معالجة اللغات الطبيعية

Natural Language Processing(NLP)

تطبيق عملي-تهيئة البيانات النصيه ــ ٥

SMS SPAM Filtering

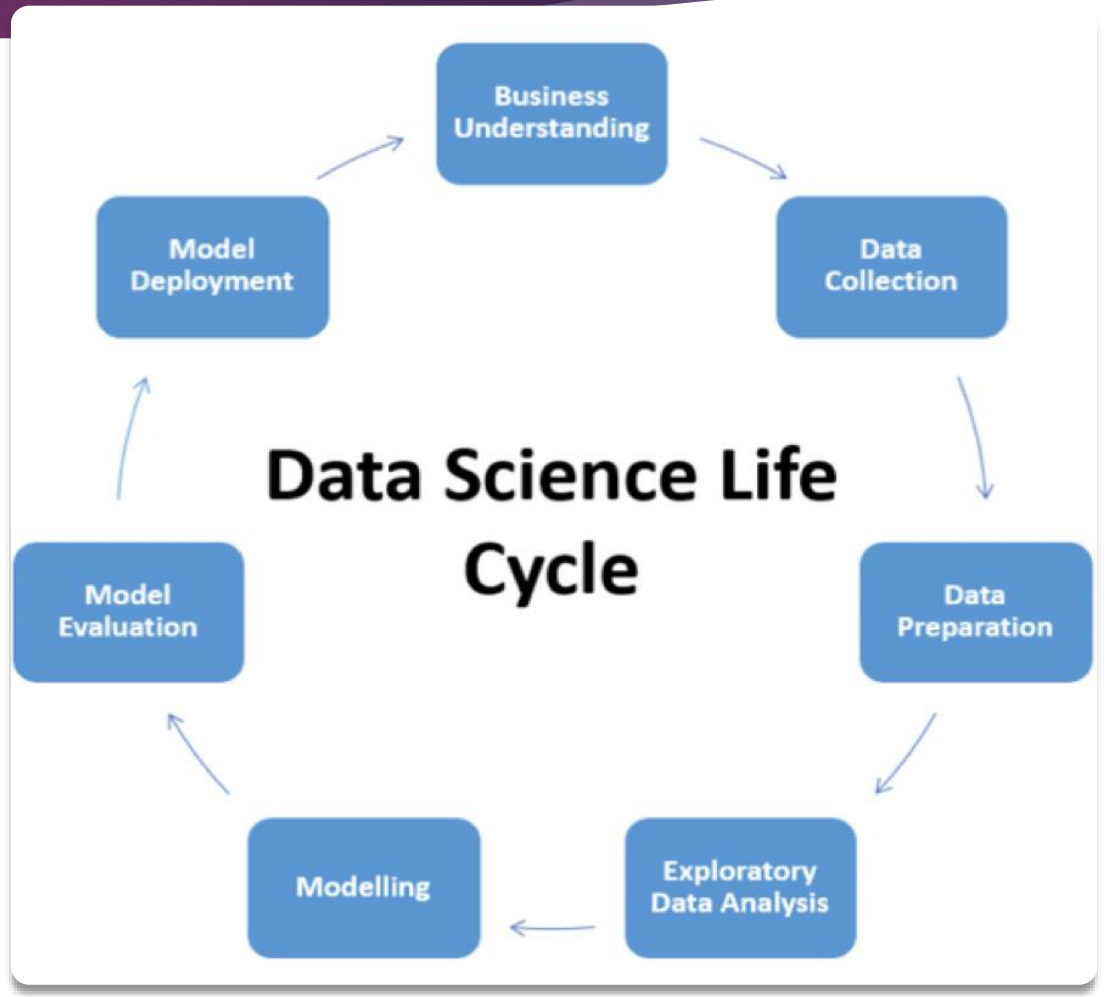
إختيار و بناء النموذج ــخوارزمية تعلم الاله

