Linear Regression

December 2019

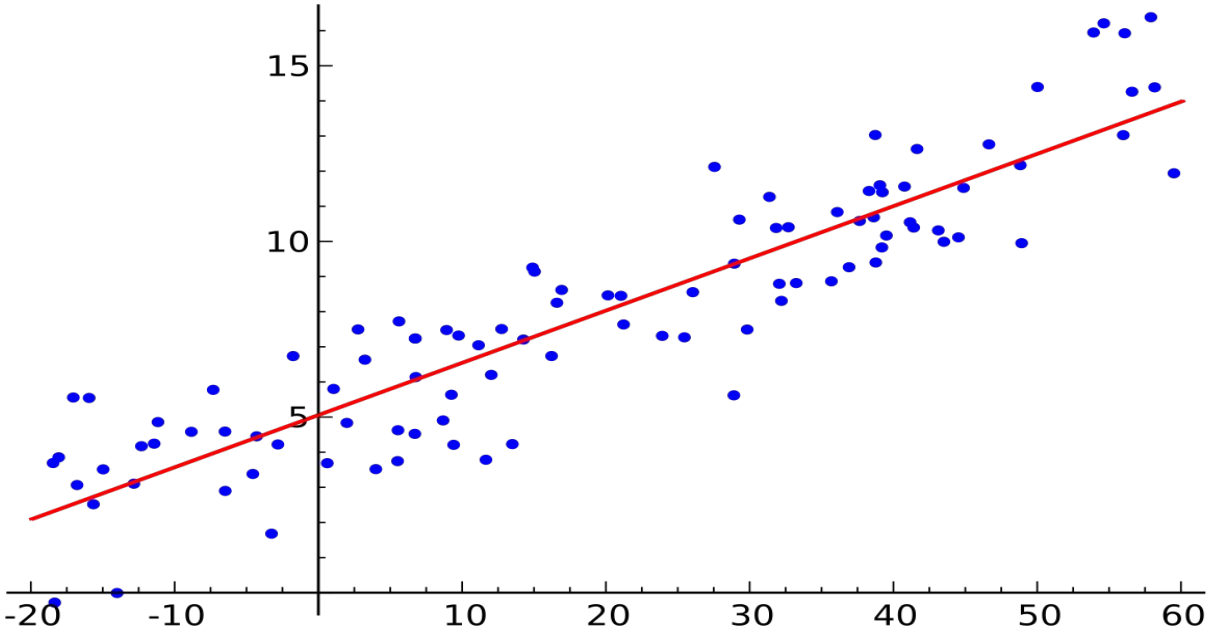
Mark Okello

Regression

- Is a technique used for studying dependence or prediction of Continuous a real variable
- Regression has been used more and more in “analytics”
- There are two types; linear and non linear Regression

