Project 1: Dimension Reduction, Predictive Modeling, and Mislabeling

Group 27:

Daniel González Muela

Francisco Boudagh

Purusothaman Seenivasen

Sky Sunsaksawat

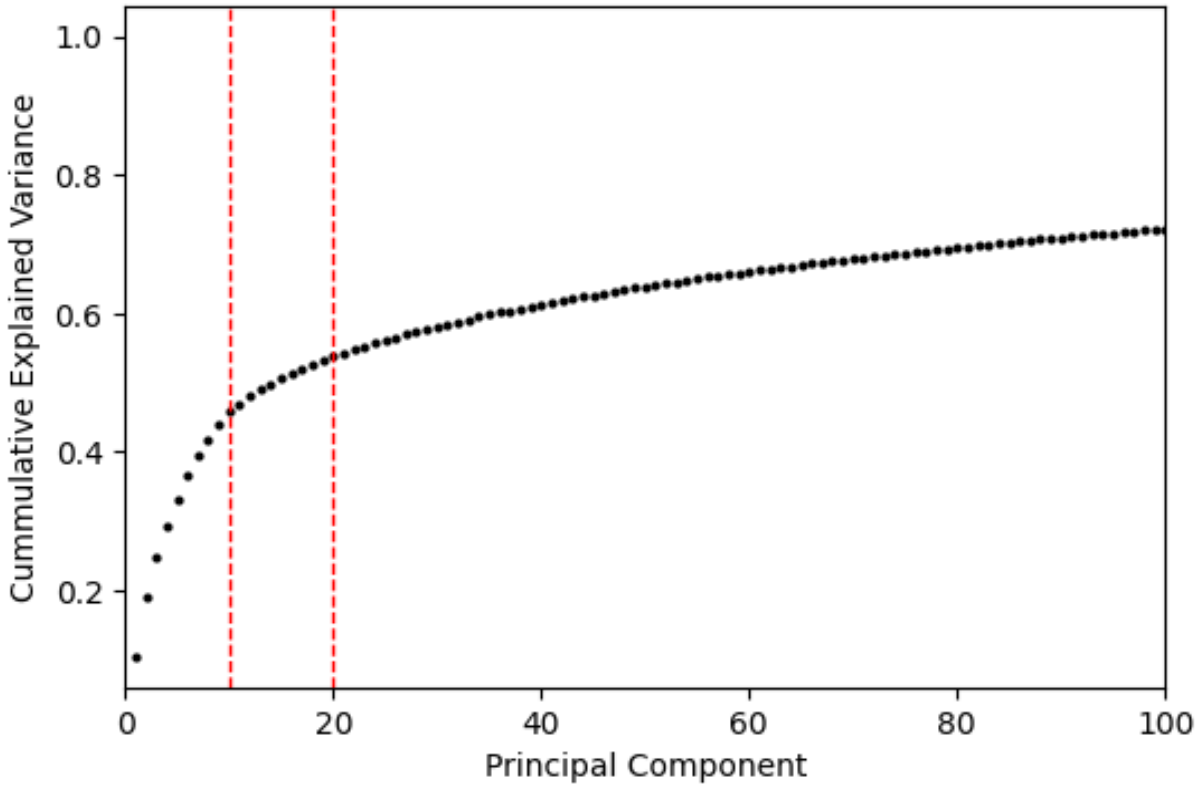
MVE441 Statistical Learning for Big Data

