CSE 537.01 ARTIFICIAL INTELLIGENCE

PROJECT TOPIC:
TRAINING CLASSIFICATION MODEL FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS

TEAM NO: 23
TEAM MEMBER: NEEL PARATKAR
SBU ID: 111483570

ABOUT THE DATA:

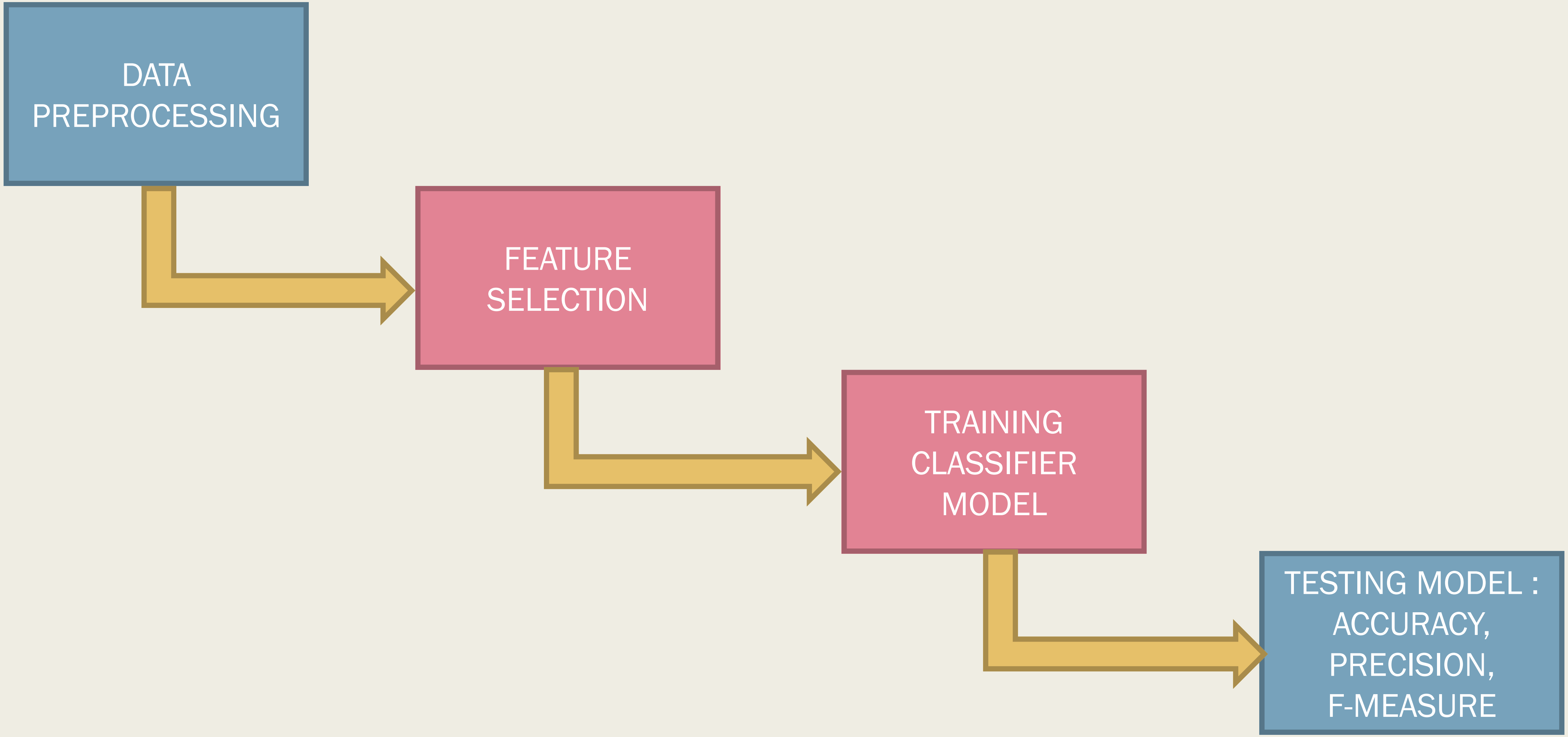
- Title: Wisconsin Diagnostic Breast Cancer (WDBC) dataset
- Creators: Dr. William Wolberg, W. Nick Street, Olvi L. Mangasarian
- Date: November 1995
- Source : https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/
- Number of Instances : 569
- Number of Attributes: 32 (ID, diagnosis, 30 real-valued input features)
- Attribute Information:
  - *ID Number*
  - *Diagnosis (M – Malignant, B – Benign)*
  - *3-32: real valued features*
- Ten real valued features:
  - *Radius (mean distances from center to points on the perimeter)*
  - *Texture (standard deviation of gray-scale values)*
  - *Perimeter*
  - *Area*
  - *Smoothness (local variation of radius length)*
  - *Compactness (perimeter^2/ area – 1)*
  - *Concavity (severity of concave portions)*
  - *Concave points (number of concave portions)*
  - *Symmetry*
  - *Fractional*
- Class Distribution: 357 Benign, 212 Malignant

