AML Project 2 - Feature Selection

Łukasz Zalewski 329532
Władek Olejnik 290593

Implemented Methods

Sequential forward floating selection - selects the most impactful features one after the other (the ones which increase classifier performance the most), and later removes a couple of least impactful features.

Mutual information scoring - scores features using the mutual information between features and the target variable.

Lasso - minimizes the sum of the squared errors between the predicted and actual values, while also adding a penalty term that depends on the absolute values of the coefficients.

Random Forest - selects features by calculating the importance of each feature based on its ability to increase the purity of the leaves in the decision trees that make up the forest.

ReliefF - an extension of the Relief algorithm, selects features by calculating a feature score for each feature based on the identification of feature value differences between nearest neighbor instance pairs.

Methodology

We evaluated each approach using the following methodology:

1. Extract relevant features using some approach
2. For selected features find best hyperparameters for KNN, RandomForest, and XGBoost, using sklearn grid search
3. Split the dataset into 10-folds and evaluate using selected features, all three classifiers and their best hyperparameters from the previous step. We use the scoring equation from the project description, which combines balanced accuracy with a penalty for a number of features. The final score is averaged over 10 folds.

Results

Dataset name	Feature selection method	Classifier	Best number of features	Score
artificial	SFFS	KNN	8	0.888
artificial	RF	KNN	12	0.8715
artificial	ReliefF	KNN	13	0.868
artificial	Lasso	RF	7	0.865
artificial	MI	XGB	19	0.734
spam	Lasso	RF	75	0.937
spam	RF	RF	60	0.925
spam	MI	RF	50	0.910
spam	ReliefF	RF	85	0.903

Final Submission

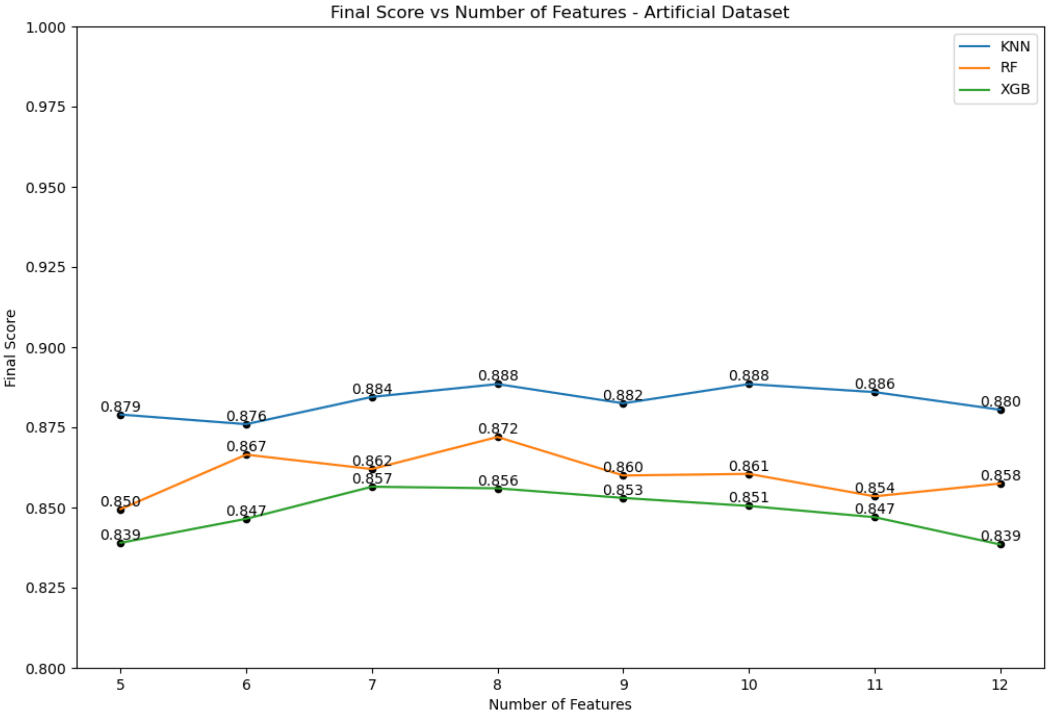
We generated predictions for artificial dataset testset using:

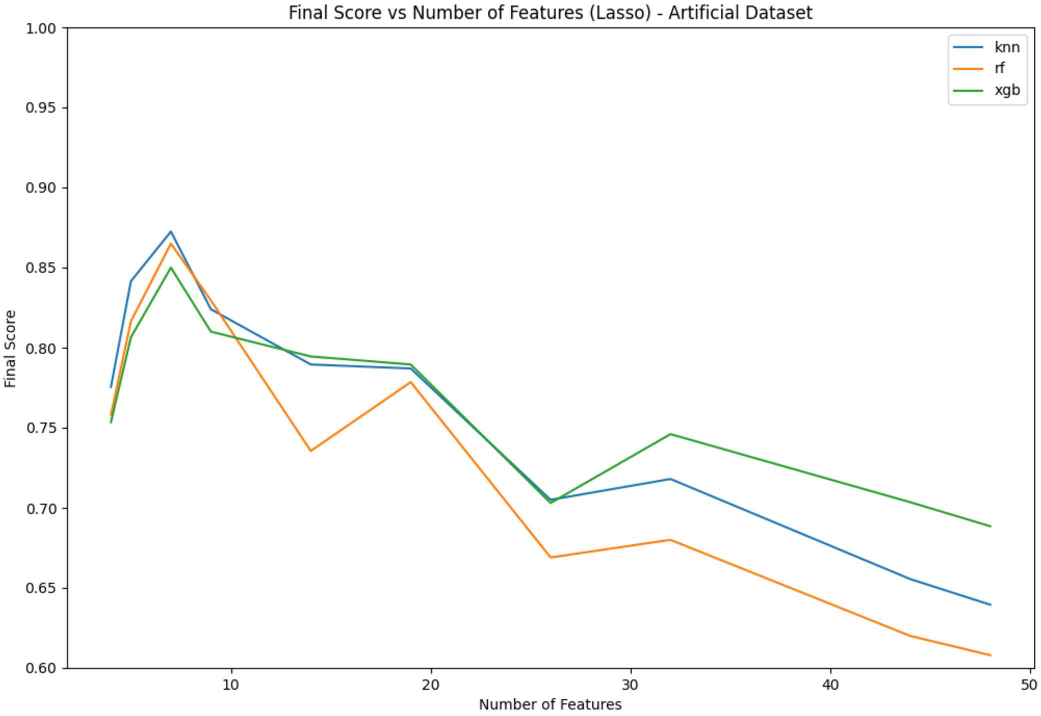
- 5 neighbors KNN
- 8 features chosen with sequential forward floating selection

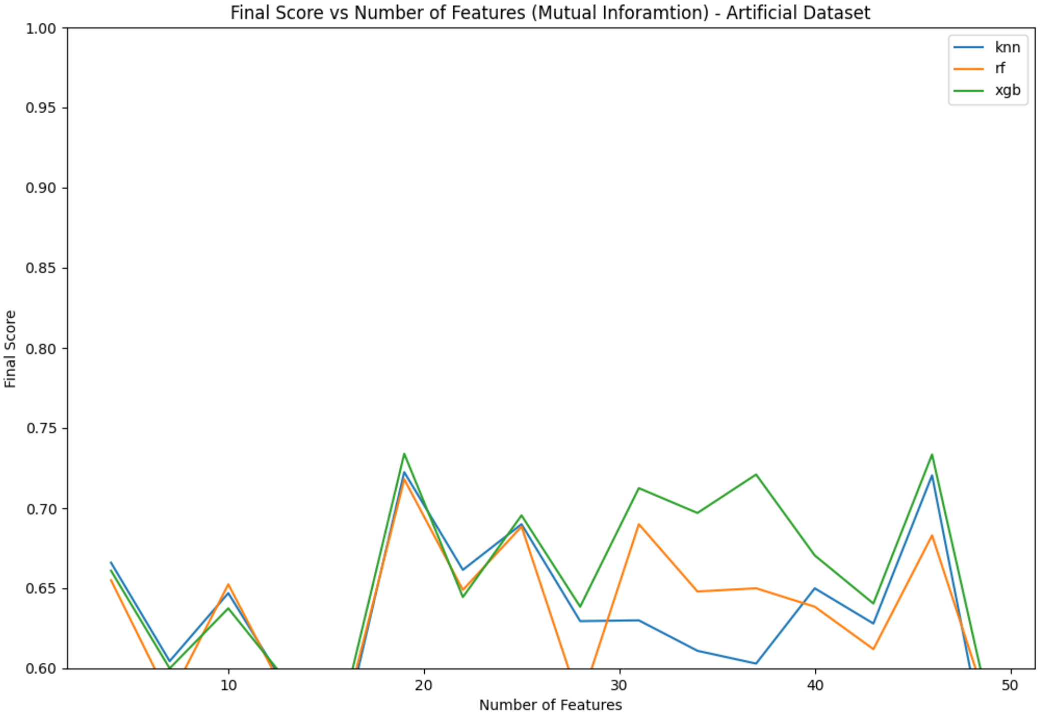
We generated predictions for spam dataset testset using:

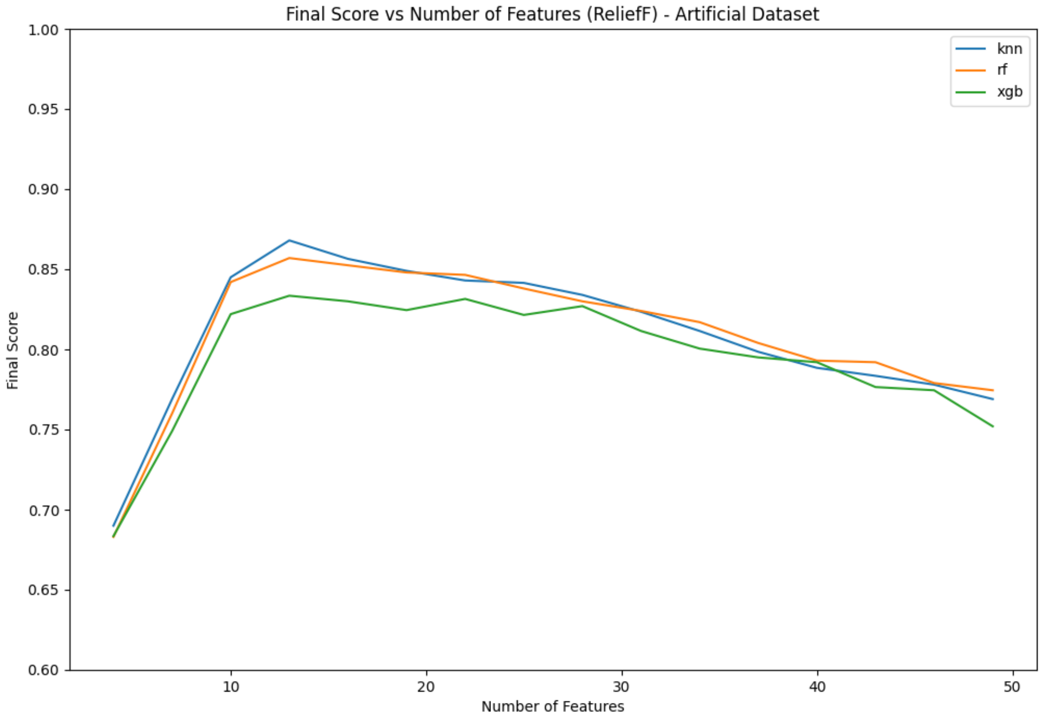
- Random Forest with 200 estimators
- 75 features selected with Lasso