18th April 2024

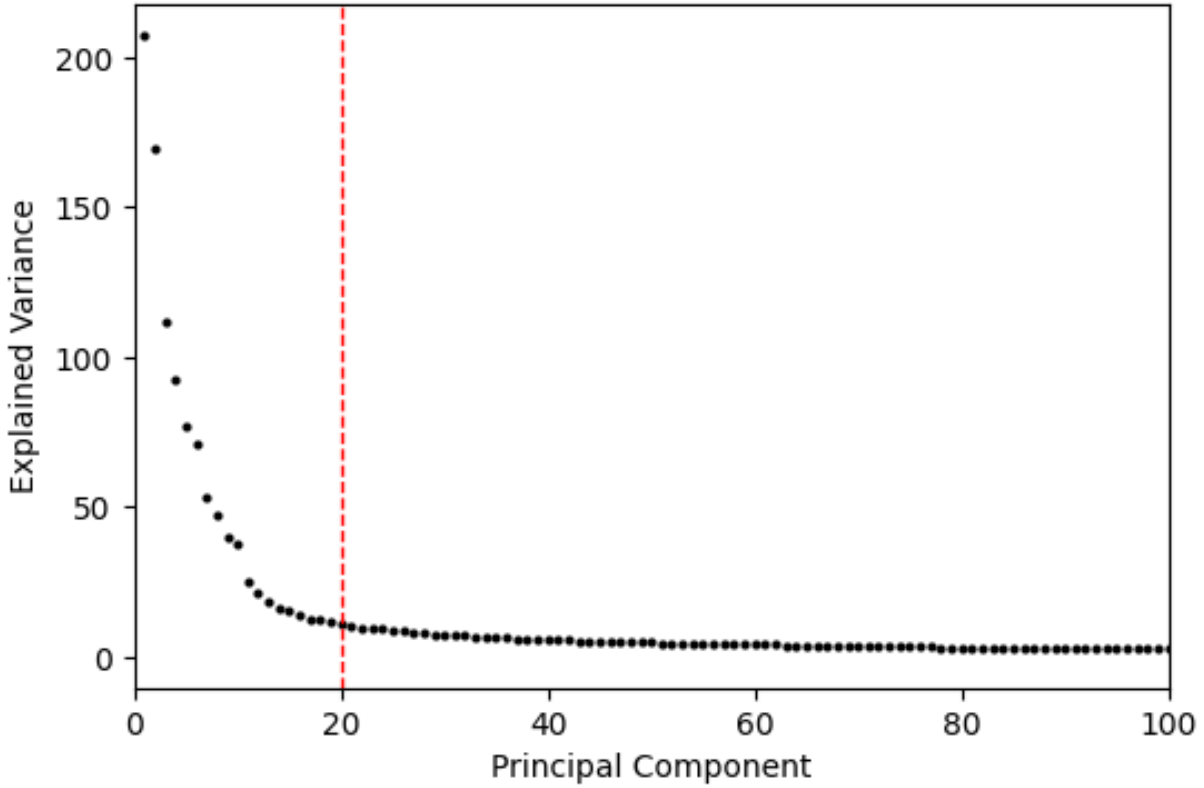
CHALMERS
UNIVERSITY OF TECHNOLOGY

Scale data and explore PCA

Ratio of explained variance



Elbow method



Both methods suggest the number of principle components around >20

3 different classifiers of different character

kNN, small k

- the closer 2 points the more similar they are

QDA

- p(x|i) ~ Normal (should be tested)
- QDA assumes that each class has its own covariance matrix

Logistic Regression

- Binary classification (extended to multiclass using softmax)
- No multicollinearity (reduced with PCA)
- Large sample (sample of 2000)
- Linear relationship of variables to log odds
- No outliers
- Independent observations

Flexible

Moderate

Rigid

Setup

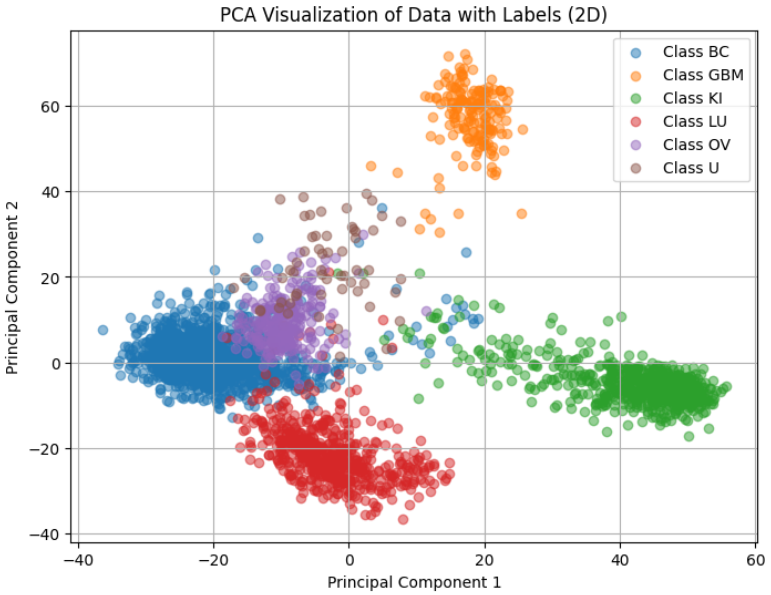
- **Training data sizes:** 50%, 65%, 80%
- **Mislabeling in training data:** 0%, 5%, 30%, 70%
- **Optimization method:** *GridSearchCV* for parameters (k in kNN, # of PCs)
- **Evaluation metric:** Recall or Sensitivity
  - Recall = TP/(TP + FN)
  - Ideal for ensuring detection of positive cases

Limitations

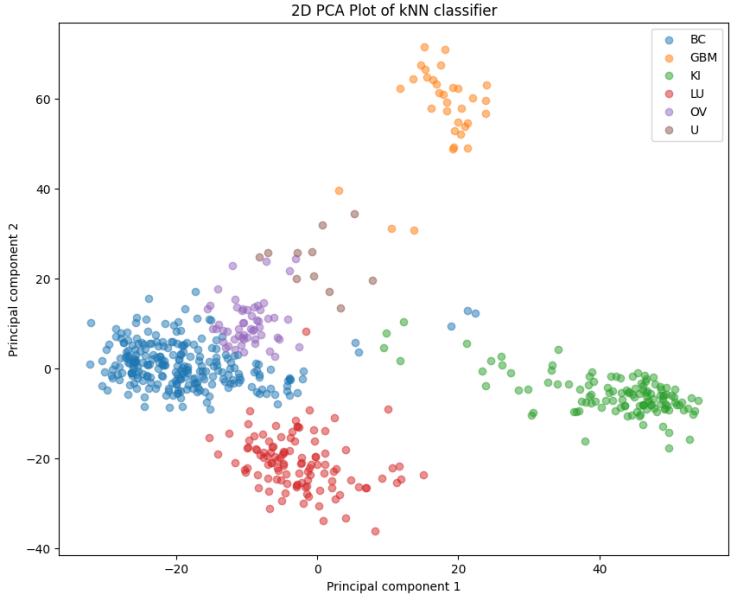
(Due to limited time and compute resources)