CLASSIFICATION MODELS

- NAÏVE BAYES:
  - *SIMPLE TECHNIQUE TO CONSTRUCT CLASSIFIERS*
  - *ASSIGN CLASS LABELS TO INSTANCES WHICH ARE REPRESENTED AS VECTOR OF FEATURES*

- LINEAR DISCRIMINANT ANALYSIS:
  - *LDA ATTEMPTS EXPRESS VARIABLES AS LINEAR COMBINATION OF OTHER FEATURES OR MEASUREMENTS*
  - *USED TO FIND A FEATURE SPACE TO PROJECT DATA IN ORDER TO MAXIMIZE CLASS SEPARABILITY*

- SUPPORT VECTOR MACHINES:
  - *SUPERVISED LEARNING MODELS*
  - *USE KERNEL TRICK FOR NON-LINEAR CLASSIFICATION*
  - ***SUPPORT VECTOR CLUSTERING WITH RBF KERNEL FUNCTION USED FOR CLASSIFICATION IN THIS PROJECT***



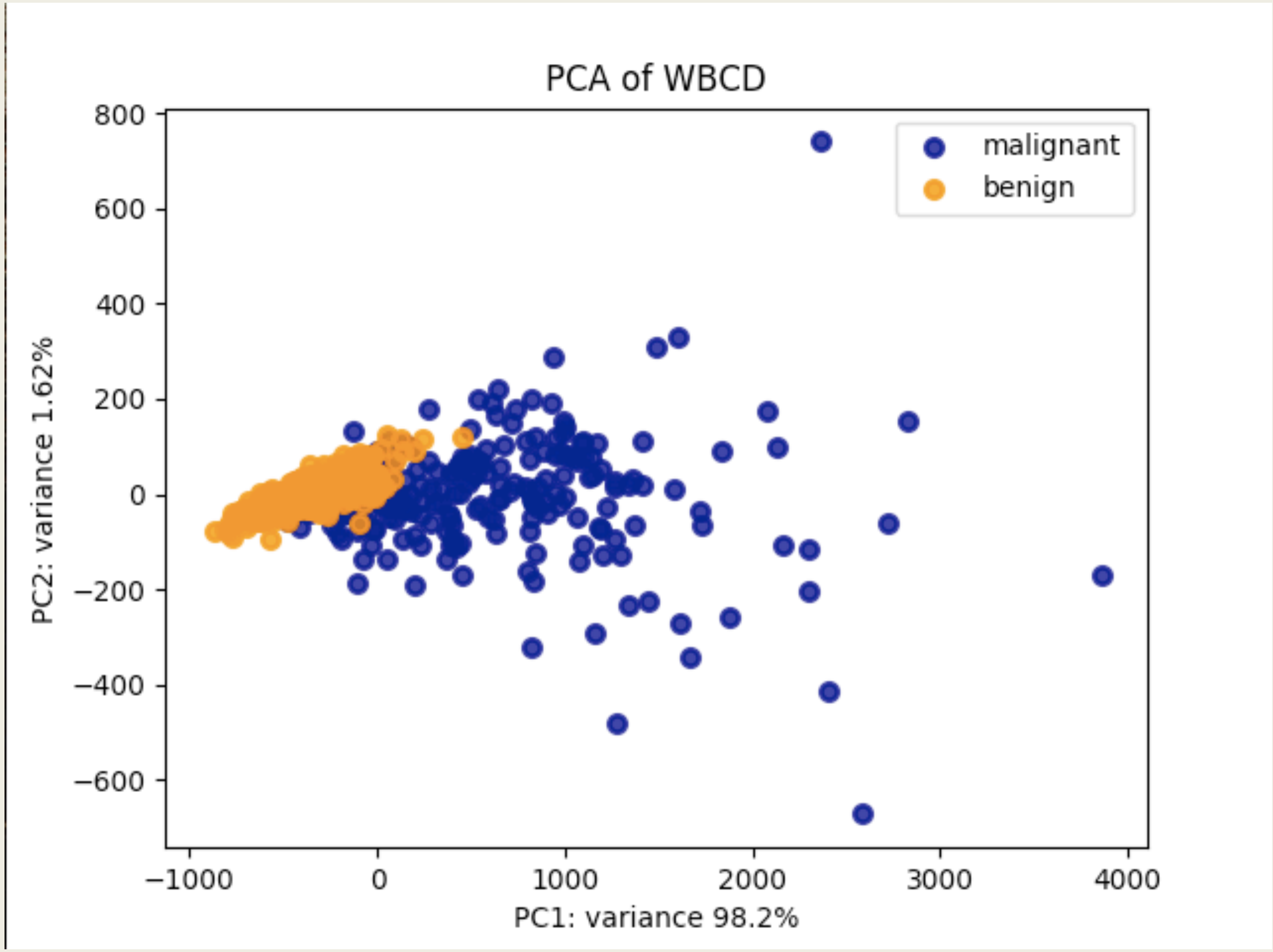
DATA PREPROCESSING

FEATURE SELECTION

TRAINING CLASSIFIER MODEL

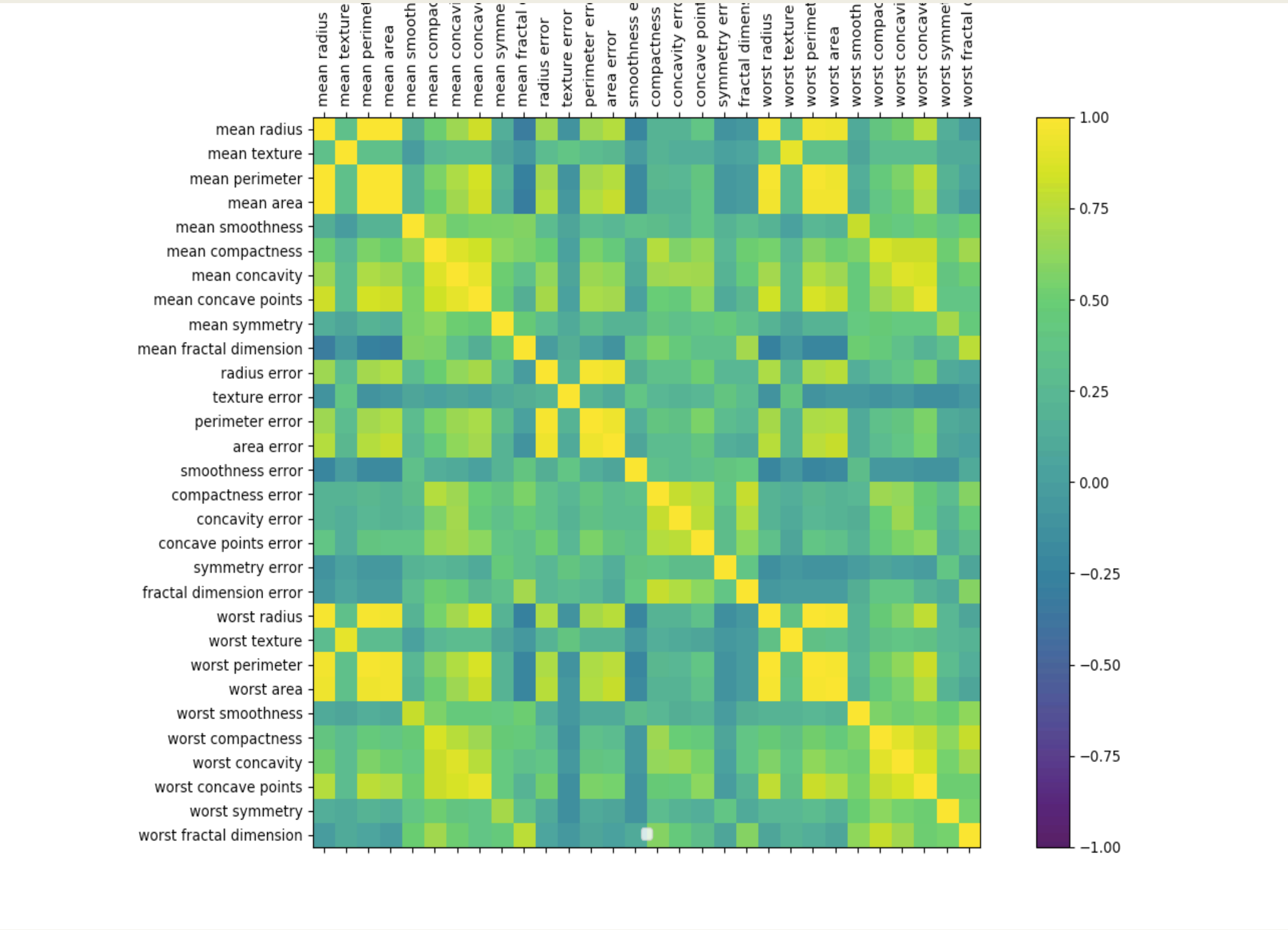
TESTING MODEL : ACCURACY, PRECISION, F-MEASURE

