Model Performance

* Feature Selection
* Cross Validation
* Dimensionality Reduction

**Feature selection**

* Are all features important?
* What is it?
  + Selecting the best feature to predict the dependent variable
* Why?
  + Improves the model accuracy
  + Reduce the training time
  + Reduce computational cost
  + Solve the problem of overfitting / underfitting
* How?
  + Univariate Selection
    - Analyze each feature and get top 10
    - SKlearn.feature\_selection 🡪 selectKbest
  + Feature Importance
    - We analyze the importance of feature
    - We plot the top 10 features
  + Correlation Matrix (Heatmap) (-1 to +1)
    - Positive
    - Negative

Cross Validation

* We split the data into train and test (eg. 70 : 30 )
* We make sure that the model is properly trained
* 2 Class – 0 and 1
* How?
  + Holdout validation Approach – Train\_Test\_Split
  + Basic K-Fold Cross Validation
  + Cons – In case of classification (Imbalanced Dataset)
  + Stratified Cross Validation
    - It makes sure that it takes all class in a good proportion
  + Leave one out cross Validation (LOOCV)
    - It takes a longer time
  + Repeated Random Test-Train-Split (ShuffleSplit)
    - It is a hybrid of Train-Test-Split and K-Fold

Dimensionality Reduction

* What is it?
  + Dimensions – No. of features/inputs/independent variables
  + Reducing the No. of features
* Why?
  + Training time decreases
  + Model Accuracy will improve
  + Get better insights (visualizations)
* How?
  + PCA – Principal Component Analysis
    - We convert multiple features in small chunks
    - They will linearly uncorrelated features by maintaining the same variation
    - These Chunks are called Components