**Which algorithm to use when:**

**If you have large no. of features or high dimension**: Then Logistic Regression should be choice because it easier to find a plane in high dimension which can divide data points.

**If you have fewer features(20-30):** Then should choose RF or GBDT because tree based model works well for smaller no of features.

**If input data is image:** then should choose CNN.

**If input data is time series or sequence**: then should choose RNN.

**Assumptions:**

**KNN:** class of a neighbour is similar to it’s class.

**Naïve-Bayes**: Features are conditionally independent.

**Logistic Regression:** Dataset is linearly seperable (If not do it via feature eng)

**Correlation matrix vs Covariance matrix:**

If variables are on similar scale then use covariance matrix,

If variable are on different scale then use correlation matrix.

**Multicollinearity test:**

Perbutation test can be done to test multicollinearity within features. Refer Notes 12 of Naïve bayes for description.

**Error Metrics:** which error metrics to use when?

**Classification:**

1. Avg Accuracy
2. Log loss
3. AUC

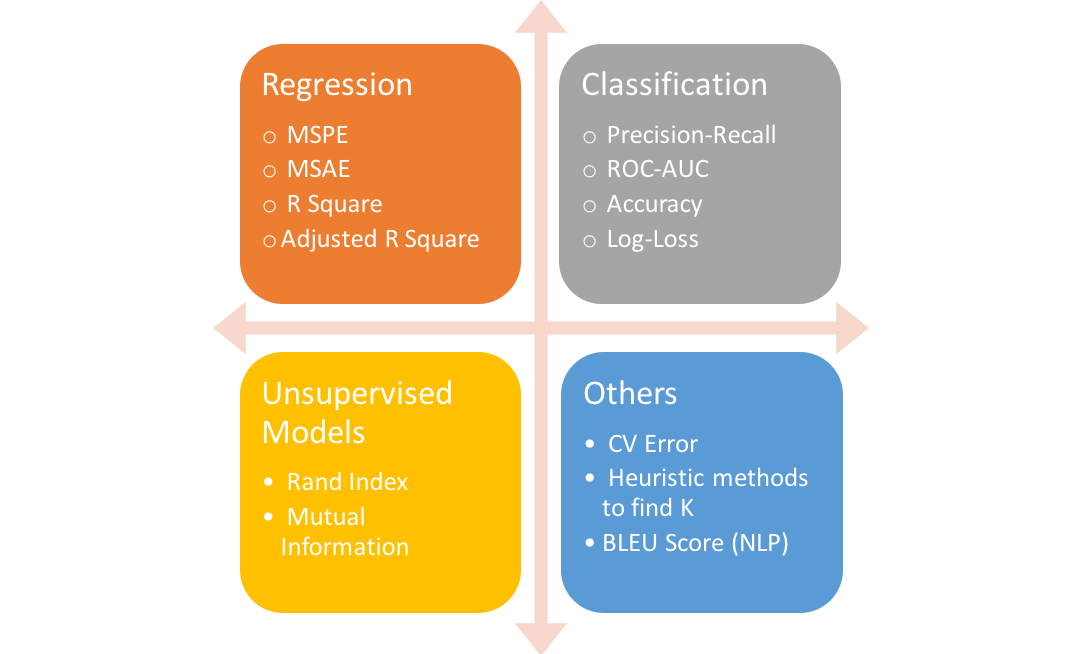
**Regression:**

1. RMSE (Root Mean Squared Error), R2 , MAD

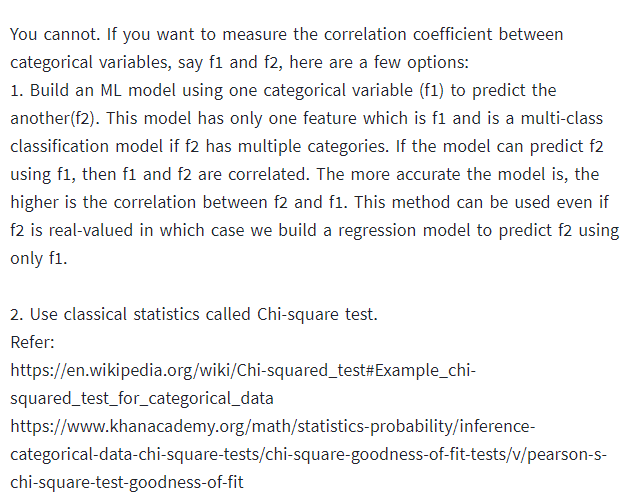
**Ranking:**

1. Precission and Recall

**for the unbalanced data set we usually use AUC , F1 score  Jaccard index metrics, for balanced we use log loss, accuracy**



**Test for collinearity among categorical or categorical and numerica features.**



<https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests/chi-square-goodness-of-fit-tests/v/pearson-s-chi-square-test-goodness-of-fit>

<https://www.mathsisfun.com/data/chi-square-test.html>

**Outliers:**

1. IQR (inter quartile range)
2. Kertosis : more the kurtosis more the chances of outliers to be there
3. Local outlier Factor(LOF)
4. Using Percentile Values

**Testing Distribution of Any feature:**

* QQ plot
* Ks-test

Difference between outliers and noise: <https://www.quora.com/What-is-the-basic-difference-between-noise-and-outliers-in-Data-mining>

Kaggle solutions:

House sales prediction: nice read

<https://www.kaggle.com/juliencs/a-study-on-regression-applied-to-the-ames-dataset>

Recommendation System

<http://blog.kaggle.com/2017/09/21/instacart-market-basket-analysis-winners-interview-2nd-place-kazuki-onodera/>

Good Github Repo

<https://github.com/dipanjanS>

Deep learning IIT madras

<https://www.youtube.com/watch?v=4TC5s_xNKSs&list=PLZEpw4xZOspJEn5s4wUj6uA1mw2_dk0ap>

to add in 15 feature eng.