Model Selection & Building

AHMAD SHHADEH

Data Science Project life cycle

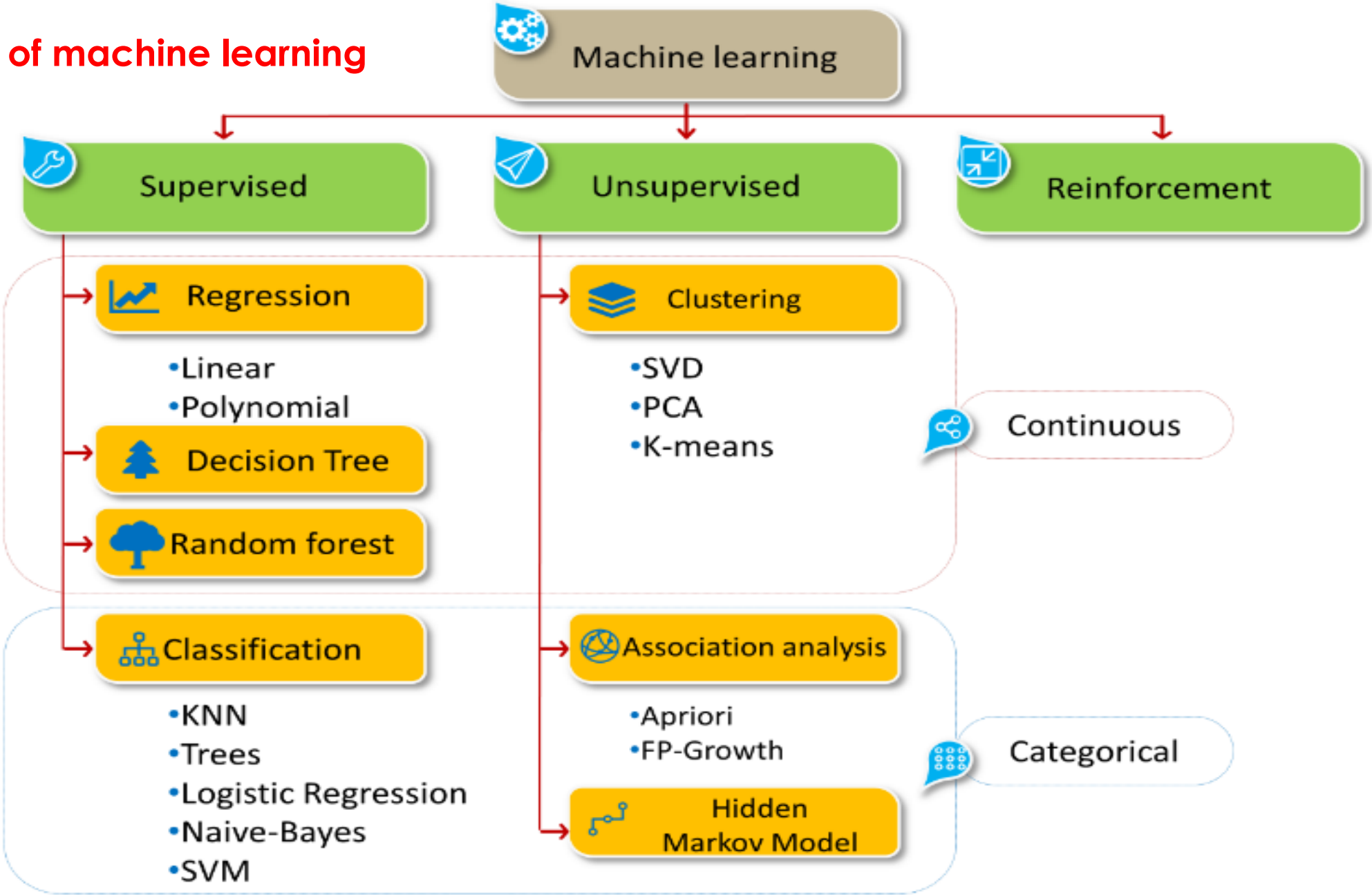
- Business Understanding
- Data Collection
- Data Preparation
- Exploratory data analytics(EDA)
- **Model Building**
- Model Evaluation
- Model Deployment

