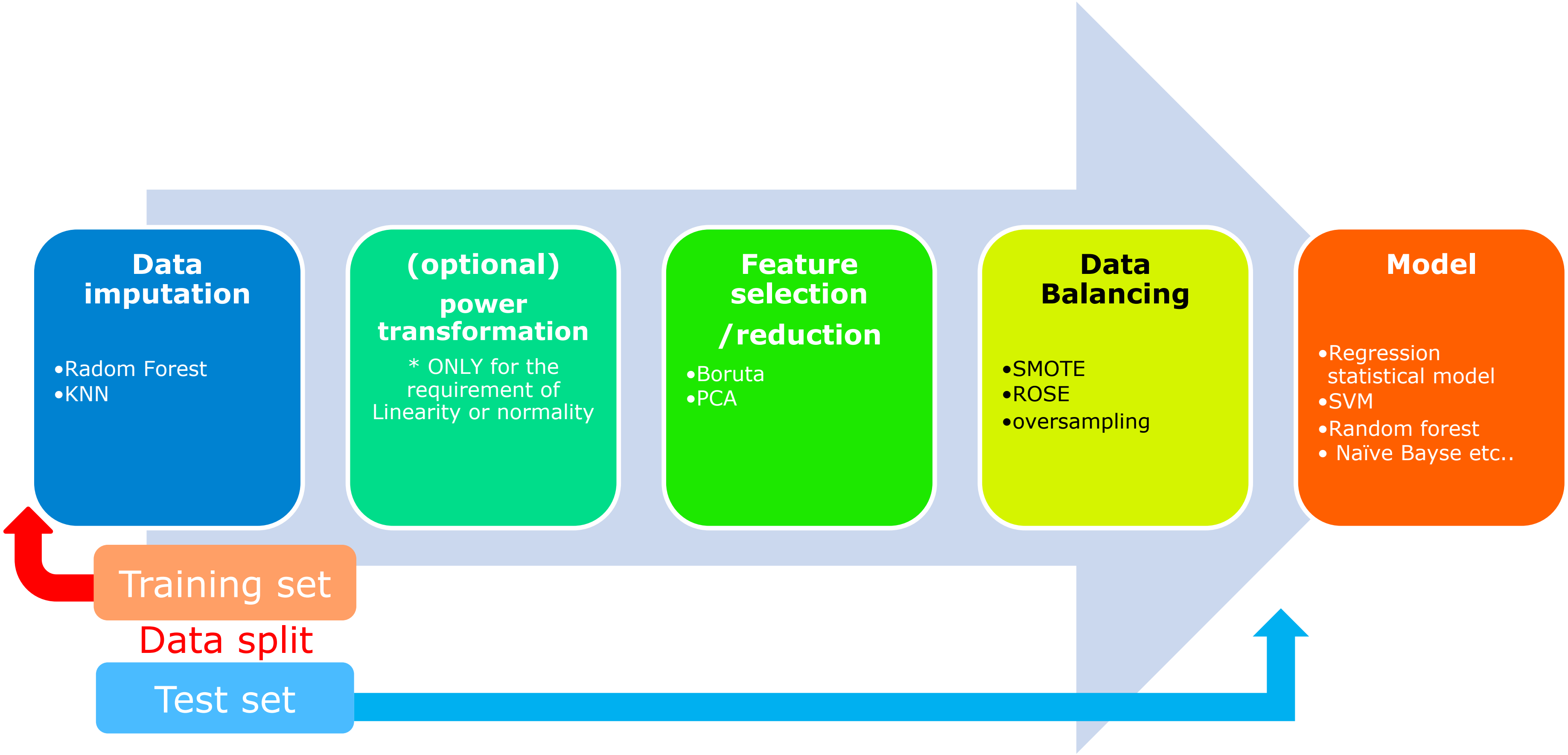
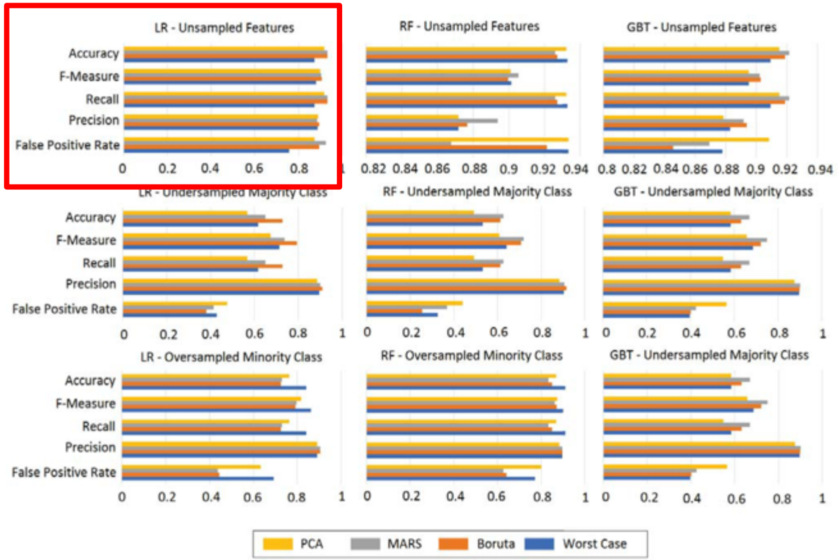