Simple Linear Regression

$Y = mX + c$ sse

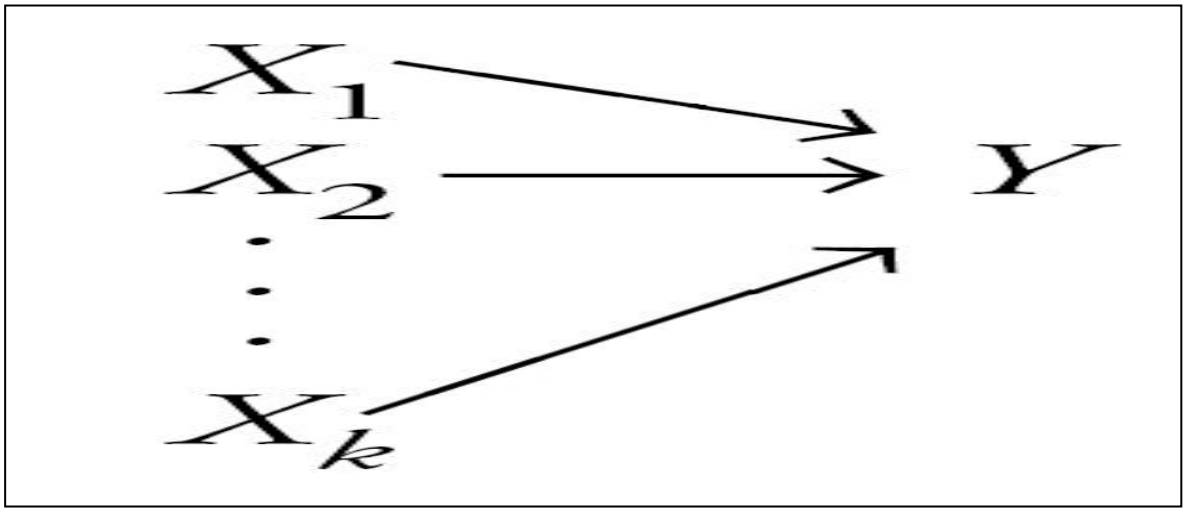


Minimising SSE - Ordinary Least Square Method