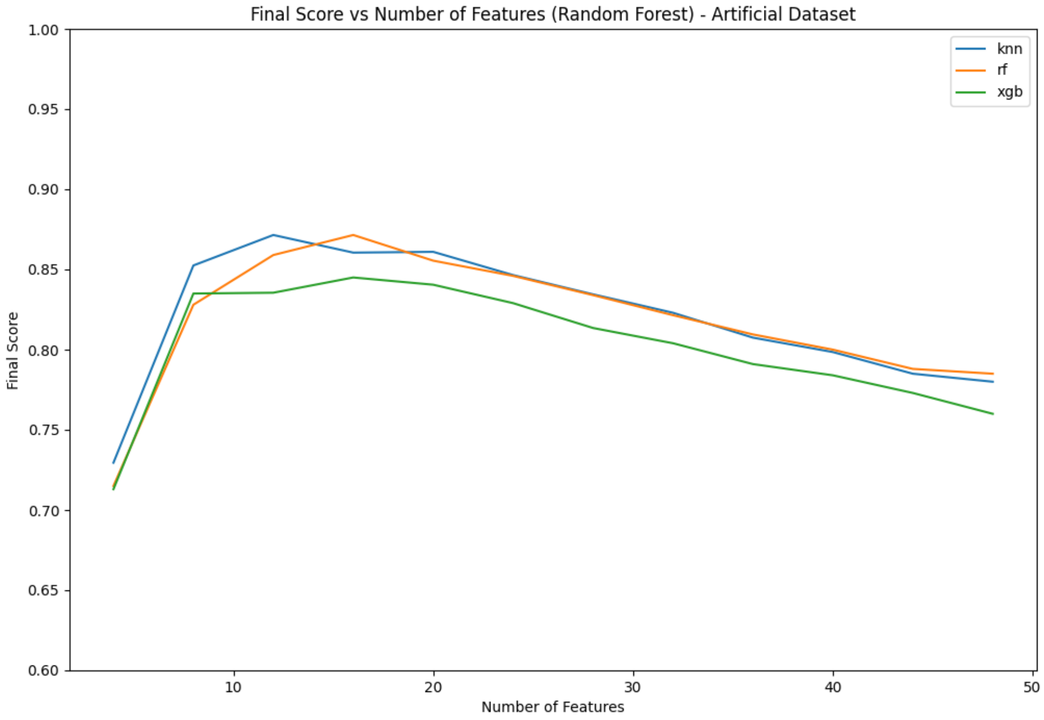
Appendix: Results Charts

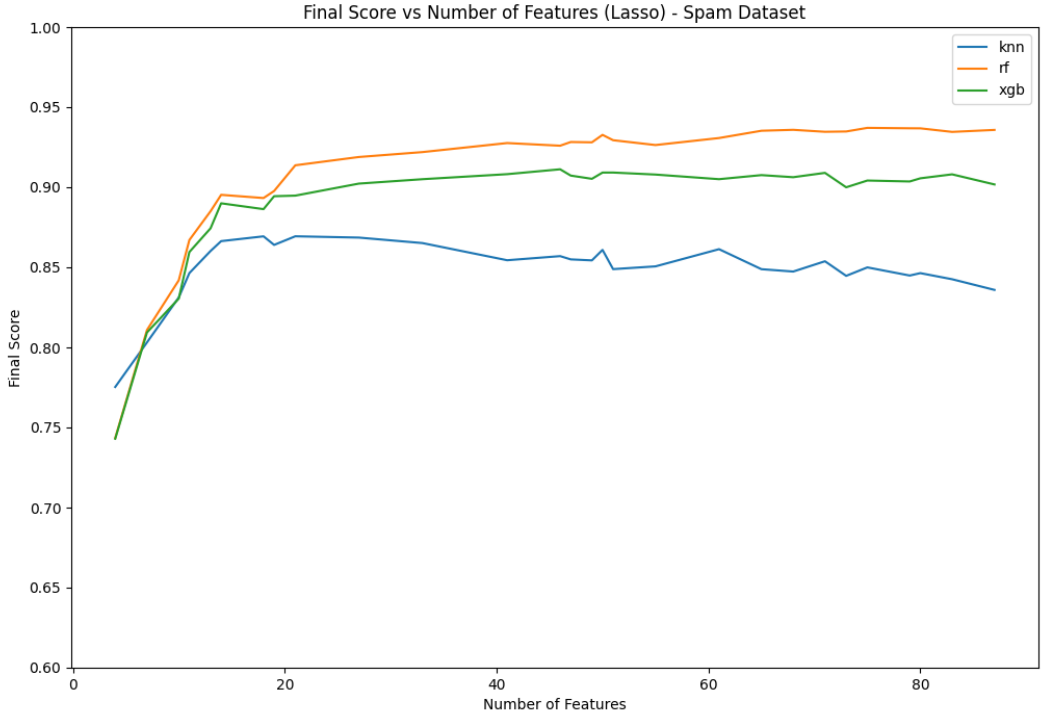


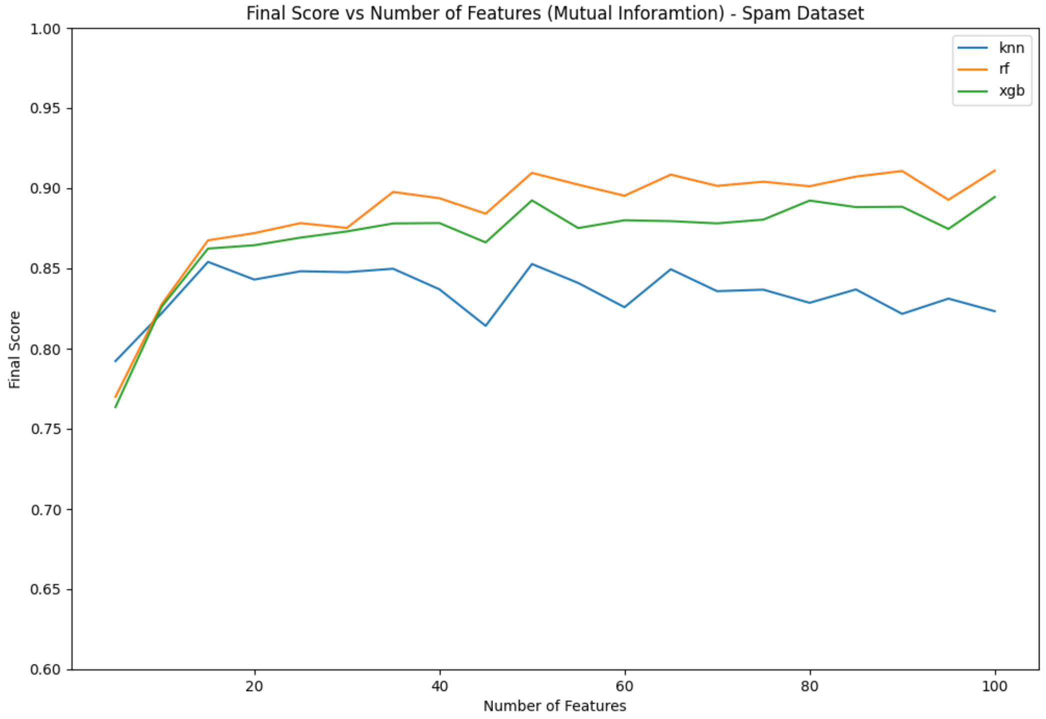

Final Score vs Number of Features (Lasso) - Artificial Dataset
Final Score
Number of Features
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
10
20
30
40
50
knn
rf
xgb