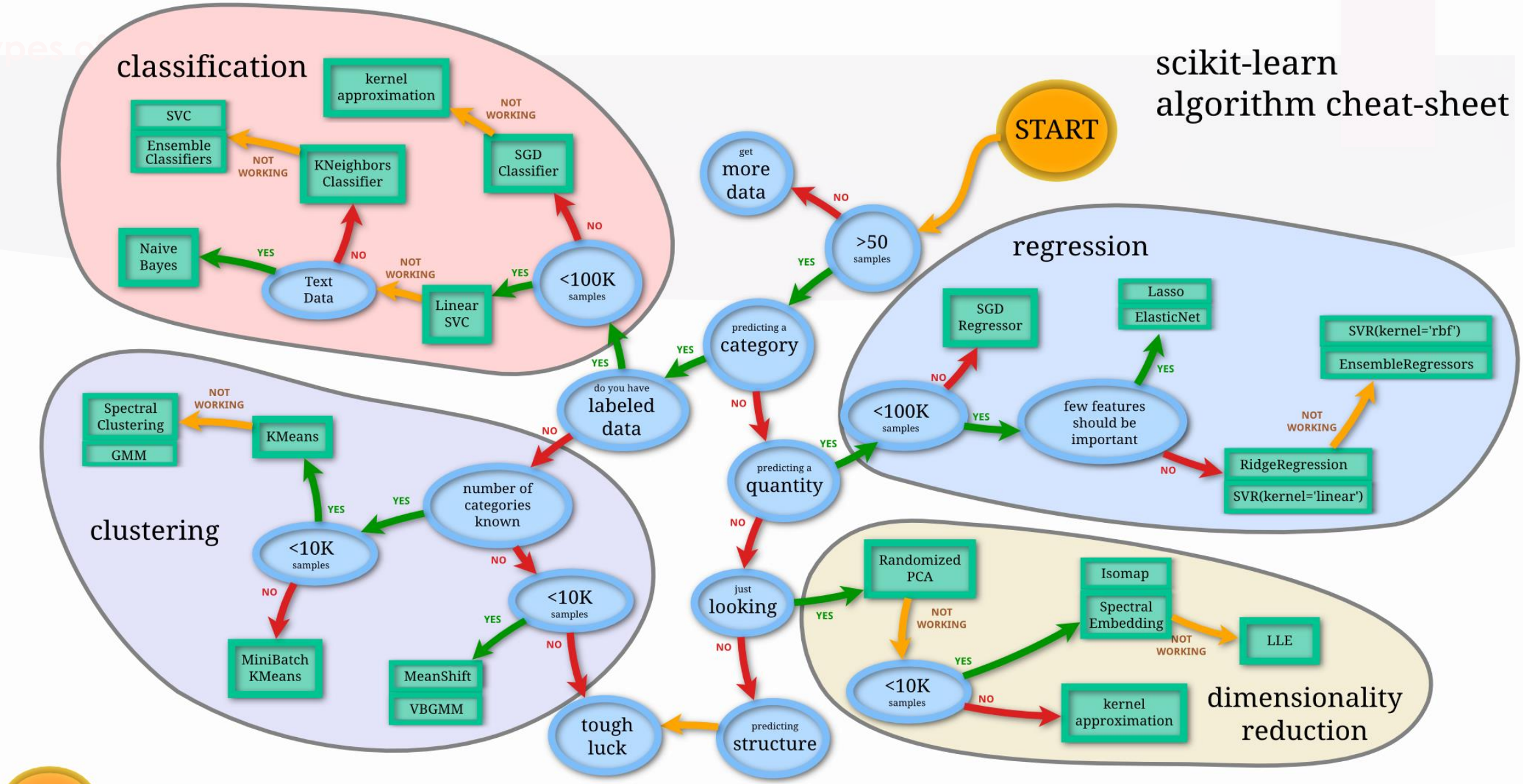


ML algorithm selection

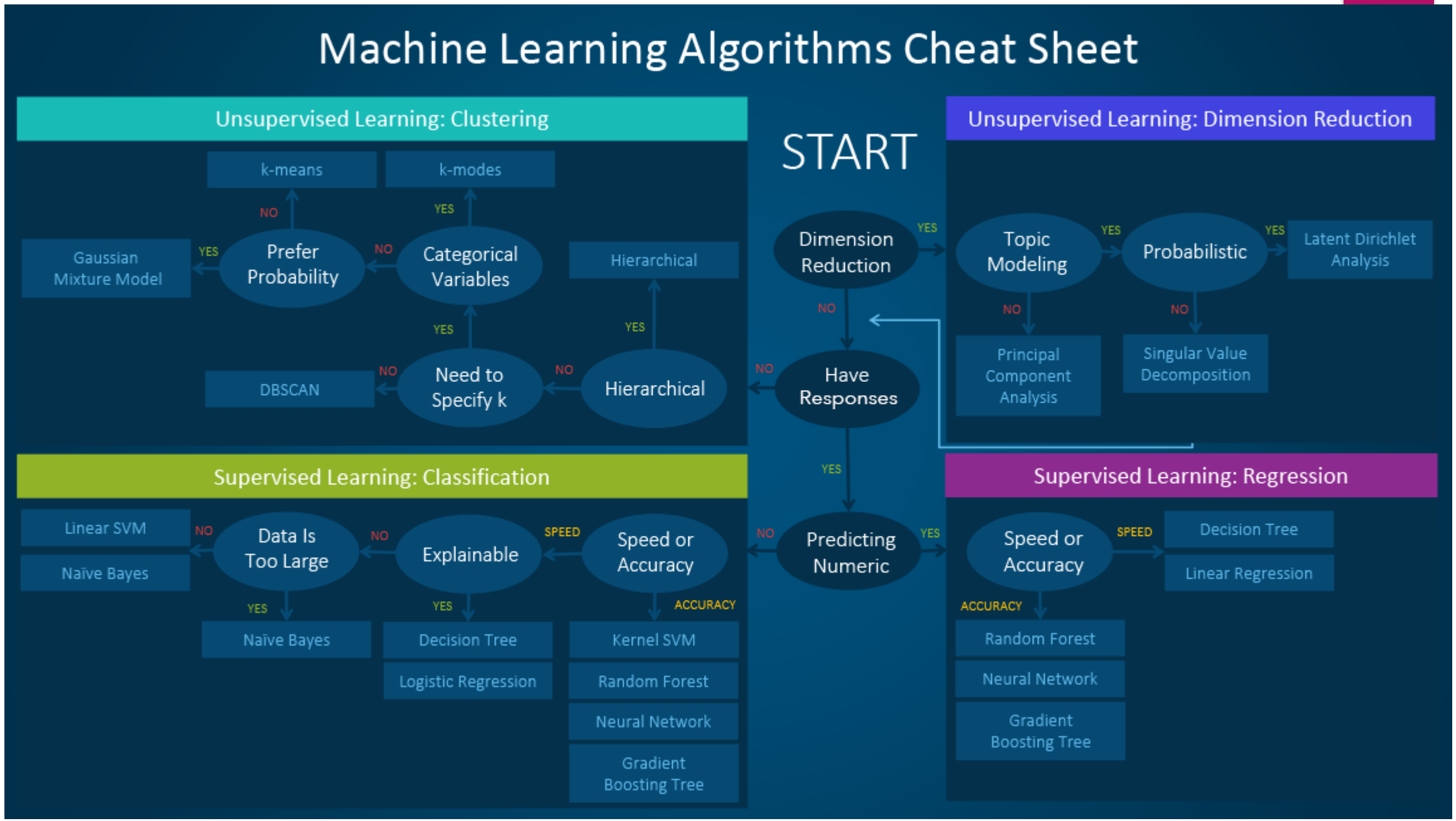
- There are **so many Machine learning** in the world ,Which machine learning Should we use ? There is **no straightforward and sure-shot way** to choose the right MLA. Determining which algorithm to use depends on many factors like:
  - The problem statement ,The kind of output we are looking .
  - Type and size of the data,
  - The available computational time/Resources(Memory, Type of processors),
  - Number of features, and observations in the data
  - …etc
- Key skills can help :
  - Machine Learning Types : Supervised and Unsupervised
  - Domain knowledge to filter down (CV, NLP, anomaly detection .. etc)
  - Data science project Pipeline