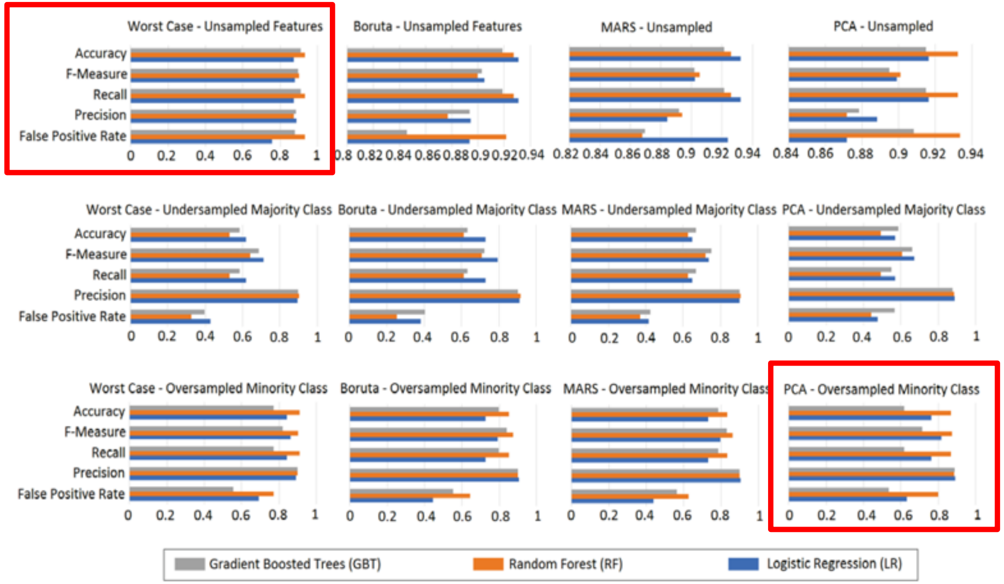
Overview of process to build a model