$$\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

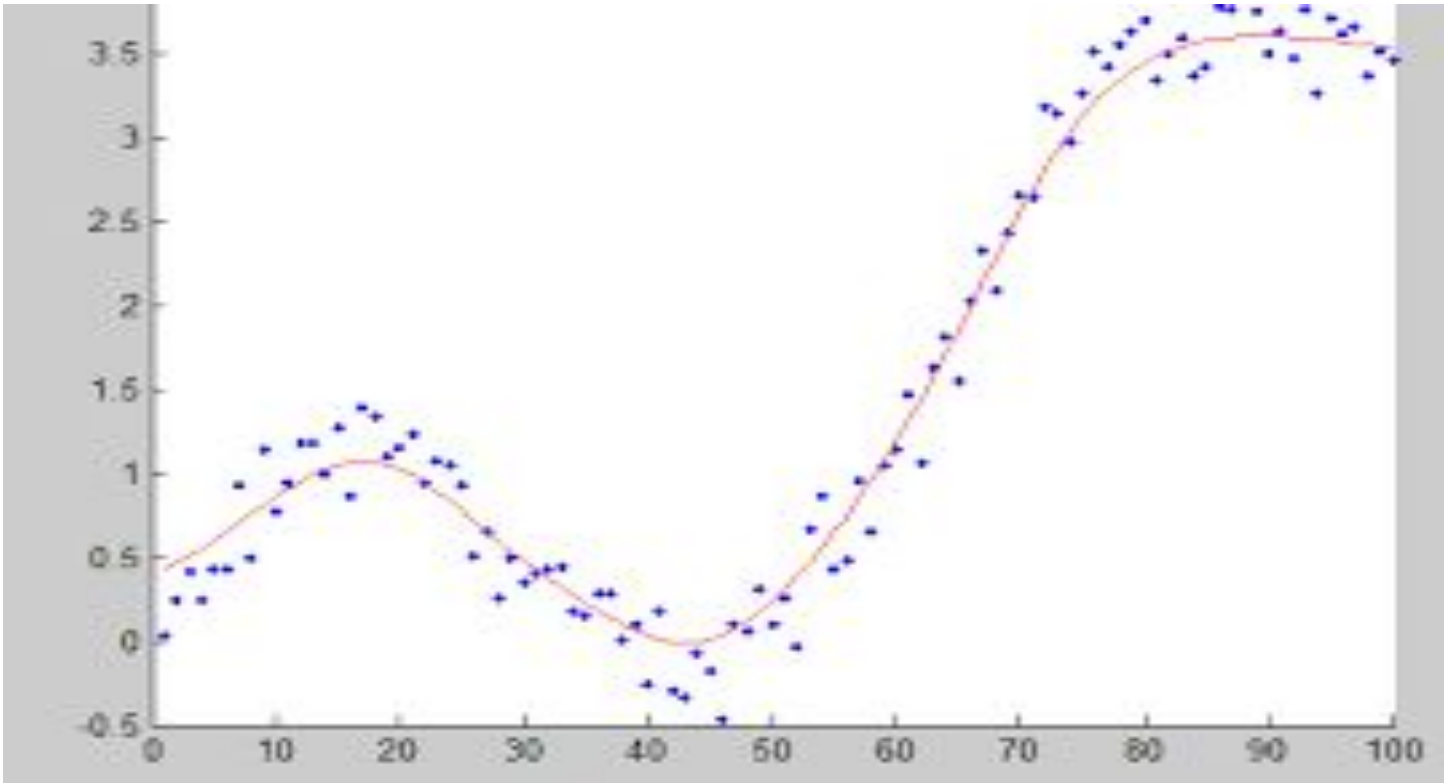
Multiple Linear Regression

$$y = c + m_1 x_1 + m_2 x_2 + \cdots m_k x_k$$