- Only 3 mislabeling levels
- Only 3 iterations per test to minimze randomness
- Few parameters given to *GridSearchCV*

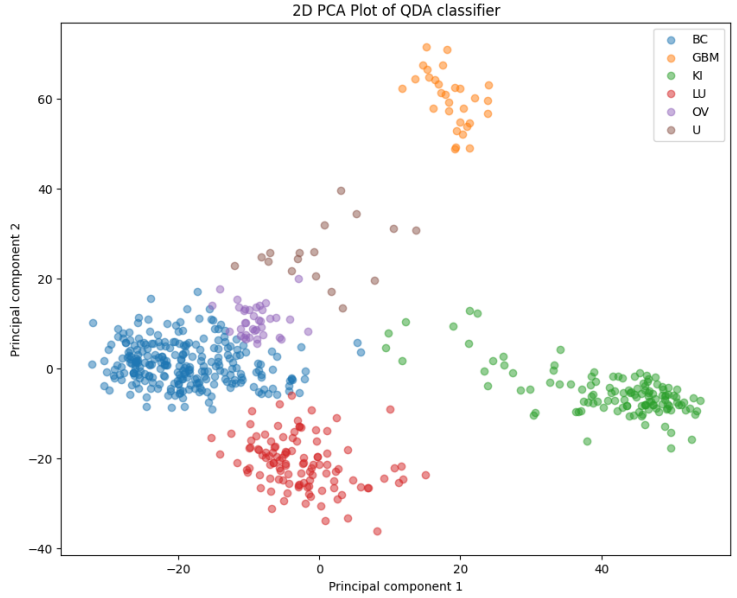
Data Visualization



Original Data

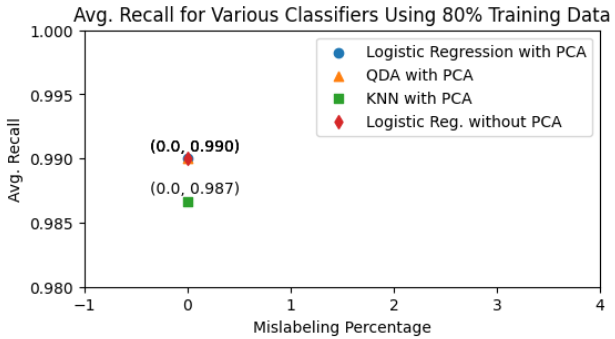


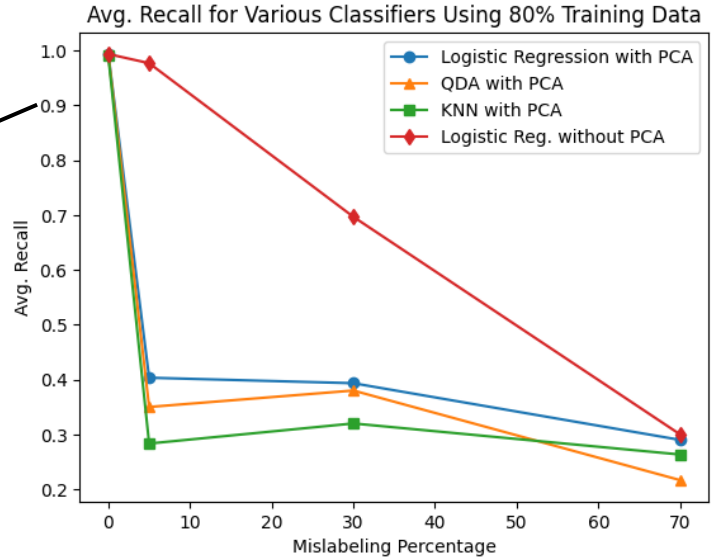
KNN

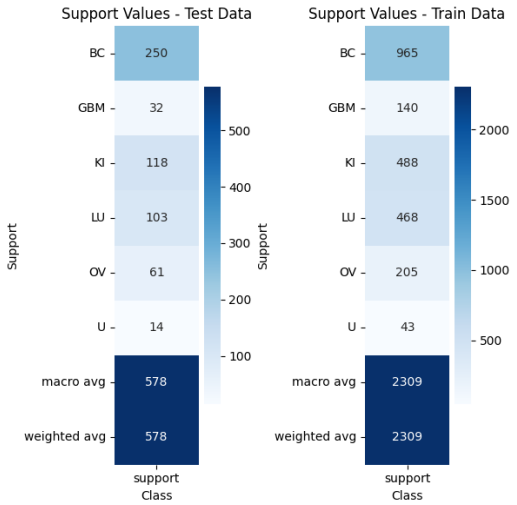


QDA

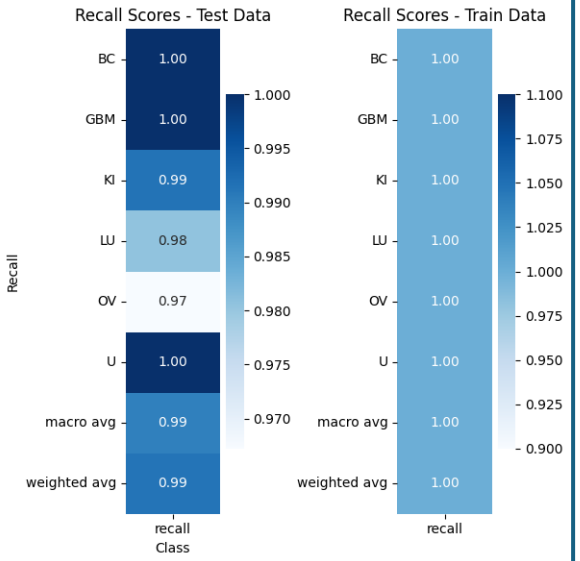