Literature review





High performance with..

- Boruta : Unsampling
- PCA : Unsampling, Oversampling

Feature selection vs. Data sampling which one should be first?

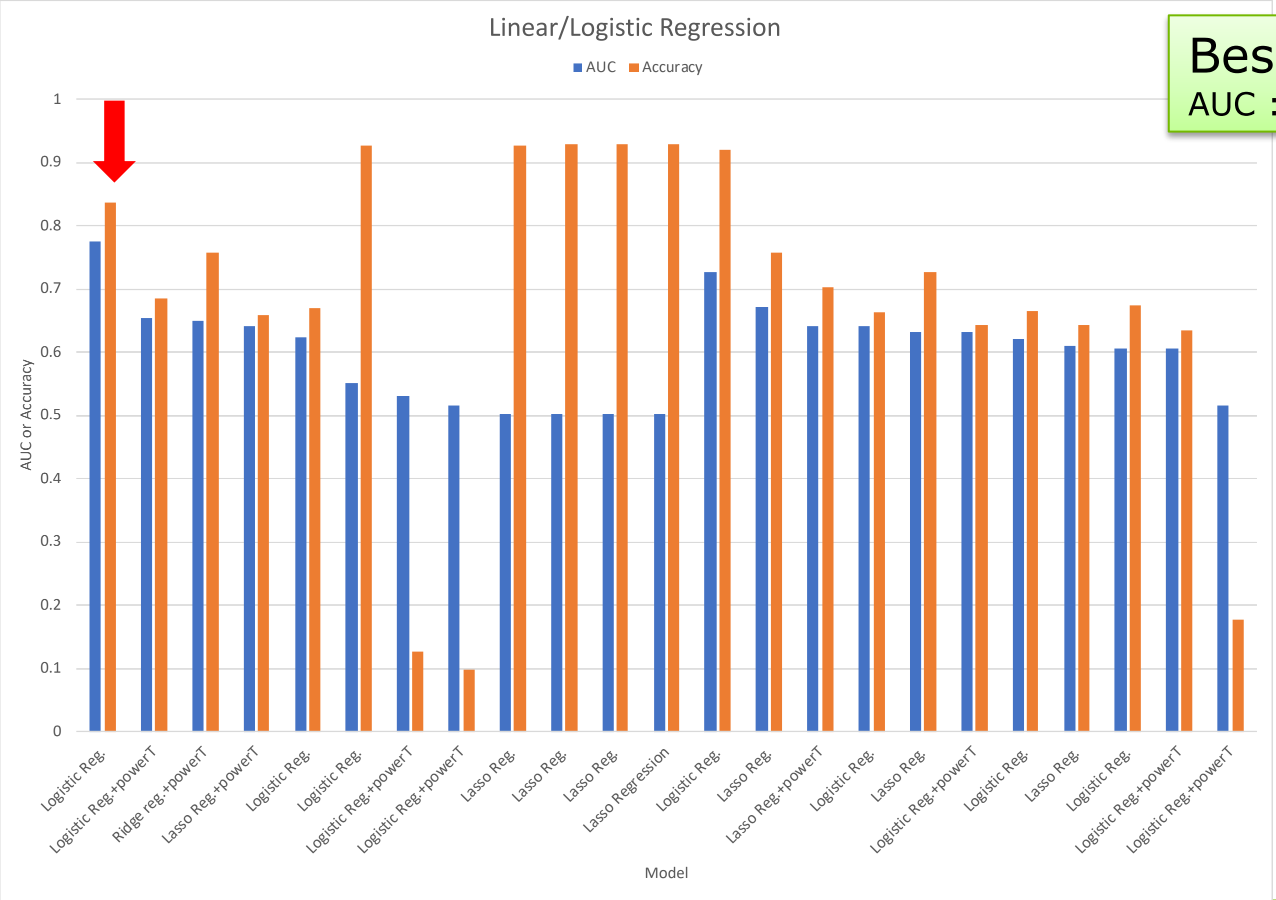
	low-dimensional data		high-dimensional data, without variable selection			high-dimensional data, with variable selection		
Classifier	NC	SMOTE	NC	CO	SMOTE	NC	CO	SMOTE
1-NN		↑		≈	↓		≈	↑
5-NN		↑		↑	↓		↑	↑
DLDA		≈		≈	≈		≈	≈
DQDA		≈		≈	↓		≈	↓
RF		↑		↑	↑		↑	≈
SVM		↑		↑	≈		↑	≈
PAM		↑		↑	≈		↑	↑
PLR-L1		↑		↑	≈		↑	≈
PLR-L2		↑		↑	≈		↑	≈
CART		↑		≈	≈		≈	≈

Figure 5 Summary of results obtained on the simulated data. Green and red color shading denote good and poor performance of the classifiers, respectively. Upwards and downwards trending arrows and the symbol ≈ denote improved, deteriorated or similar performance of the classifier when comparing SMOTE or adjusted classification threshold (CO) with the uncorrected analysis (NC).