Types of machine learning





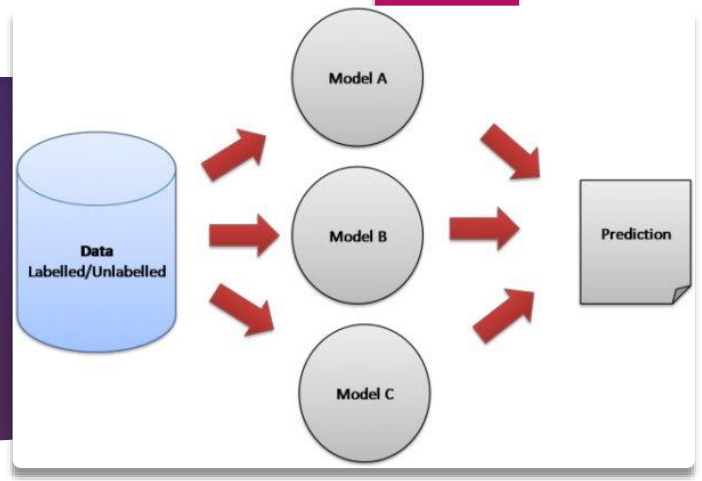
https://scikit-learn.org/stable/tutorial/machine_learning_map/



Source : https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use/

Random Forest

Introduction to Ensemble Learning



- **Ensemble** :A technique that **create multiple models** and then combine them to produce a better result
- ***Why learn one classifier when you can learn many***
- Example: I have created a short movie about machine learning and need to get a feedback before making it public
  - ***1st Model:*** *asking two of my friends*
  - ***2nd Model:*** *asking 5 colleagues on the machine learning domain*
  - ***3rd Model*** *:creating a small survey and get feed back from 20 people*
- The responses, would be more **generalized** and **diversified** since we have people with different skill set and different relationships ,This is a better approach to get honest ratings
- With these examples, you can infer that a diverse group of people are likely to make better decisions as compared to individuals.

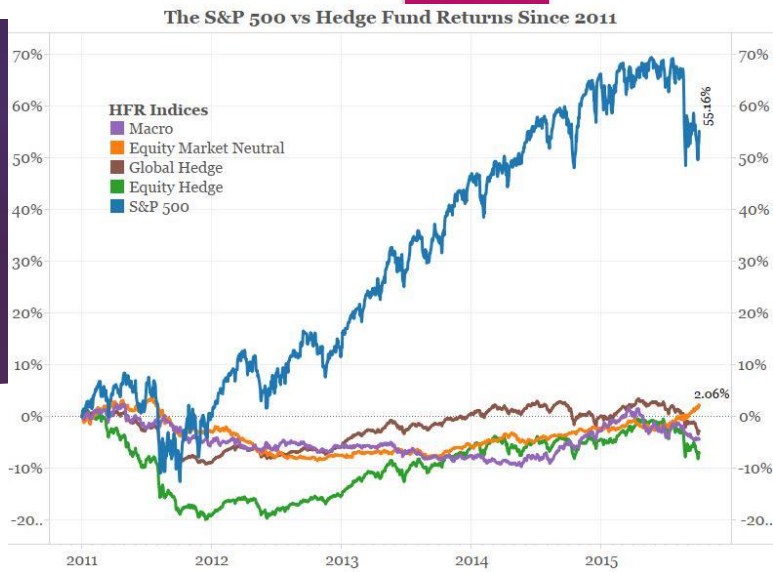
Random Forest in real life

- You want to purchase a new car!!!!!!
  - **will you walk up to the first car shop and purchase one based on the advice of the dealer? OR**
  - **You would browser a few web sites/mobile App (haraj ,souq..etc)**
    - check the posted reviews
    - Compare different car models, checking for their features and prices.
  - **You will also ask some of your friends for their opinion.**

In Summary , you **wouldn't directly reach a conclusion**, but will instead make a decision considering the **opinions of other people** as well.

- Who want to be millionaires?
  - Asking the audience option

Random Forest in real life



- *In the world of finance and investments,*
  - *The basic concepts of the investments is to build a **bunch of uncorrelated models**,*
  - *Each with a **positive expected return**,*
  - *then put them **together in a portfolio** to earn massive alpha (alpha = market beating returns).*
- *Last example for online shopping*
  - *We rely on **multiple sources** (never trust a solitary Amazon review), and therefore, **not only is a decision tree** intuitive, but so is the idea of combining them in a random forest.*

In short : A Multiple number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

What is Random Forest?