Recall Evaluation on 80% Training Data





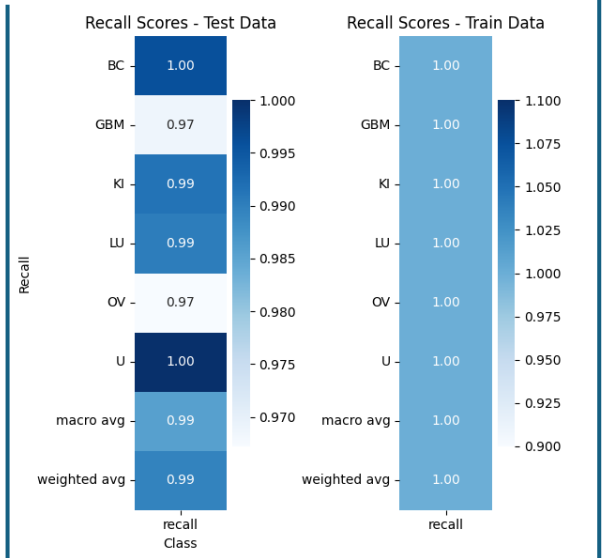


Wtih, Logistic regression performed better than QDA and KNN



Raw Logistic regression

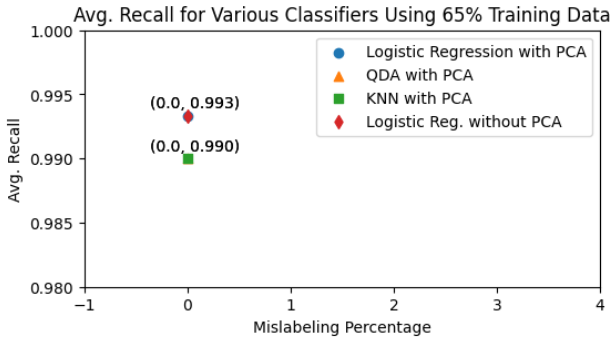
KNN with 40 PCA with 7 neighbours

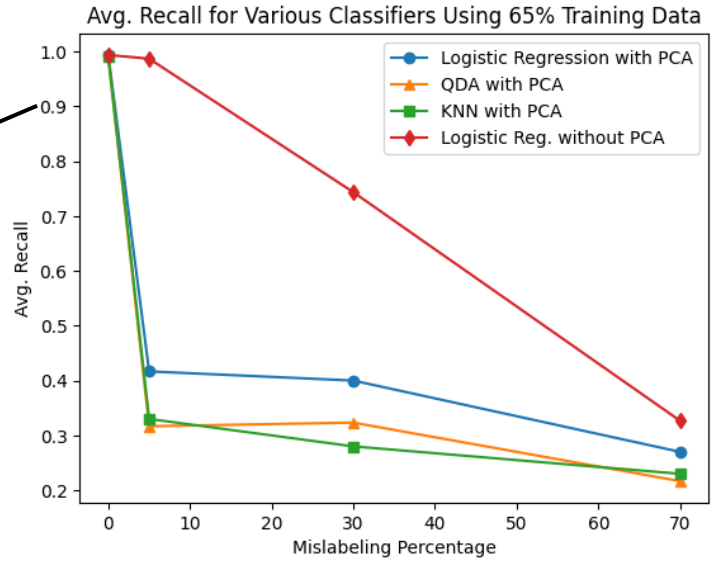


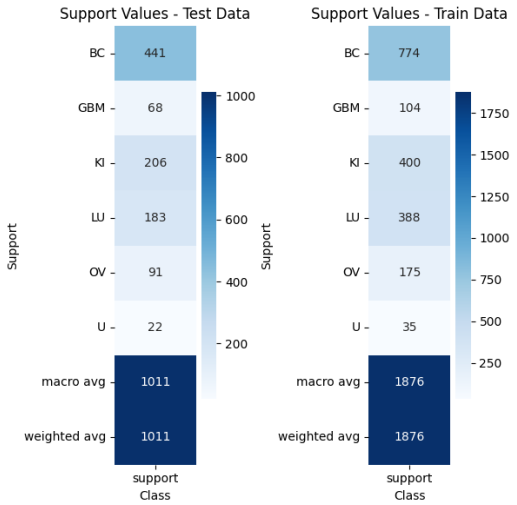