Feature selection should be done before Sampling(e.g. SMOTE)

- Classifier could be biased by minority group
- High dimensional data make highly correlated balancing result

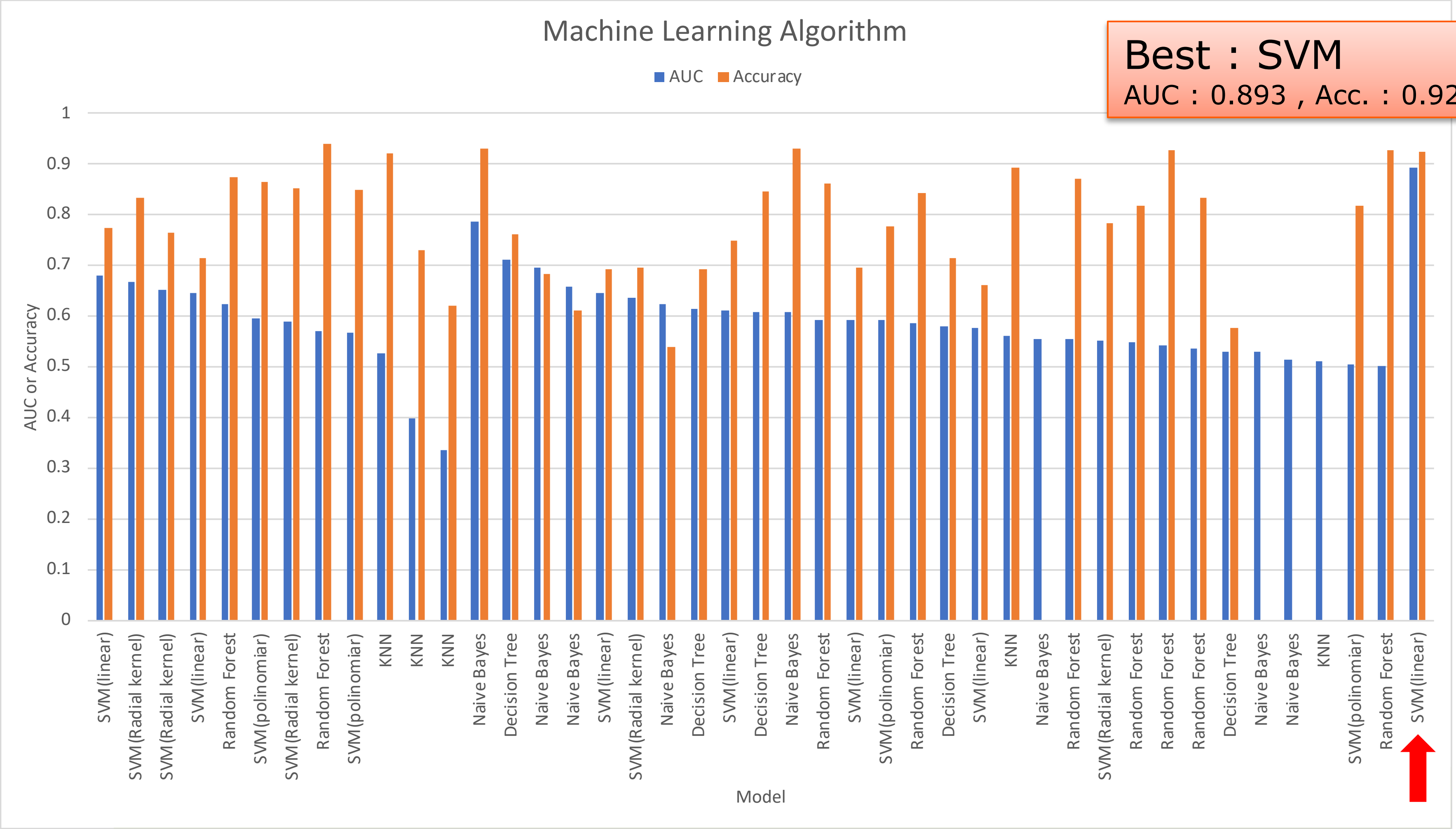
Results : Linear / Logistic Regression after F. selection



Best : Logistic Reg.