PRINCIPAL COMPONENT ANALYSIS



- *PC1 Variance : 98.2 %*
- *PC2 Variance : 1.62 %*

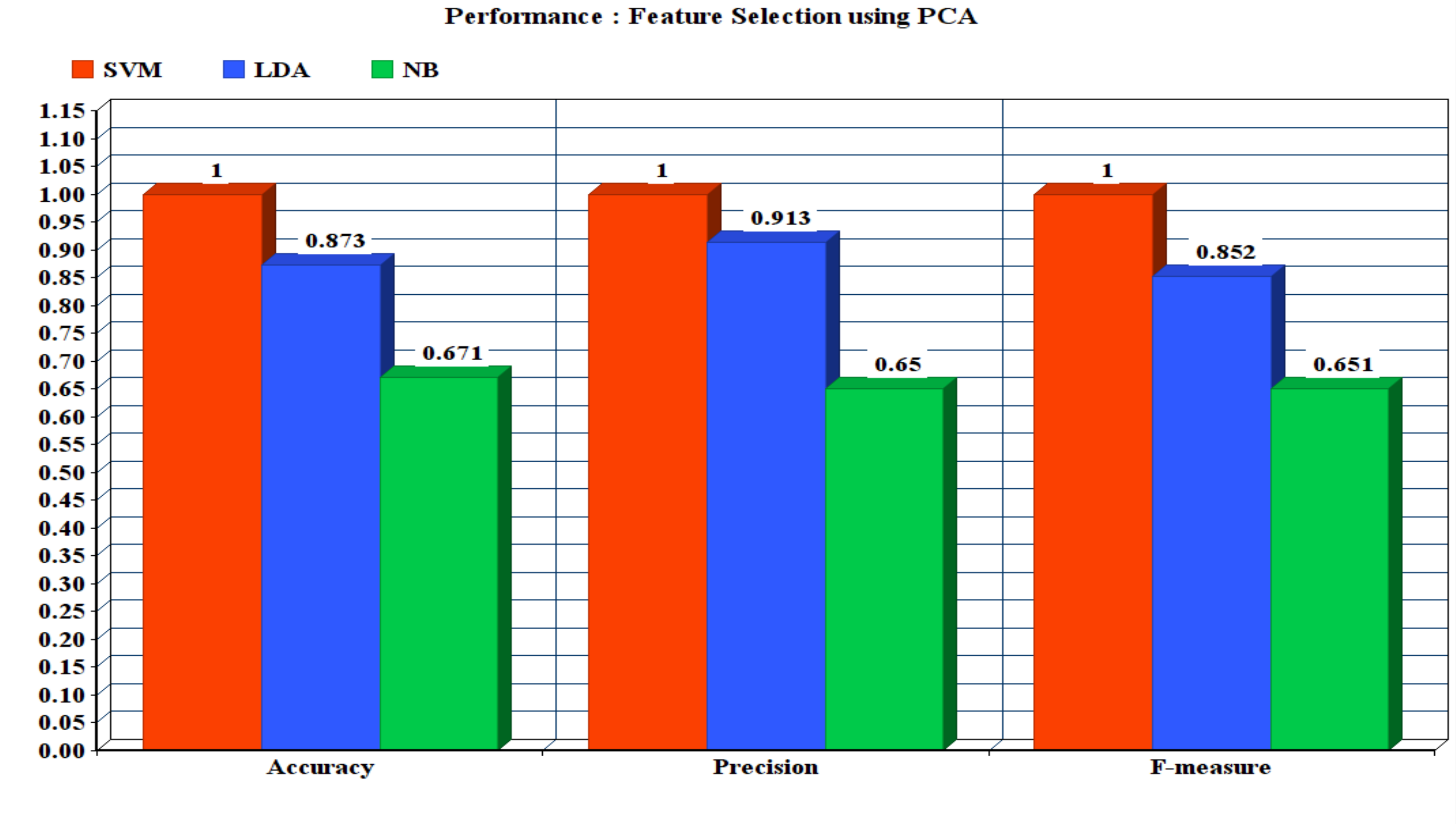