- **Random forest** is an ensemble learning method that constructs a **collection of decision trees** and then **aggregates the predictions** of each tree to determine the final prediction.
- **A decision tree** is the building block of a random forest and is an intuitive model.
- Combining the **weak models** which are produced by individual decision trees to get **a strong model**
- **In general**, the higher the number of trees in the forest gives the **high accuracy results**.

Random Forest Characteristics

- **Advantage :**
  - Very **versatile and powerful** machine learning algorithm.
  - Can solve both type of problems i.e. **classification and regression**, Accepts various types of inputs as well, may it be **ordinal or continuous** data.
  - Easily handles **outliers**, **missing values**, **skewed data**, the data doesn't even have to be on the same scale.
  - Less likely to **overfit than** some of the other machine learning models.
  - Providing the feature **importance by identifying the most significant variables** so it can use for dimensionality reduction methods.
- **Disadvantage:**
  - It surely does a good job at classification but **not as good as for regression** problem
  - Random Forest can feel like a **black box approach** for statistical modelers - you have very little control on what the model does.

Random Forest for our project(SMS Spam Filtering)

- *For our project :SMS Spam filtering*
  - We can build a random forest with 100 decision trees in it.
  - Then each decision trees are built independently, and it will predict either **spam** or **ham**.
  - Assume 70 of those decision trees predict **spam** and 30 predict **ham**.
  - Then will apply simple voting method for the trees.
    - Max Voting
    - Averaging
    - Weighted Averaging
  - Then the final prediction of the random forest model will be spam if we applied the **Max Voting** method.
  - In our project will use the simplest voting technique which is **Max Voting**

Max voting method

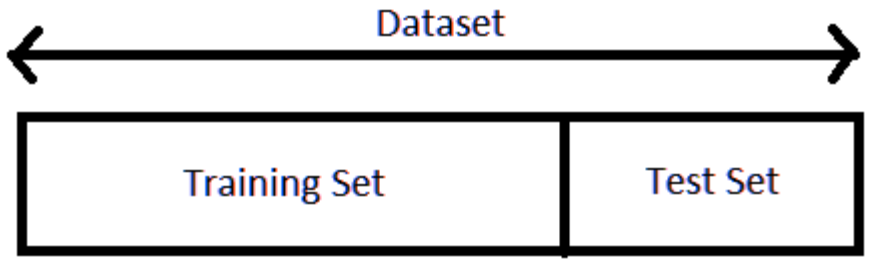
- The max voting method is generally used for **classification problems**.
- In this technique, multiple models are used to make predictions for each data point. The predictions by **each model are considered as a 'vote'.**
- The predictions which we get from the **majority** of the models are used as the final prediction.
- For example, when you asked 5 of your colleagues to rate your movie (out of 5); we'll assume three of them rated it as 4 while two of them gave it a 5. Since the majority gave a rating of 4, the final rating will be taken as 4.

sklearn.ensemble.RandomForestClassifier

Parameter/Attribute/Method	Descriptrion
n_estimators (par)	The number of trees in the forest., default=100.
n_jobs, (par)	default=None ,The number of jobs to run in parallel. fit, predict, decision_path and apply are all parallelized over the trees. None means 1 unless in a joblib.parallel backend context. -1 means using all processors
fit(X_train, y_train) (Method)	Build a forest of trees from the training set (X, y).
predict(X)test) (Method)	Predict class for X.
Feature Importances (Attributes)	indicate what predictor variables the random forest considers most important,it can be used for feature engineering by building additional features from the most important. We can also use feature importances for feature selection by removing low importance features.

Split Training and Testing Data Sets

- ***Machine learning is training an algorithm on a set of known examples with a clear goal of generalizing to unseen examples.***
- The train-test split is a **technique for evaluating the performance** of a machine learning algorithm, It can be used for classification or regression problems and can be used for any supervised learning algorithm.
- The procedure involves taking a dataset and **dividing it into two subsets**.
  - **Train Dataset**: Used **to fit** the machine learning model.
  - **Test Dataset**: Used **to evaluate** the fit machine learning model: not used to train the model
- The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.



Thank You