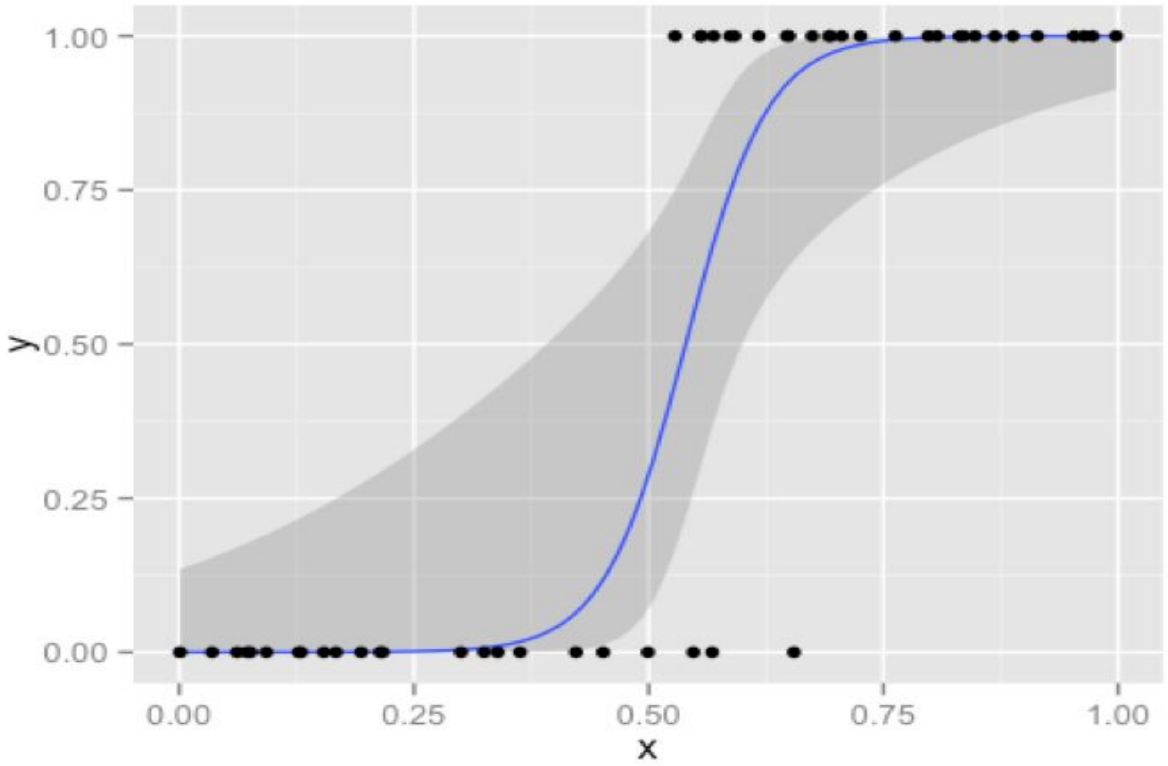


Polynomial Regression

$$Y = C + m_1X + m_2X^2 + \ldots + m_kX^k$$