CORRELATION BASED FEATURE SELECTION



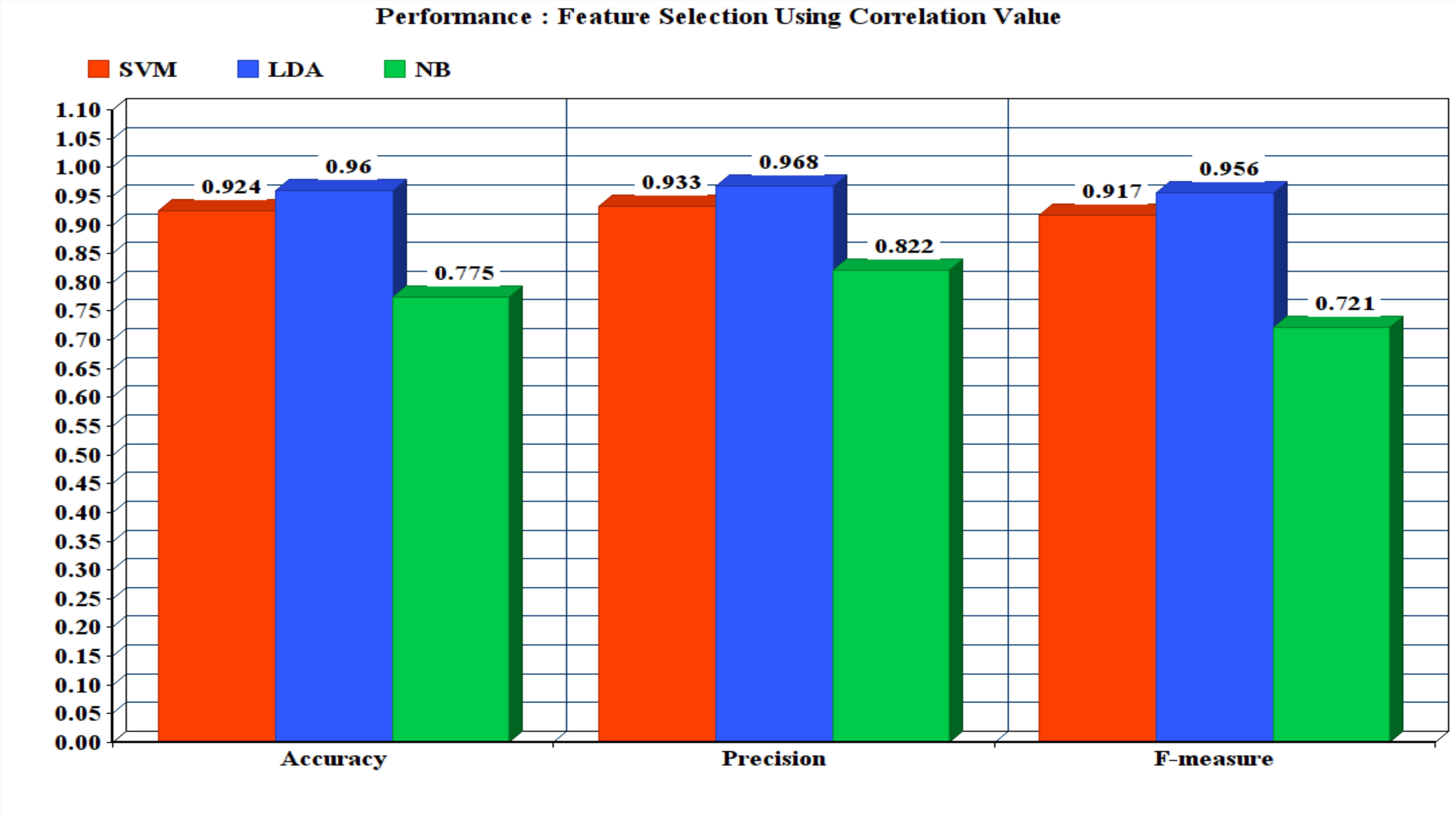
- *Features remaining after removing ones with correlation:*

