BDI Python Code Clinic

18/05/2021

More Machine Learning in Python with scikit-learn
Session 2

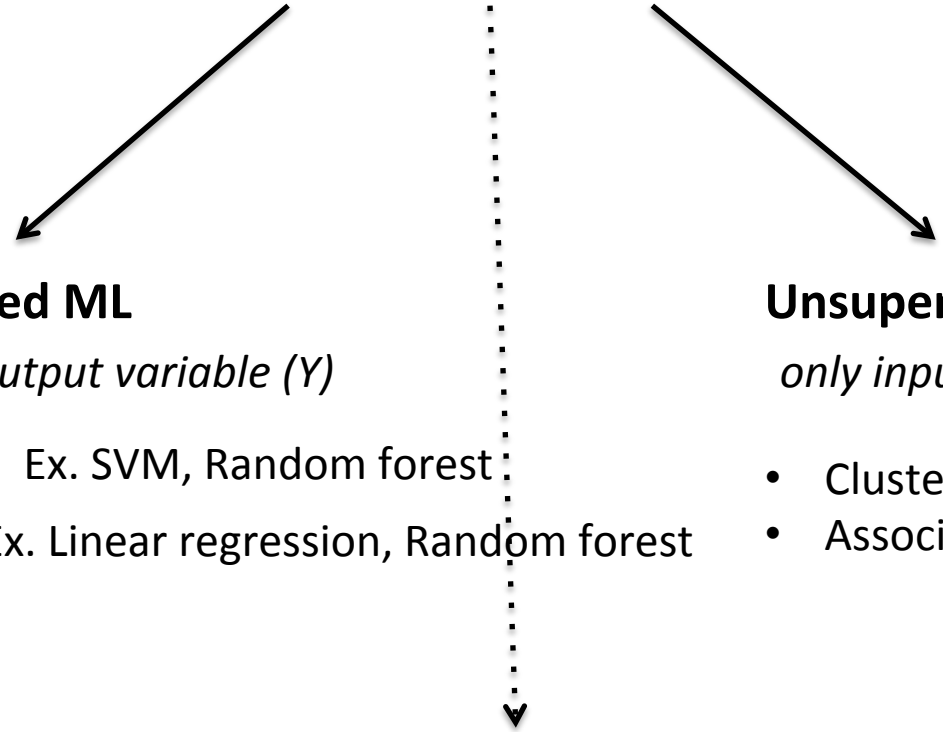
Irina Chelysheva

Before we start…

List of the Python SciPy libraries required for this tutorial:
scipy
numpy
matplotlib
pandas
sklearn

+my Python scripts from
https://github.com/Chelysheva/ML_Python_course

ML problems



Supervised ML

input variables (x) + output variable (Y)

- Classification — Ex. SVM, Random forest
- Regression — Ex. Linear regression, Random forest

Unsupervised ML

only input data (X)

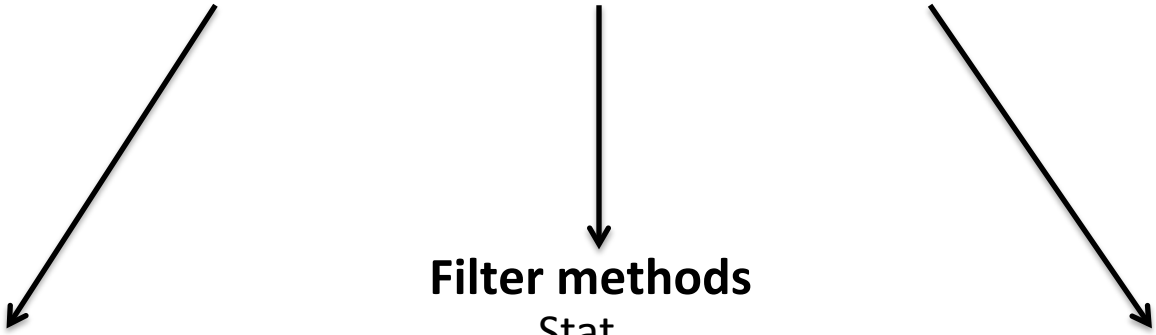
- Clustering — Ex. K-means
- Association — Ex. Apriori

Semi-supervised ML

large amount of input data (X)
+ only some of the data is labeled (Y)

Feature selection

Feature selection is primarily focused on removing non-informative or redundant predictors from the model.
— Page 488, Applied Predictive Modeling, 2013.



Filter methods
Stat

Wrapper methods

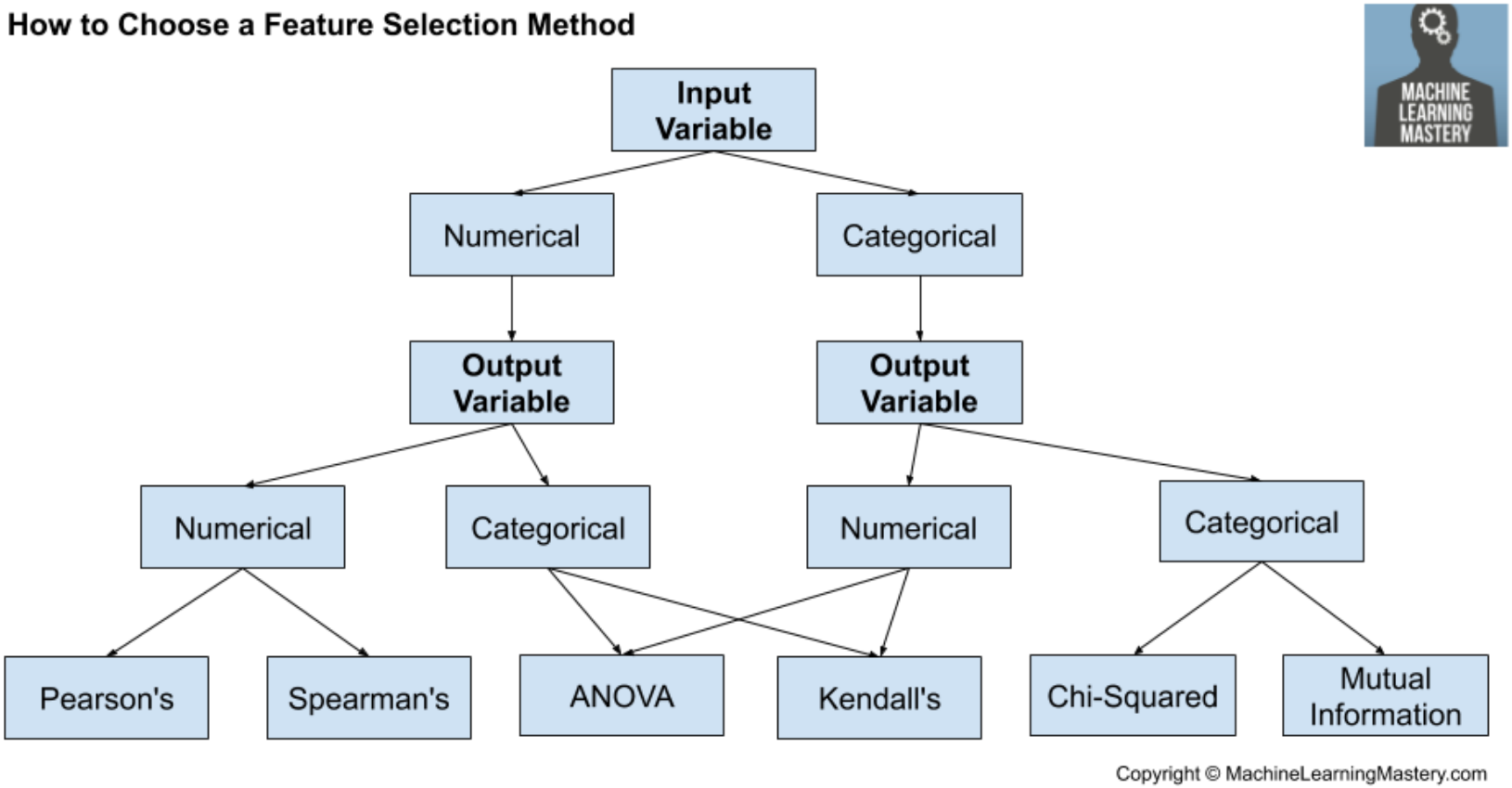
Backward Elimination, Forward Selection, Bidirectional Elimination and Recursive Feature Elimination (RFE)