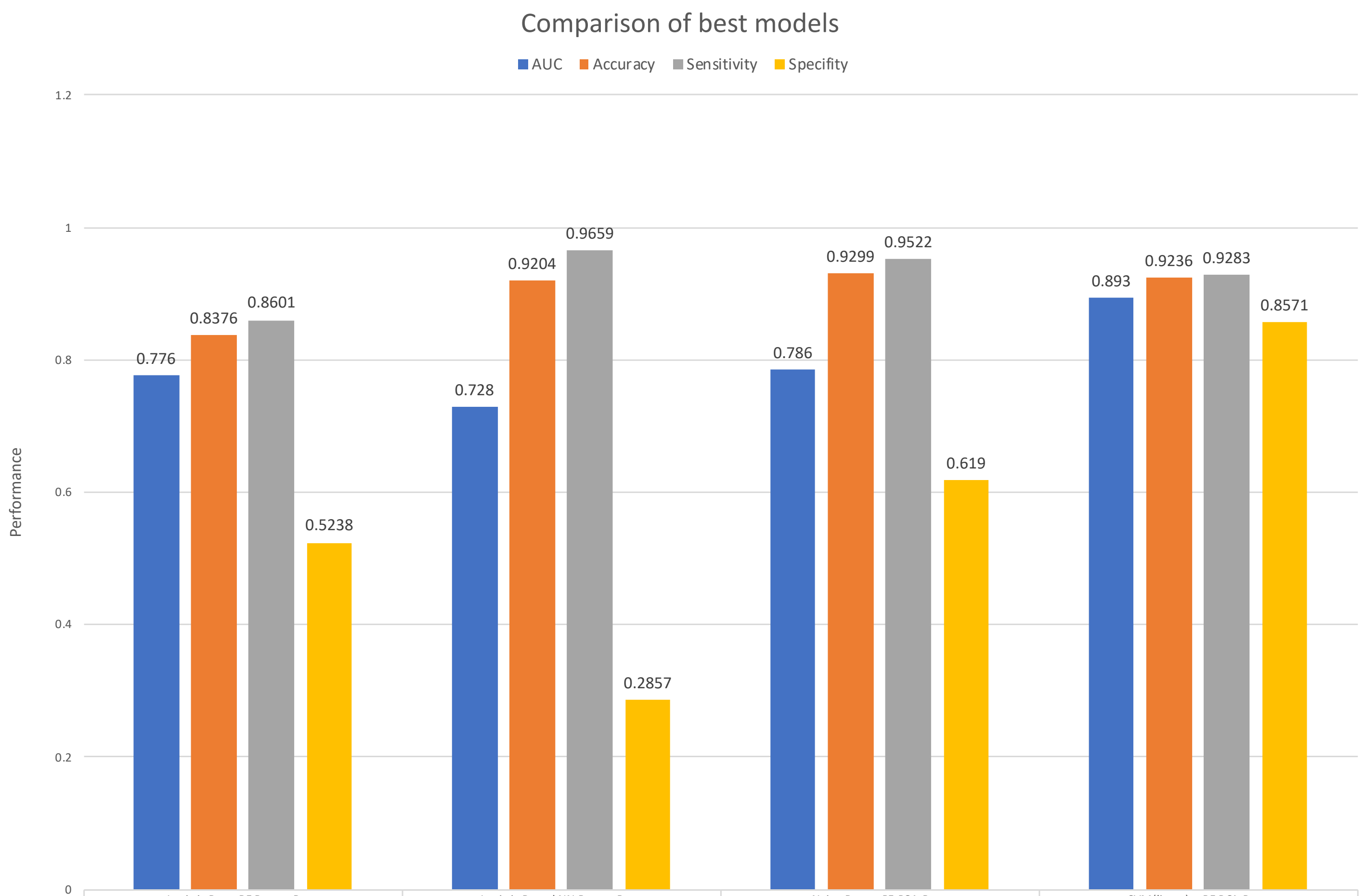
AUC : 0.728 , Acc. : 0.9204

Results : Machine Learning Algorithm after F.selection



Best : SVM

AUC : 0.893 , Acc. : 0.9236



	Logistic Reg.+RF,Boruta,Rose	Logistic Reg.+kNN,Boruta,Rose	Naive Bayes+RF,PCA,Rose	SVM(linear) + RF,PCA,Rose
AUC	0.776	0.728	0.786	0.893
Accuracy	0.8376	0.9204	0.9299	0.9236
Sensitivity	0.8601	0.9659	0.9522	0.9283
Specifity	0.5238	0.2857	0.619	0.8571