Logistic regression with PCA 40

QDA with 20 PCA

Recall Evaluation on 65% Training Data

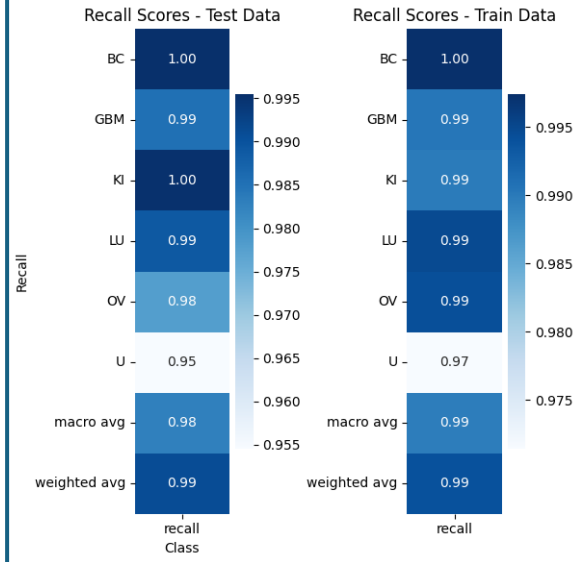




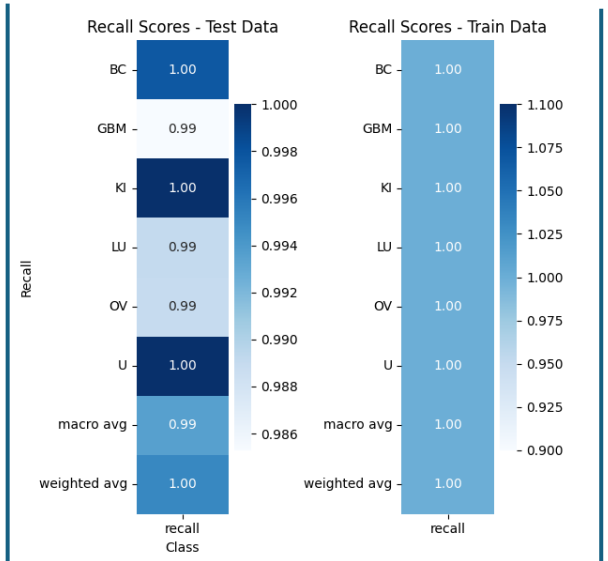


With PCA, Logistic regression performed better than QDA and KNN

Raw Logistic regression



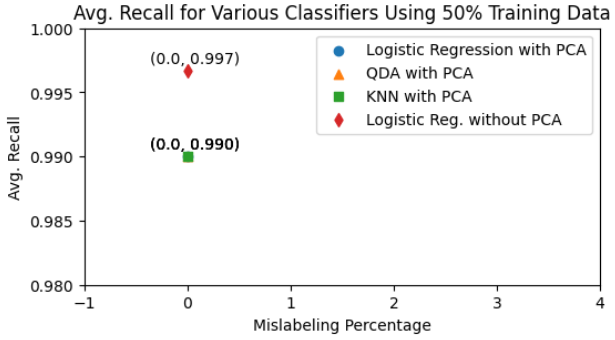
KNN with 30 PCA with 5 neighbours

