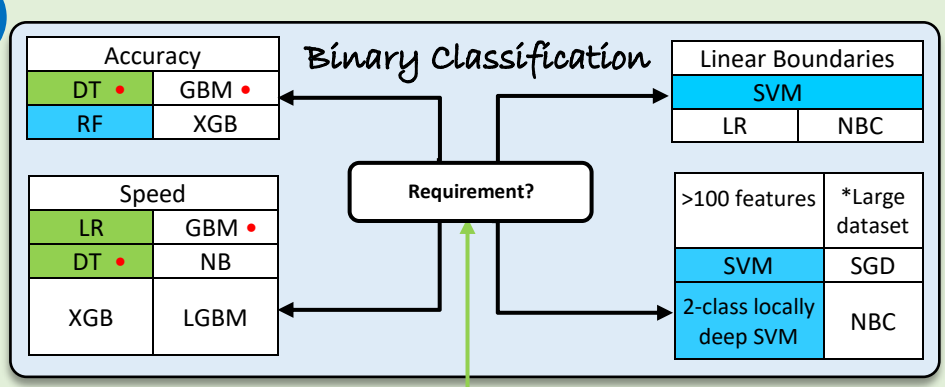
- ✓ Algorithm selection can be done as per modelling requirement from the flow chart (only for labelled data)
- ✓ For the common algorithms their characteristics mention under one type of problem will hold good for other problems as well
- ✓ If you are dealing with Text Classification, please refer to multiclass and binary classification as the algorithms used are same
- ✓ If (no. of features > no. of samples) please get more data before proceeding

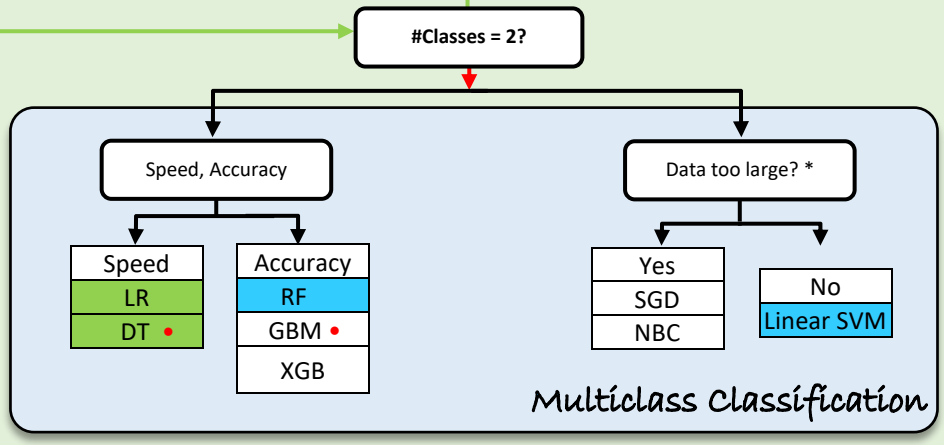
*For data too large, the limit beyond which it can be said that the data at hand can be considered large or not completely depends on the problem at hand, type of dataset and other such specific factors. However, to give an example, a dataset containing 100k datapoints can be considered large.

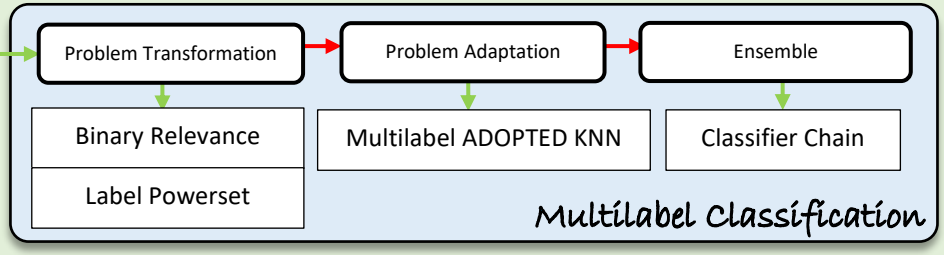












LR- Linear/Logistic Regression, **SVM-** Support vector, **DT-** Decision tree, **RF-** Random Forest, **GBM-** Gradient boosting machine, **SGD-** Stochastic gradient descent, **XGB-** Extreme Gradient Boosting, **LGBM-** Light GBM, **NB-** Naïve Bayes , **EM-**Exponential Smoothening

References: 1. https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/ 2. https://scikit-learn.org/stable/tutorial/machine_learning_map/
3. https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet