**SCIKIT-LEARN TUTORIAL**

**Scikit-learn** is a powerful and easy-to-use Python library for classification, dimensionality reduction

* **Supervised learning**: classification, regression (e.g., SVM, Random Forest)
* **Unsupervised learning**: clustering, dimensionality reduction
* **Model evaluation** and preprocessing

**pip install Scikit-learn**

**MODELS FOR LEARNING**

* LogisticRegression (for classification problems, particularly binary classification)
* KNeighborsClassifier (for similarity between data points)
* RandomForestClassifier (on multiple subset of dataset create multiple decision)
* GaussianNB (data follows Gaussian distribution / normal distribution , probability of event to happen)
* Svm (separate data into different classes based on fine/ optimal hyperplane)
* train\_test\_split ( divides the data into training , testing phase )
* accuracy\_score (to calculate the accuracy of datasets)

**FEATURES SELECTION MODELS**

* **Select k best** : Select features according to the k highest scores. (k : int or "all", default=10)
* **Select Percentile:** Select features according to a percentile of the highest scores (percentile : int, default=10)
* Percent of features to keep.
* **Select FDR :** Select the p-values for an estimated false discovery rate. (alpha : float, default=5e-2)
* The highest uncorrected p-value for features to keep
* **Select FPR :** Select the pvalues below alpha based on a FPR test. It controls the total amount of false detections.(alpha : float, default=5e-2)
* Features with p-values less than alpha are selected.
* **Select FWE :** Select the p-values corresponding to Family-wise error rate. (alpha : float, default=5e-2)
* The highest uncorrected p-value for features to keep.

WE CANNOT USE FDR, FWE, FPR SELECTION METHOD WITH MUTUAL\_CLASSIF SCORING TECHNIQUE BECAUSE IT DOESNOT SUPPORT IT

**SCORING TECHNIQUES**

* **f\_classif :** ANOVA F-value between label/feature for classification tasks.
* **chi2 :** Chi-squared stats of non-negative features for classification tasks.
* **f\_regression :** F-value between label/feature for regression tasks.
* **mutual\_classif:** Mutual information (MI) 1 between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.