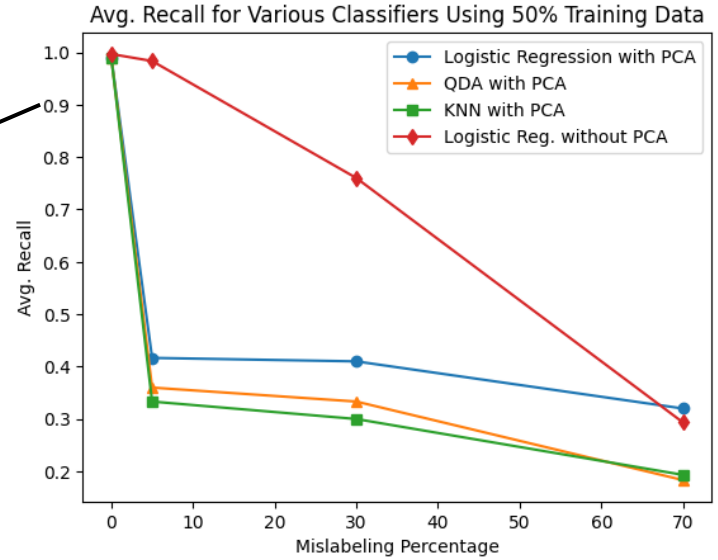


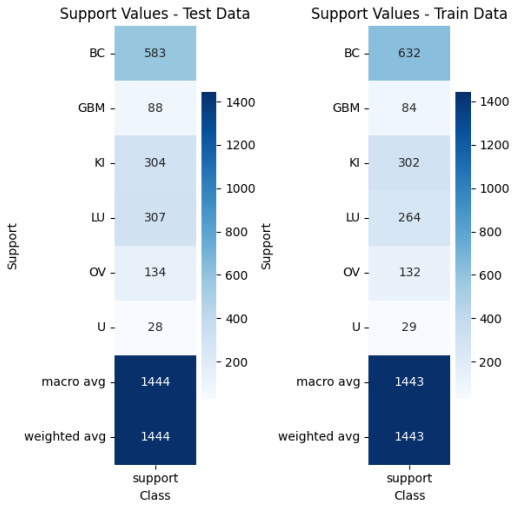
Logistic regression with PCA 40

QDA with 20 PCA

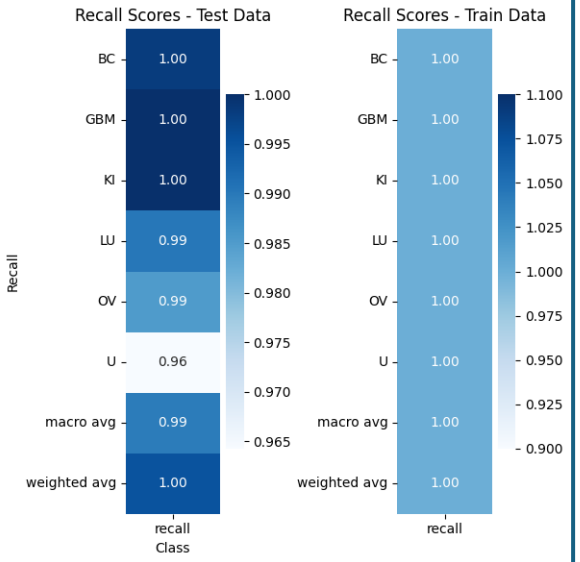
Recall Evaluation on 50% Training Data



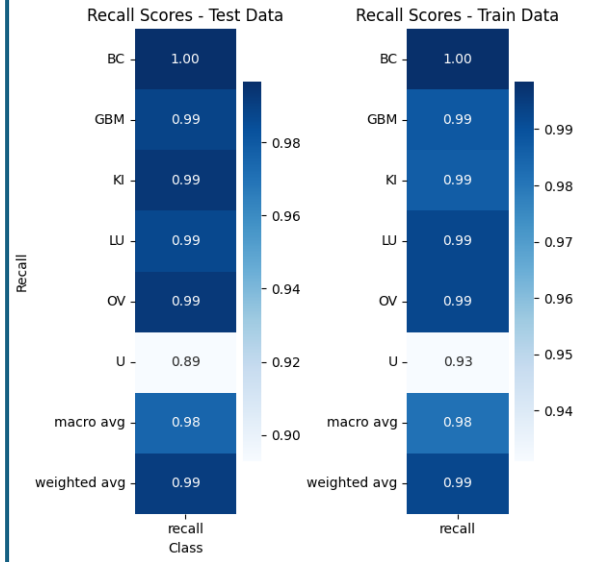




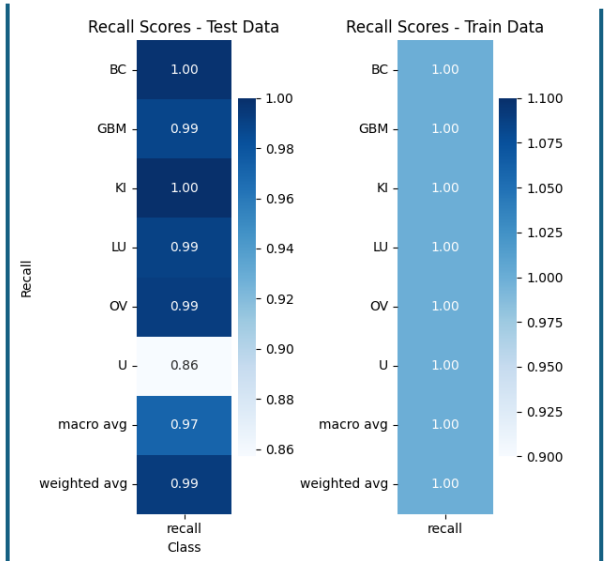
With PCA, Logistic regression performed better than QDA and KNN