Embedded methods

Regularization:
LASSO, Elastic Net, Ridge Regression

Filter methods

Image Source: Analytics Vidhya

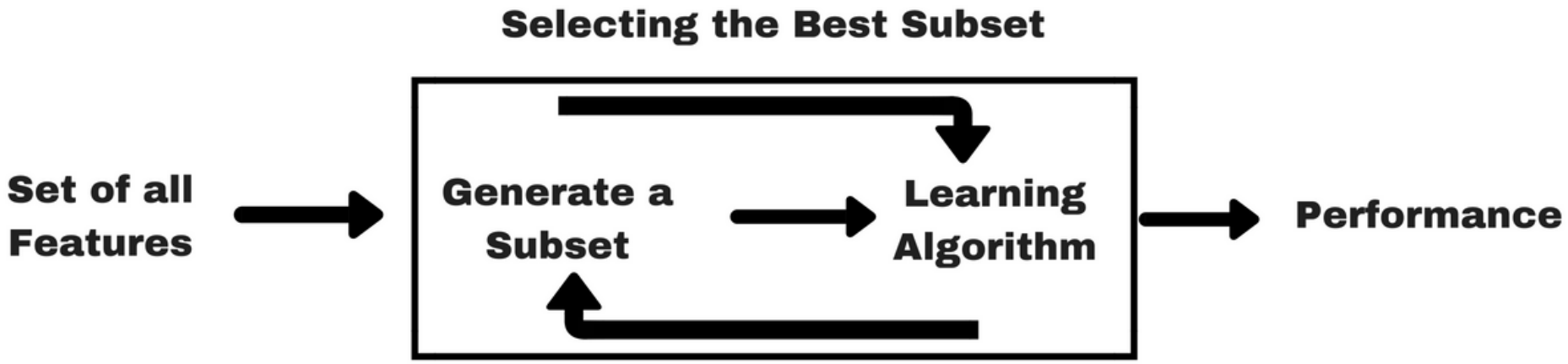


Wrapper methods

A wrapper method needs one machine learning algorithm and uses its performance as evaluation criteria.

Feed the features to the selected Machine Learning algorithm and based on the model performance you add/remove the features.

It is an iterative and computationally expensive process but it is more accurate than the filter method.



Wrapper methods

- **Step Forward Selection**

We start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

- **Backward Elimination**

We start with all the features and remove the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

- **Recursive Feature Elimination (RFE)**

It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

Embedded methods

It takes care of each iteration of the model training process and carefully extracts those features which contribute the most to the training for a particular iteration. Regularization methods (LASSO and Ridge) are the most commonly used embedded methods which penalize a feature given a coefficient threshold.

We will do feature selection using Lasso regularization. If the feature is irrelevant, lasso penalizes its coefficient and make it 0. Hence the features with coefficient = 0 are removed and the rest are taken.

