Data Challenges on Prediction of Treatment of Breast Cancer

Lecturers: - Jean-Philippe vert
- Yulong JIAO

Group Members: - Jeremiah FADUGBA
- Belona SONNA

PLAN

- Data
- Techniques/ Model used
- Techniques that didn’t work
- Final Model
- Summary

THE DATA

```
First Class precentage: %64.130435
Second Class precentage: %35.869565
```

THE ACCURACY IS NOT SUFFICIENT

THE F1 SCORE

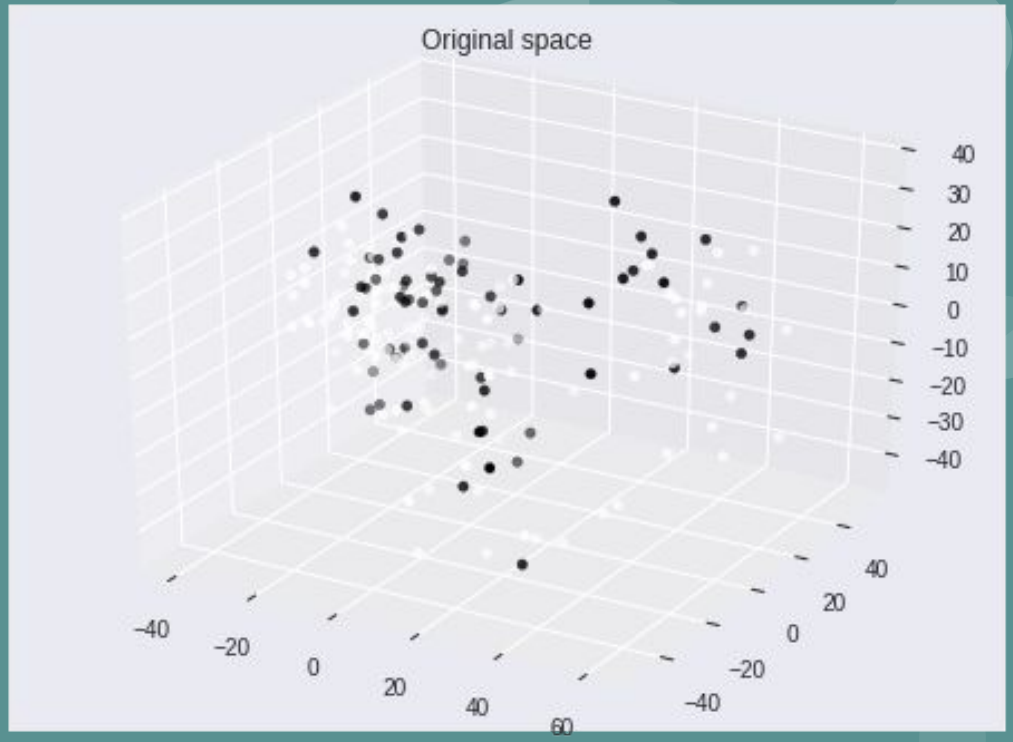
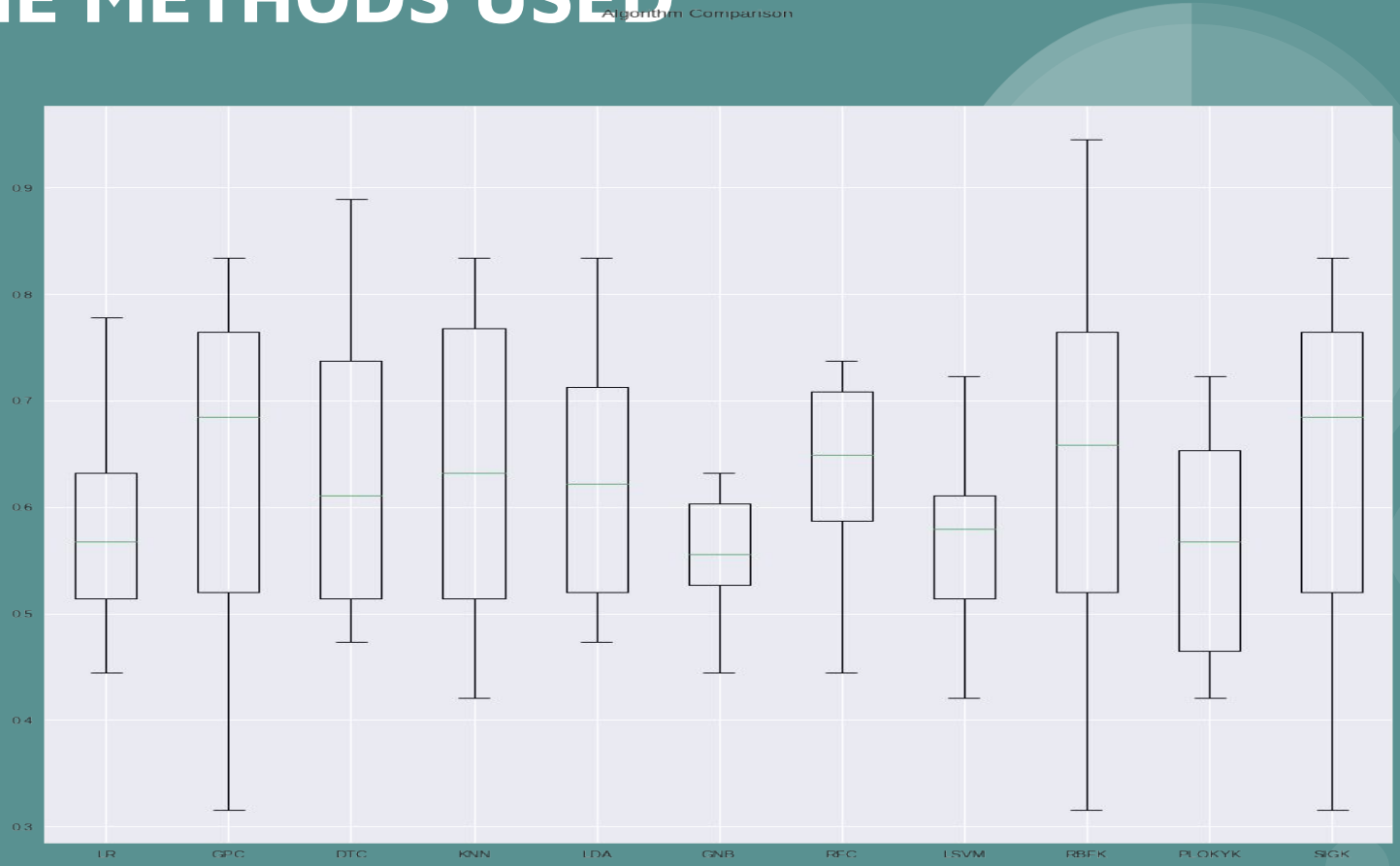