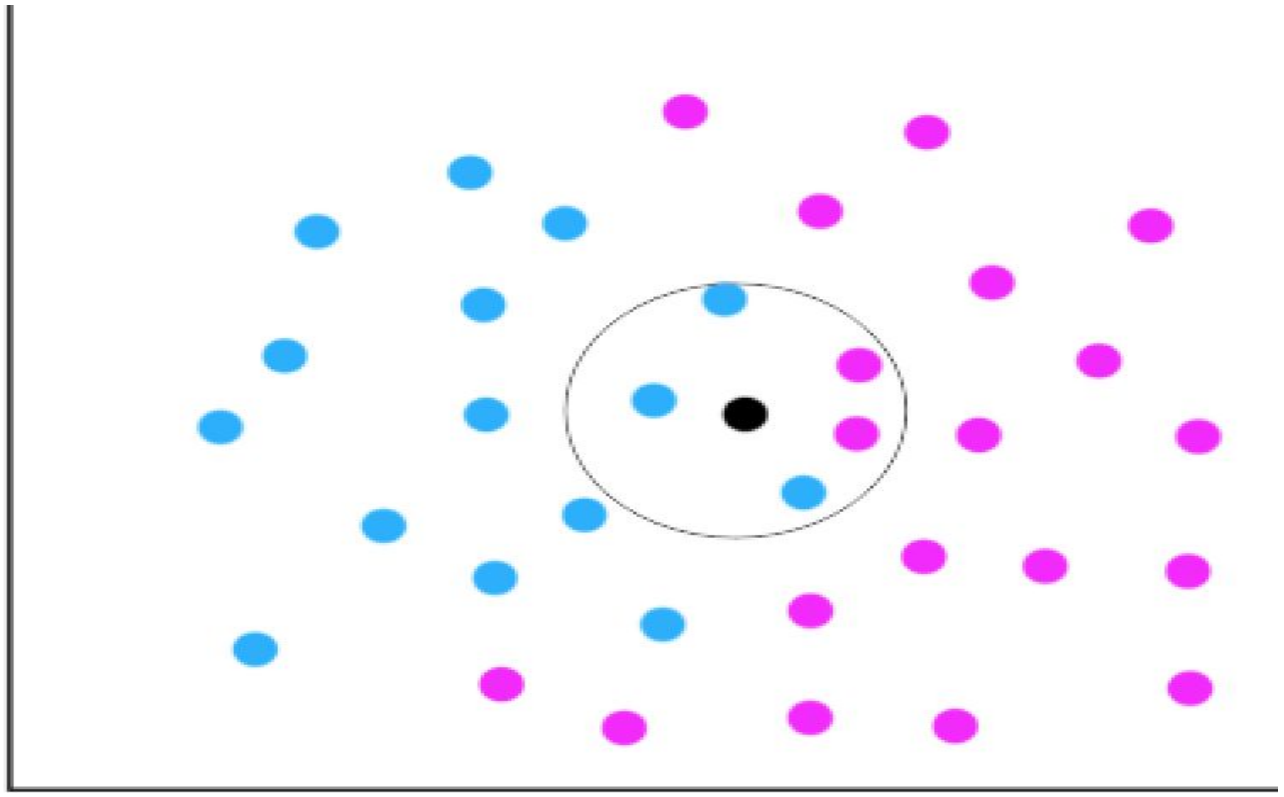


Logistic Regression



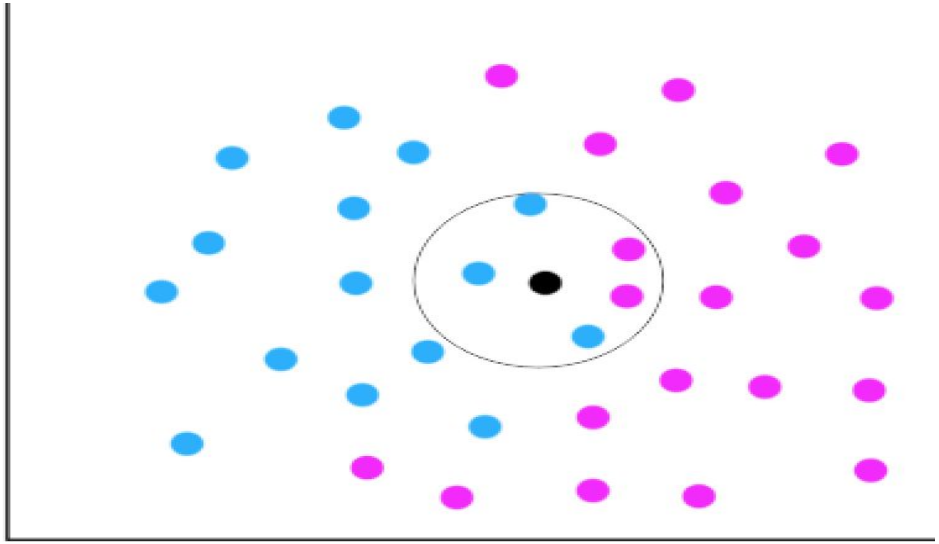
Break

Classification KNN



Naive Bayes

- Bayes Theorem