Indicator

Confusion Matrix : Log. Regression

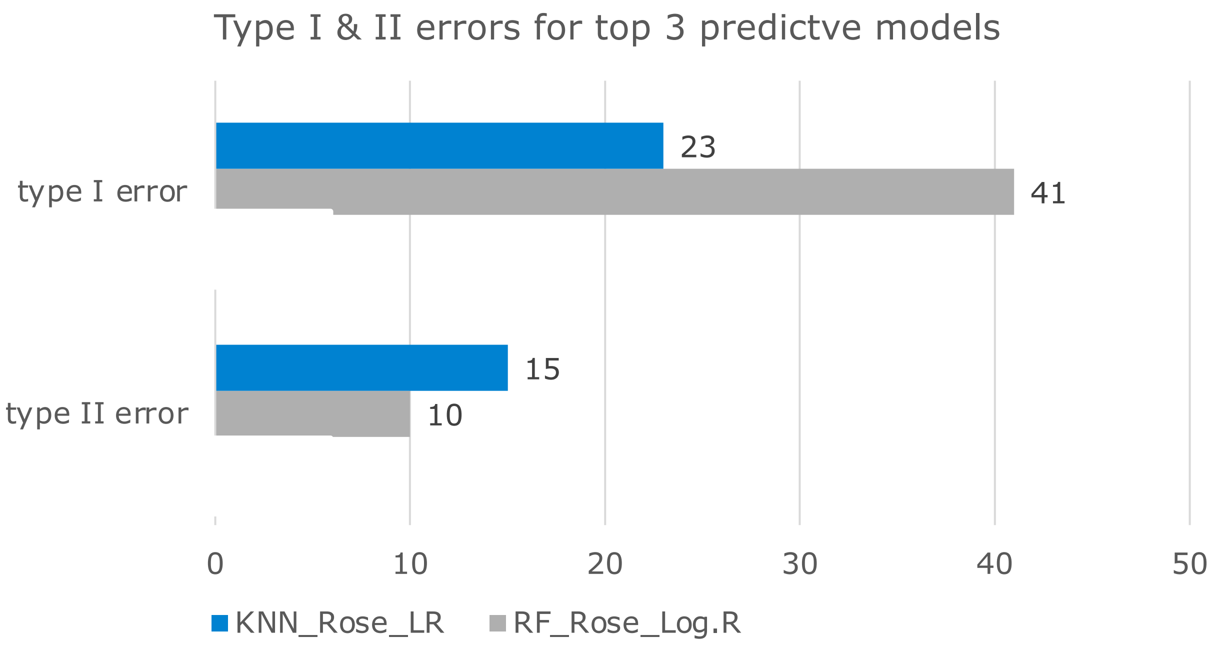
Definition: Consider a semi conductor for faulty detection

1 True = Fail
-1 False = Pass

			predicted class *Positive* 1	*Negative* -1	total
actual / observed class	*Positive*	1	TP	FN	
	Negative	-1	FP	TN	
		total			314

	RF_Rose_Boruta_Log.Reg	KNN_Rose_Boruta_Log.Reg
Faulty equipment and positive tested (TP)	252	283
Faulty equipment but negative tested - Type II error (FN)	41	10
Good quality equipment but positive tested - Type I error (FP)	10	15
Good quality equipment and negative tested (TN)	11	6

Confusion Matrix : Log. Regression



Summary

Parameter	PCA		Boruta	
	SVM	Naïve Bayes	Logistic Regression	Logistic Regression
AUC	0.89	0.78	0.78	0.73
ACCURACY	0.92	0.93	0.84	0.92
SENSITIVITY	0.93	0.95	0.86	0.84
SPECIFCITY	0.85	0.62	0.52	0.29
Error I	3	8	41	23
Error II	21	14	10	15
Best Predictive Model	1º	2º	3º	4º
Best Business Model	4º	3º	2º	1º
Conclusion	Difficult to implement	Difficult to implement	A lot of semiconductors are discarded despite of being OK. Higher costs but less risk.	It is not statistically the best model. Fewer working semiconductors are discarded but has more risk.

Good Practices learnt

Programming:

- ✓ Naming convention and consistency.
- ✓ Saving intermediate outputs for reusability.
- ✓ Switching steps and sequences for improving results.
- ✓ Segregation of scripts for readability and reusability.
- ✓ Computer power availability

- ✓ How to handle imbalanced dataset
- ✓ Effect of each data handling techniques to imbalanced dataset
- ✓ Practice with various classifiers