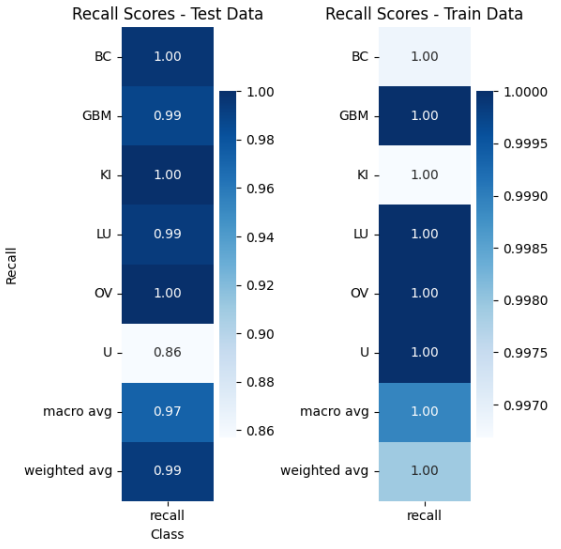
Raw Logistic Regression



KNN with 30 PCA with 5 neighbours



Logistic Regression with PCA 20



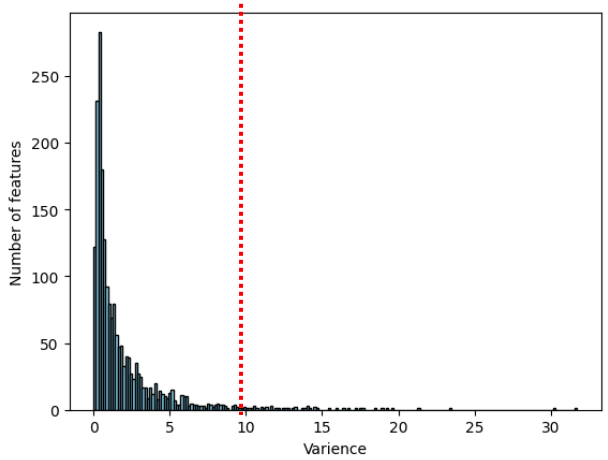
QDA with 20 PCA

Feature Selection

Keeping features with maximum variance

Training data size (% of dataset)	% of Mislabeling data			
	0%	5%	30%	70%
80%	**0**	**0**	**10**	**10**
65%	**0**	**0**	**10**	**10**
50%	**0**	**0**	**10**	**10**

The table shows the maximum variance threshold for feature selection that optimizes predictive performance in each condition

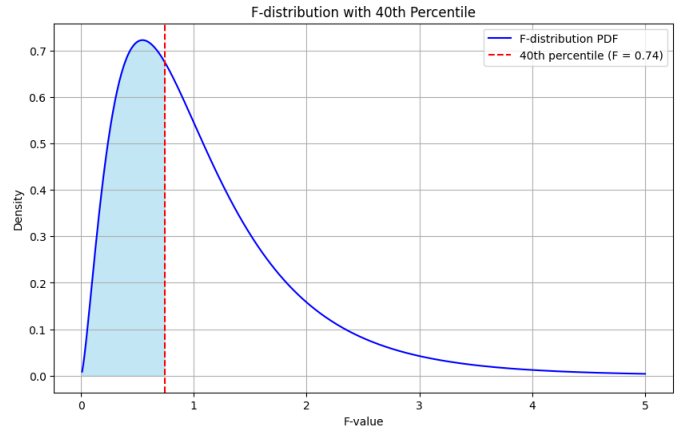


Keeping feature with ANOVA F-test

- the score is obtained by comparing the variances of each feature to the target variable.

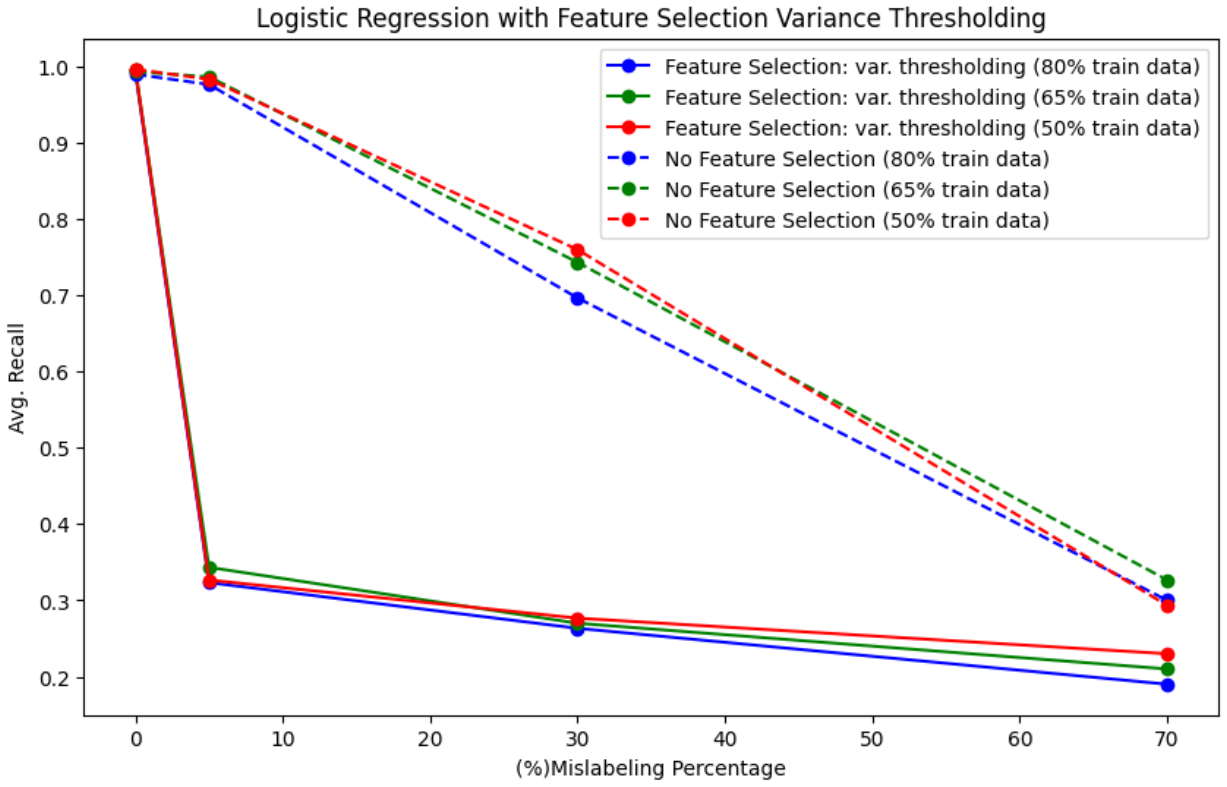
Training data size (% of dataset)	% of Mislabeling data			
	0%	5%	30%	70%
80%	**40th**	**90th**	**5th**	**5th**
65%	**40th**	**80th**	**5th**	**5th**
50%	**60th**	**60th**	**5th**	**5th**