Fig: KPCA with 3 components

SOME METHODS USED



Algorithms used.

Top 5 Models that still failed

Model	F1 Score	Accuracy
Random Forest	0.56	60%
Kernel Ridge Regression	0.55	62%
Support Vector Machine (Linear)	0.59	61%
K Nearest Neighbour	0.57	62%
Logistic Regression (with kernel)	0.61	63%

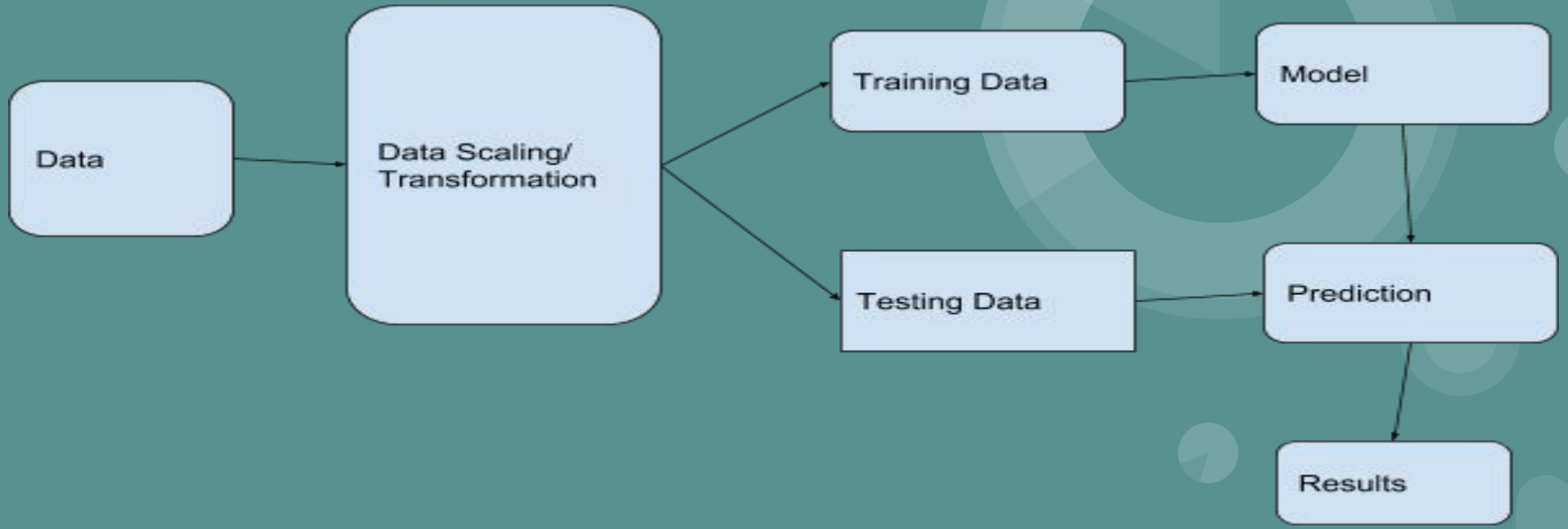
Techniques we tested.

Label Propagation

Gaussian Random Projection

RBFSampler

Final Pipeline Model

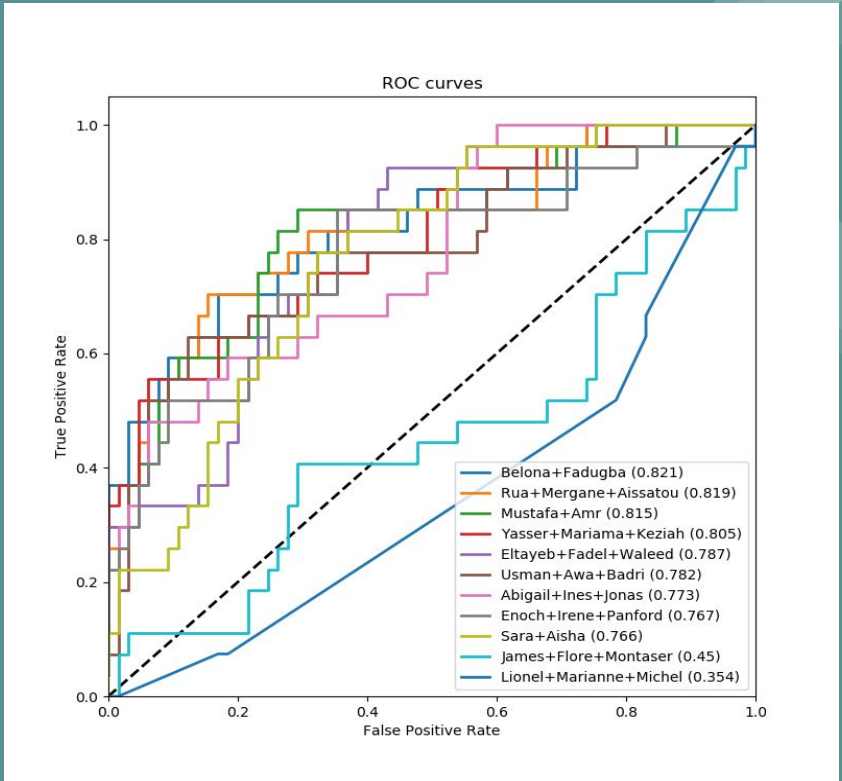


Model:
Support Vector Machine (with RBF)

We need to :
>> Have high F1 score for the class with -1
>> Select the model with high F1 score for all models.

	Precicion	Recall	F1 Score
-1	0.67	0.97	0.79
1	0.43	0.27	0.31

Results of the challenge on test data



Summary

- Gaussian Random Projection works by reducing the dimension and retaining pairwise distance
- Kernel approximation would have been better suited if there were more sample points
- Label Propagation sounds interesting but unfortunately didn't work with our data. wasn't suitable
- Support Vector Machine works on this data with a radial basis function kernel.
- Cross Validation didn't help here.

Always choose a simple model over complicated ones