- > 0.9 = 20
- > 0.8 = 13
- > 0.7 = 10

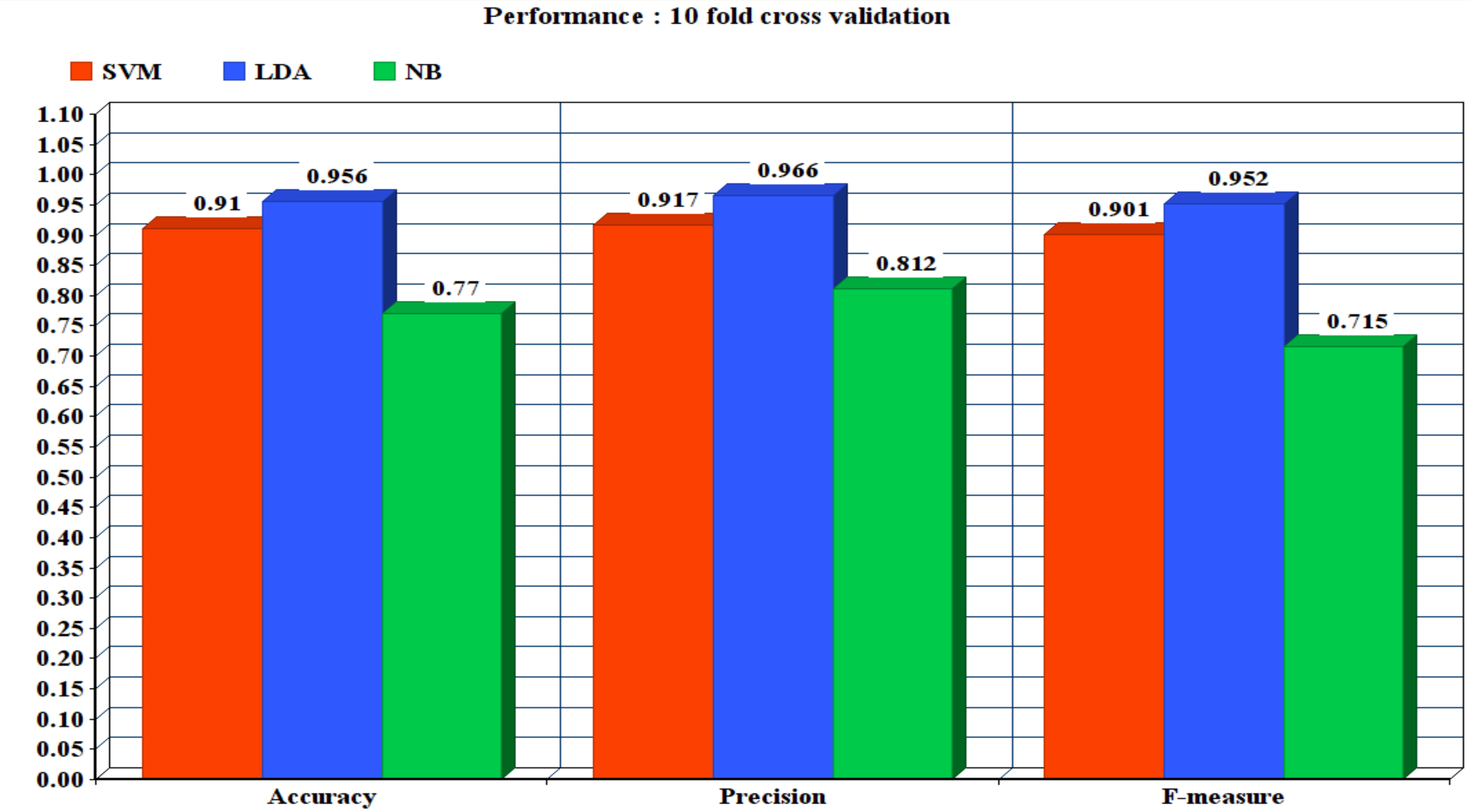
MODEL PERFORMACE : PCA BASED FEATURE SELECTION



MODEL PERFORMACE : CORRELATION BASED FEATURE SELECTION (IGNORE CORR > 0.9)



MODEL PERFORMACE : 10-fold cross validation


Performance : 10 fold cross validation
🟥 SVM
🟦 LDA
🟩 NB
Accuracy
Precision
F-measure