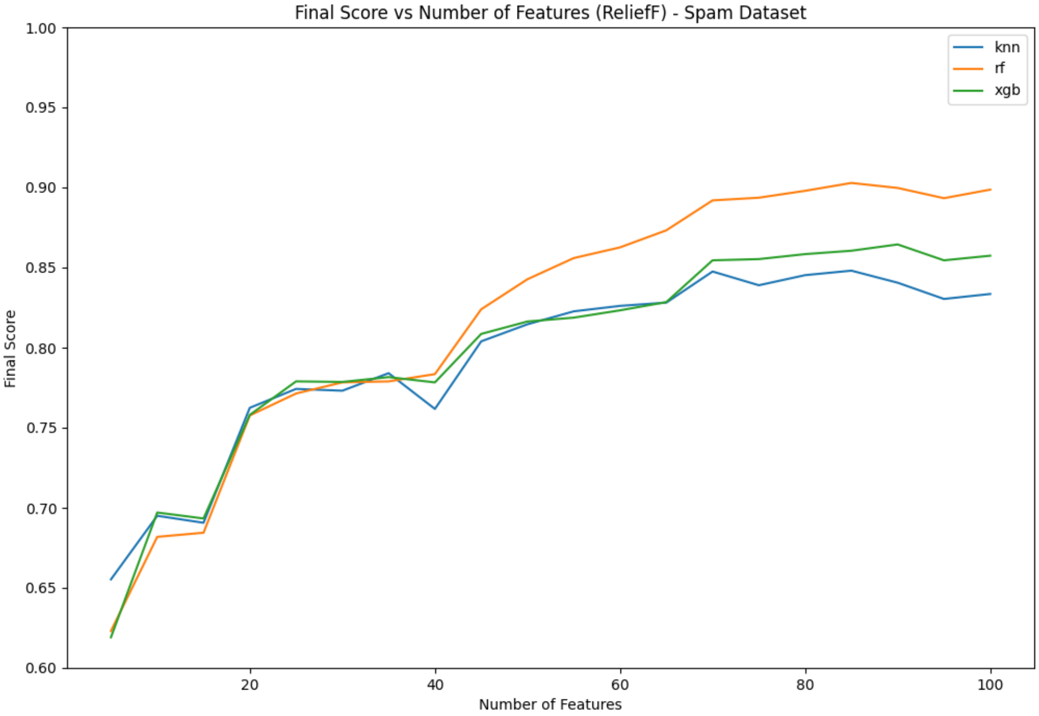

Final Score vs Number of Features (Mutual Inforamtion) - Artificial Dataset
Final Score
Number of Features
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
10
20
30
40
50
knn
rf
xgb

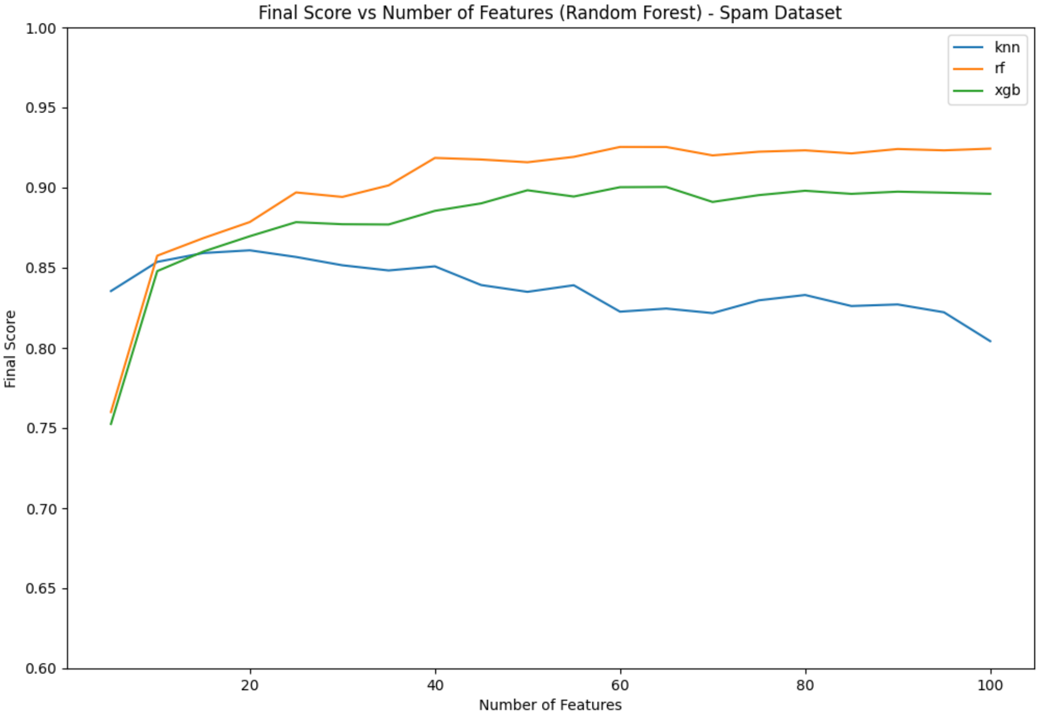

Final Score vs Number of Features (ReliefF) - Artificial Dataset
knn
rf
xgb
Final Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
10
20
30
40
50
Number of Features


Final Score vs Number of Features (Random Forest) - Artificial Dataset
knn
rf
xgb
Final Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
10
20
30
40
50
Number of Features


Final Score vs Number of Features (Lasso) - Spam Dataset
Final Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
20
40
60
80
Number of Features
knn
rf
xgb


Final Score vs Number of Features (Mutual Inforamtion) - Spam Dataset
Final Score
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
20
40
60
80
100
Number of Features
knn
rf
xgb


Final Score vs Number of Features (ReliefF) - Spam Dataset
Final Score
Number of Features
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
20
40
60
80
100
knn
rf
xgb


Final Score vs Number of Features (Random Forest) - Spam Dataset
Final Score
Number of Features
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
20
40
60
80
100
knn
rf
xgb