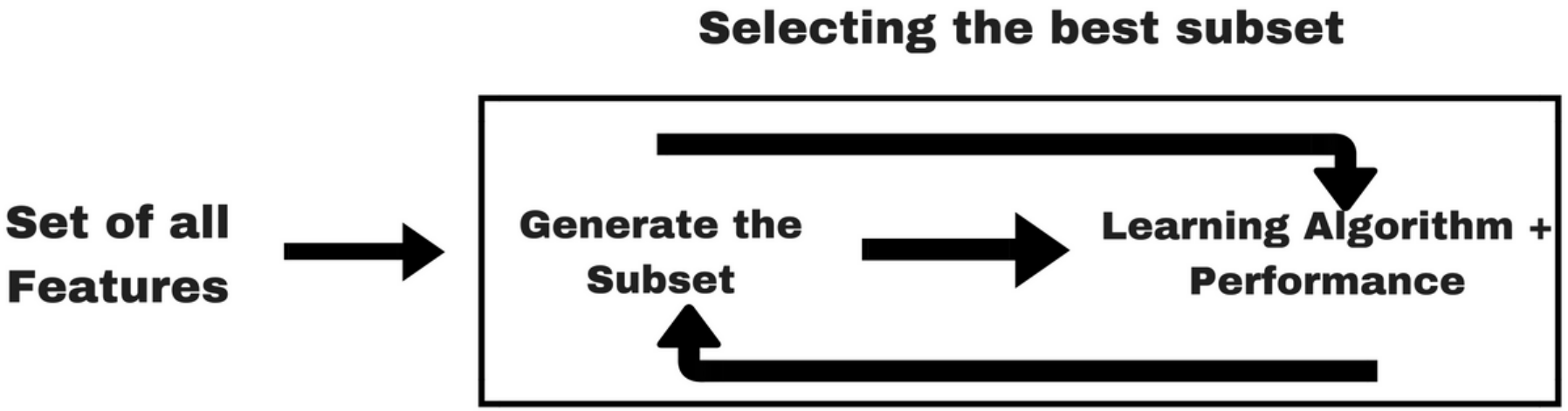


Image Source: Analytics Vidhya

Which feature selection method to use?

Filter method is less accurate. It is great for exploratory data analysis and can also be used for checking multi co-linearity in data. Good for extensive datasets with 100 or 1000s features.

Wrapper and Embedded methods give more accurate results but as they are computationally expensive. These methods are suitable for data with less total number features (<=50).

But there is no single right way!

Famous Boston dataset for regression analysis

- Boston house prices dataset
- Data Set Characteristics:
- Number of Instances: 506
- Number of Attributes: 13 numeric/categorical predictive.
- Median Value (attribute 14) is usually the target.
- Attribute Information (in order): - CRIM per capita crime rate by town - ZN proportion of residential land zoned for lots over 25,000 sq.ft. - INDUS proportion of non-retail business acres per town - CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) - NOX nitric oxides concentration (parts per 10 million) - RM average number of rooms per dwelling - AGE proportion of owner-occupied units built prior to 1940 - DIS weighted distances to five Boston employment centres - RAD index of accessibility to radial highways - TAX full-value property-tax rate per $10,000 - PTRATIO pupil-teacher ratio by town - B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town - LSTAT % lower status of the population
- - MEDV Median value of owner-occupied homes in $1000's

Today's plan

- Dataset: Boston dataset from sklearn
- Regression problem: predicting the "MEDV" column

- Apply different feature selection methods
  - Filter method: Pearson correlation
  - Wrapper methods: Backward Elimination, RFE
  - Embedded Method: LASSO regularization