- P(A / B) = (P(B/ A) * P(A)) / P(B)



Decision Tree

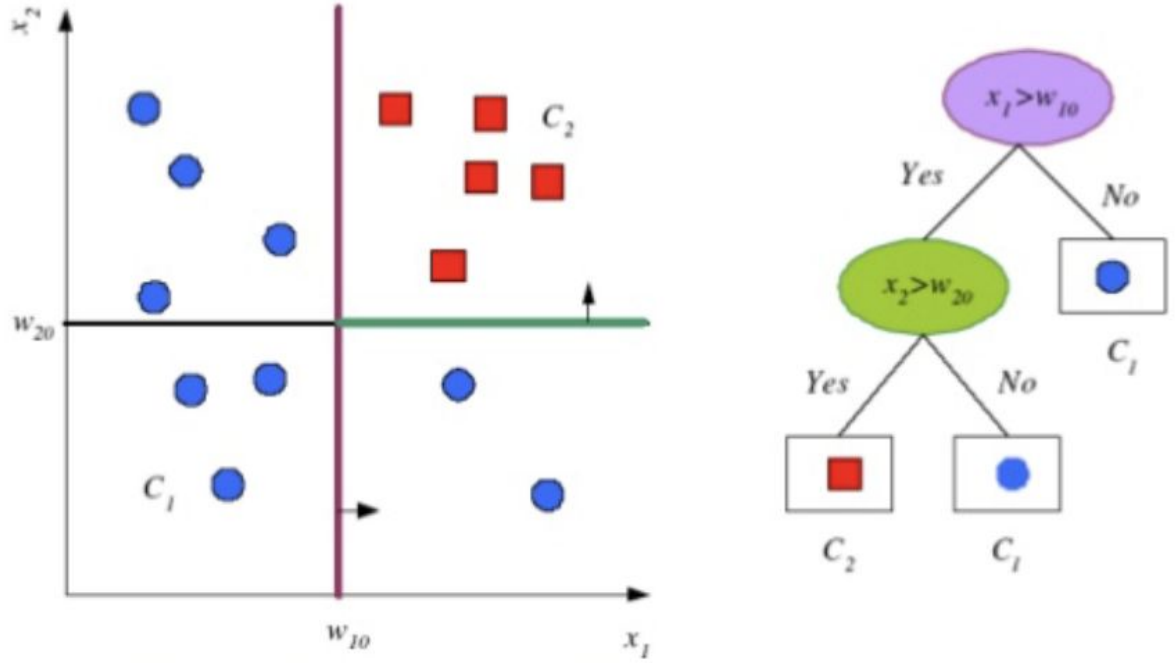
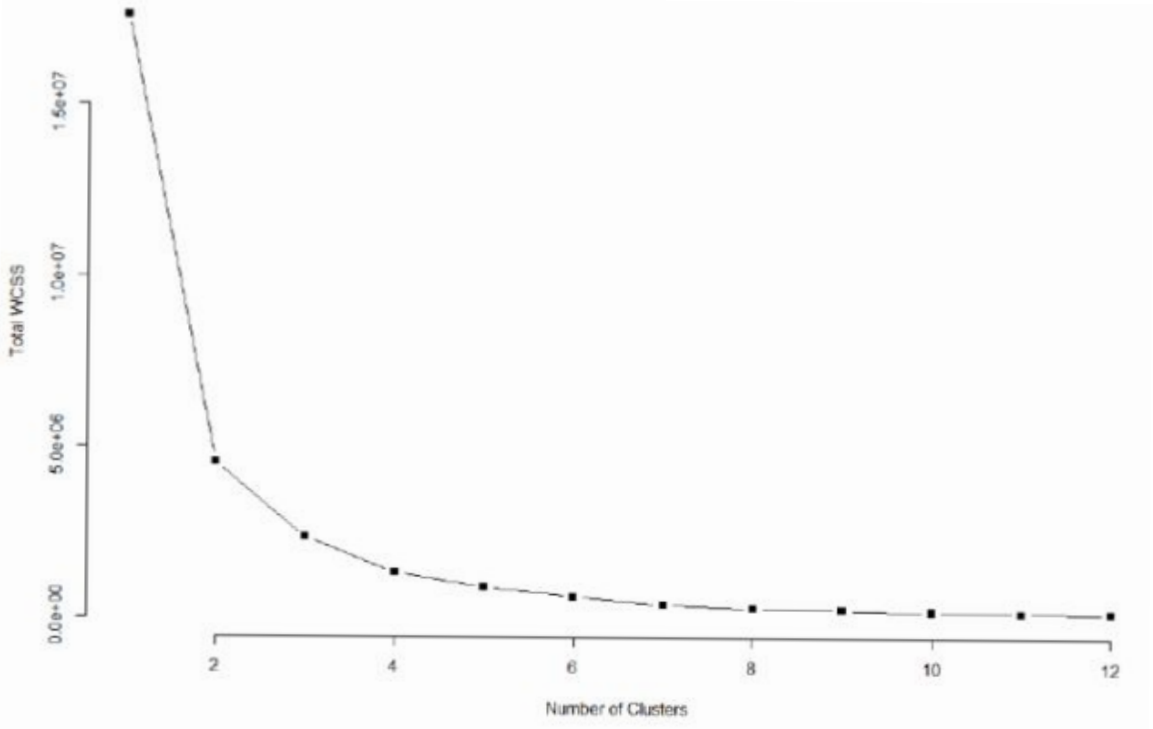