The table show the percentile for features selection that optimize predictive performance in each condition



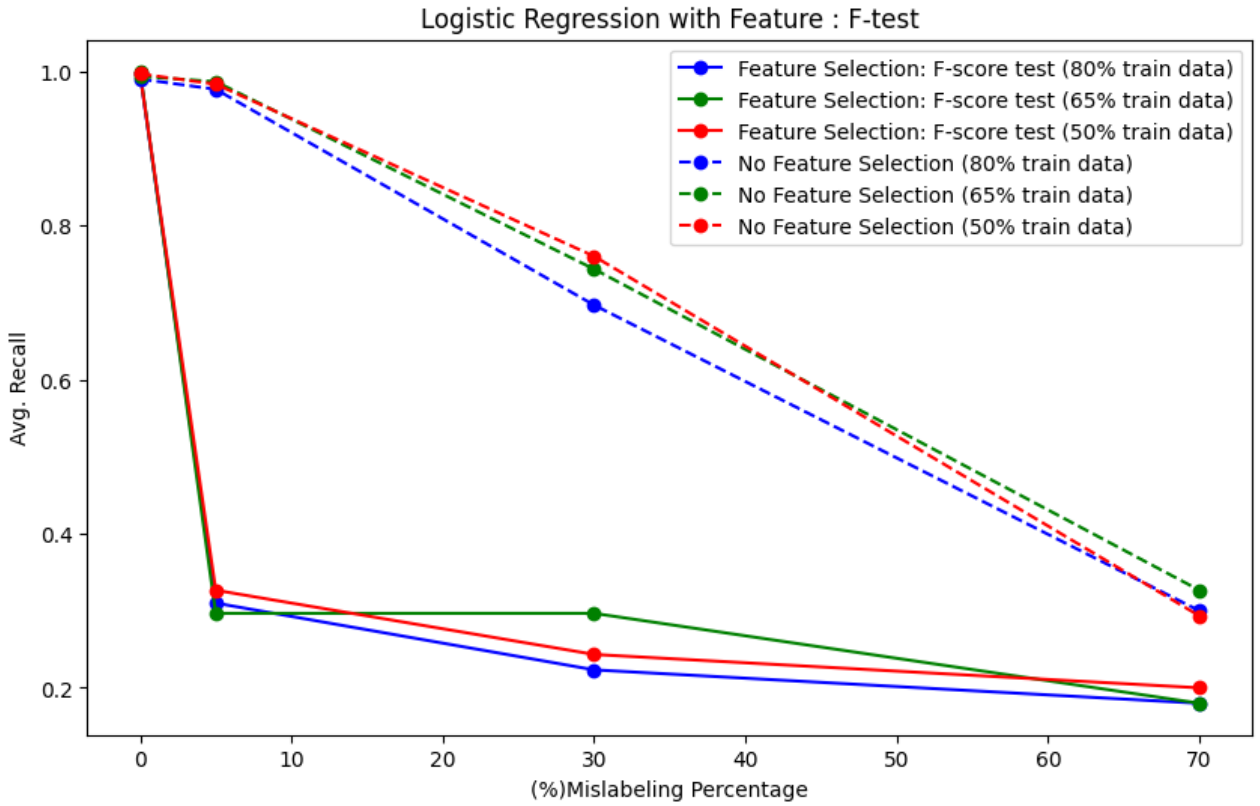
Recall Evaluation on Feature Selection with Maximum Variance and F test

Keeping features with maximum variance



Keeping feature with ANOVA F-test

- the score is obtained by comparing the variances of each feature to the target variable.



Conclusion and Key Findings

Top performer: Logistic Regression

- Indicates (a likely) linear relationship between the variables.

Stability against mislabeling (noise):

- Logistic regression without feature selection shows the highest robustness to mislabeling.
- Regularization techniques help to prevent overfitting.

Impact of mislabeling:

- Even a small percentage of mislabeling significantly impacts performance.
- This effect is amplified when feature selection is applied.