+ SelectKbest in sklearn

- Select the best ML regression algorithm

4 Linear ML Algorithms:
Linear Regression
Ridge Regression
LASSO Linear Regression
Elastic Net Regression

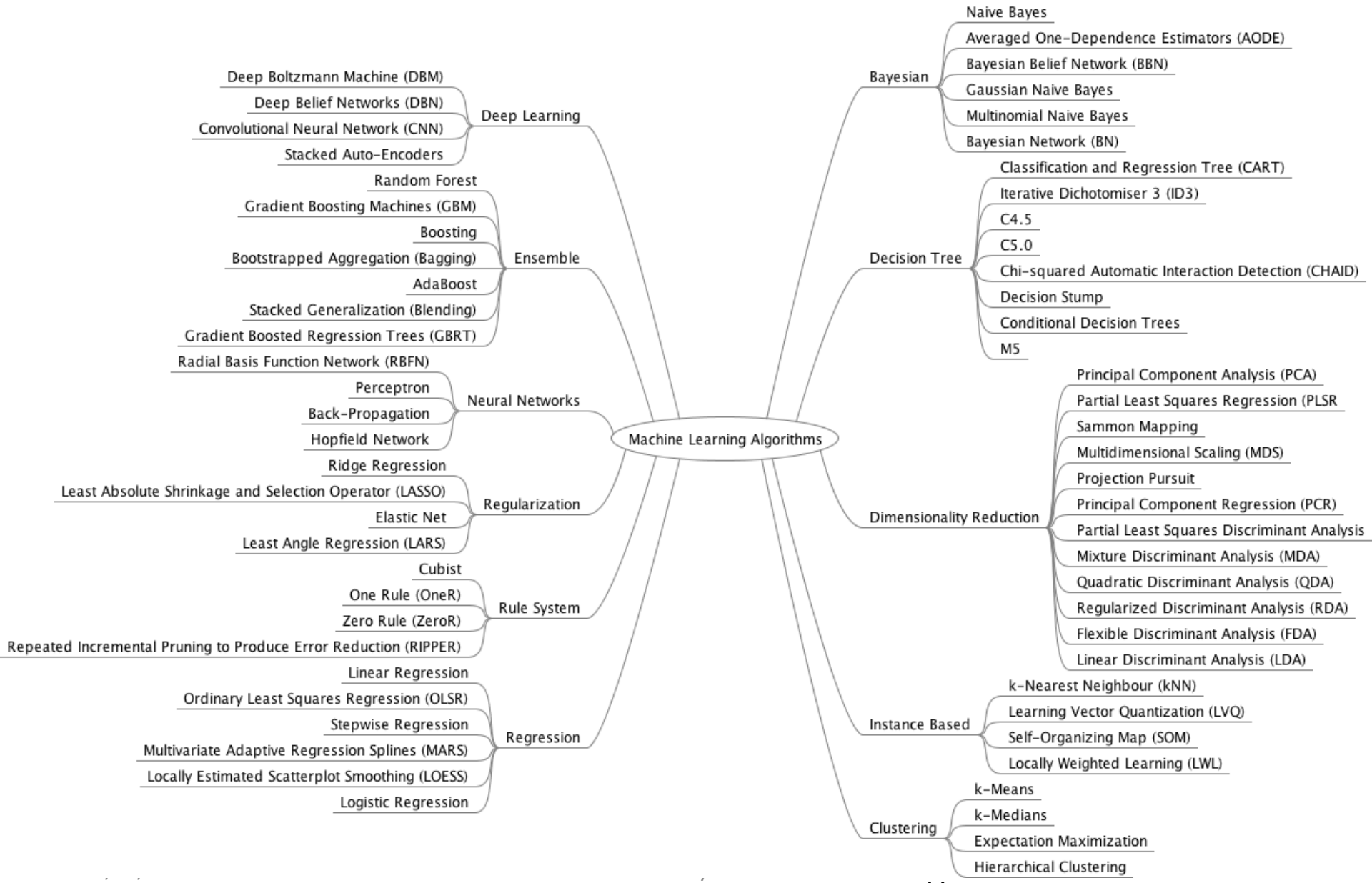
3 Nonlinear ML Algorithms:
K-Nearest Neighbors
Classification and Regression Trees
Support Vector Machines

If you look for practice…

Good, extensive public dataset for regression ML problem:

https://github.com/owid/covid-19-data

Overview of ML algorithms



https://machinelearningmastery.com