Image Source : Machine Learning for Language Technology

Random Forest

Image Source : Forest Stock 40 by Sed-rah-Stock

Clustering - K Means



Dimensionality Reduction

- DR is a data science method for reducing the number of dimensions(**number of columns in a dataframe**) while maintaining most of the information from the **original set**
- DR is essential when tackling modern DS problems where the number of features is high
- Many problems involve thousands of features for each training instance. Meaning **extremely slow training and much harder** to find a good solution - curse of dimensionality

Why Dimensionality Reduction

- Several features making it harder to train
- Reducing noisy signals
- Improve performance

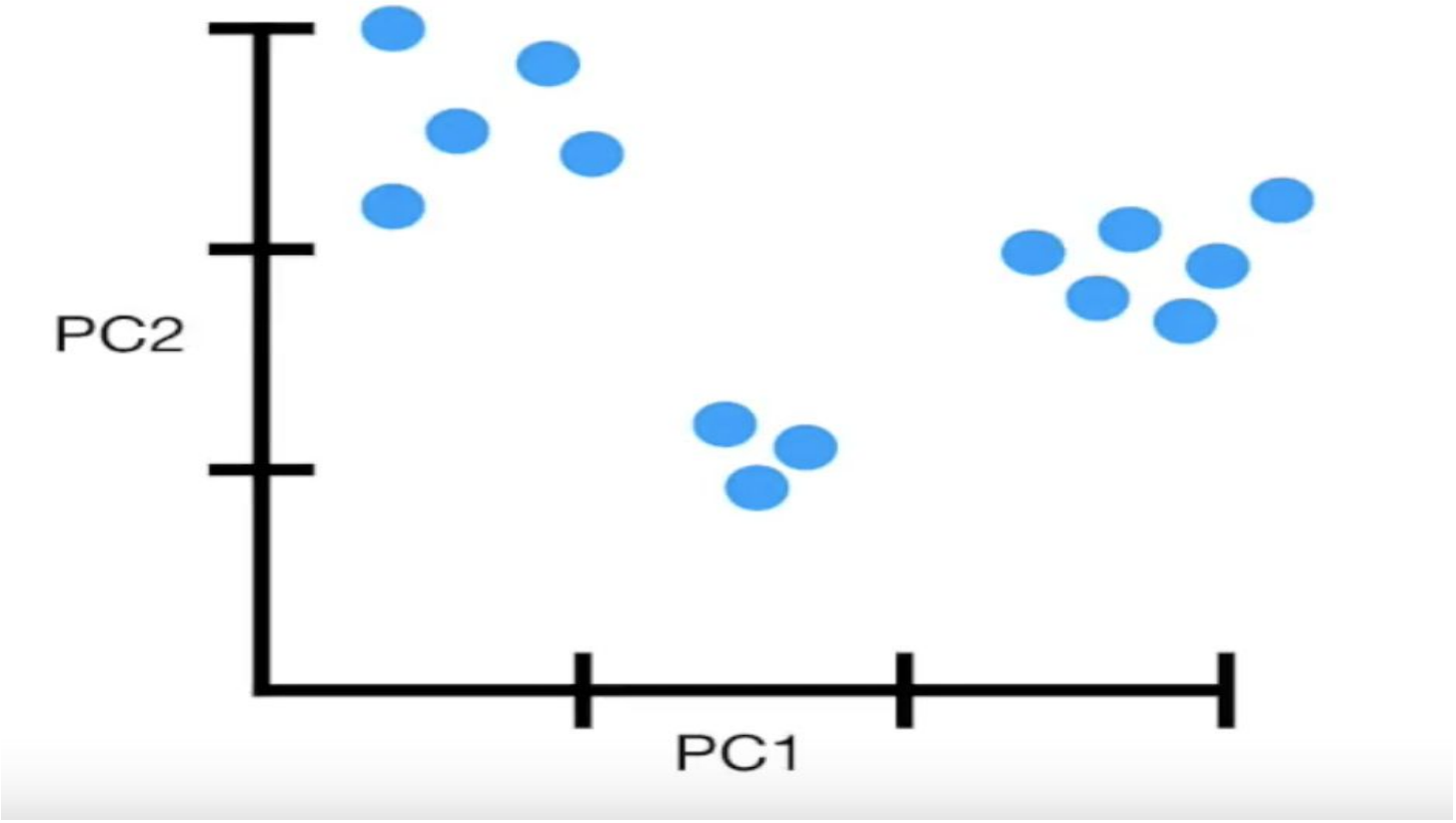
Disclaimer

- Reducing dimensionality Can lead to some information (just like **compressing an image to JPEG can degrade its quality**), so even though it will speed up training, it may also make your system perform slightly worse. It also makes your pipelines a bit **more complex** and thus **harder to maintain**

Feature Selection

- Select a significance level for the model e.g 0.05
- Fit the model with all the variables
- Look for the variable with the highest p value and > significance level, otherwise finish
- Remove that Variable
- Fit the model without that/those variables

Feature Extraction PCA
Highly Correlated Clustered



Practical Process Flow

- Define the problem to solve
- clean, manipulate, understand, feature scale and split the data (data preprocessing or wrangling)
- Train the model
- Visualize the predictions
- Evaluate model performance
- Model Improvement