Design of high bulk moduli high entropy alloys using machine learning

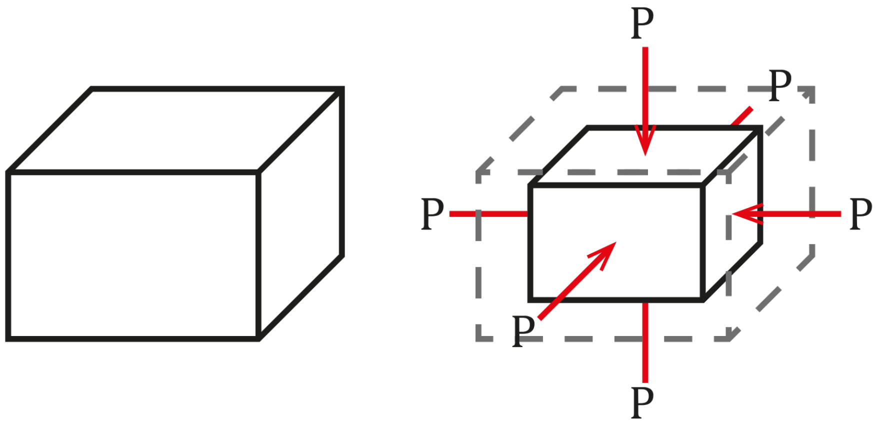
Professor: Dr. Parastar

Presenters: Erfan Nasrollahi – Setayesh Khalili

Introduction to Chemometrics course
Spring 2024

High Entropy Alloy

- Composed of 5 or more principal elements in approximately equal atomic percentage
- High entropy due to random distribution stabilize the solid solution phase
- Extraordinary properties (eg. High bulk modulus)



$$k = -V\frac{dP}{dV}$$

k = bulk modulus

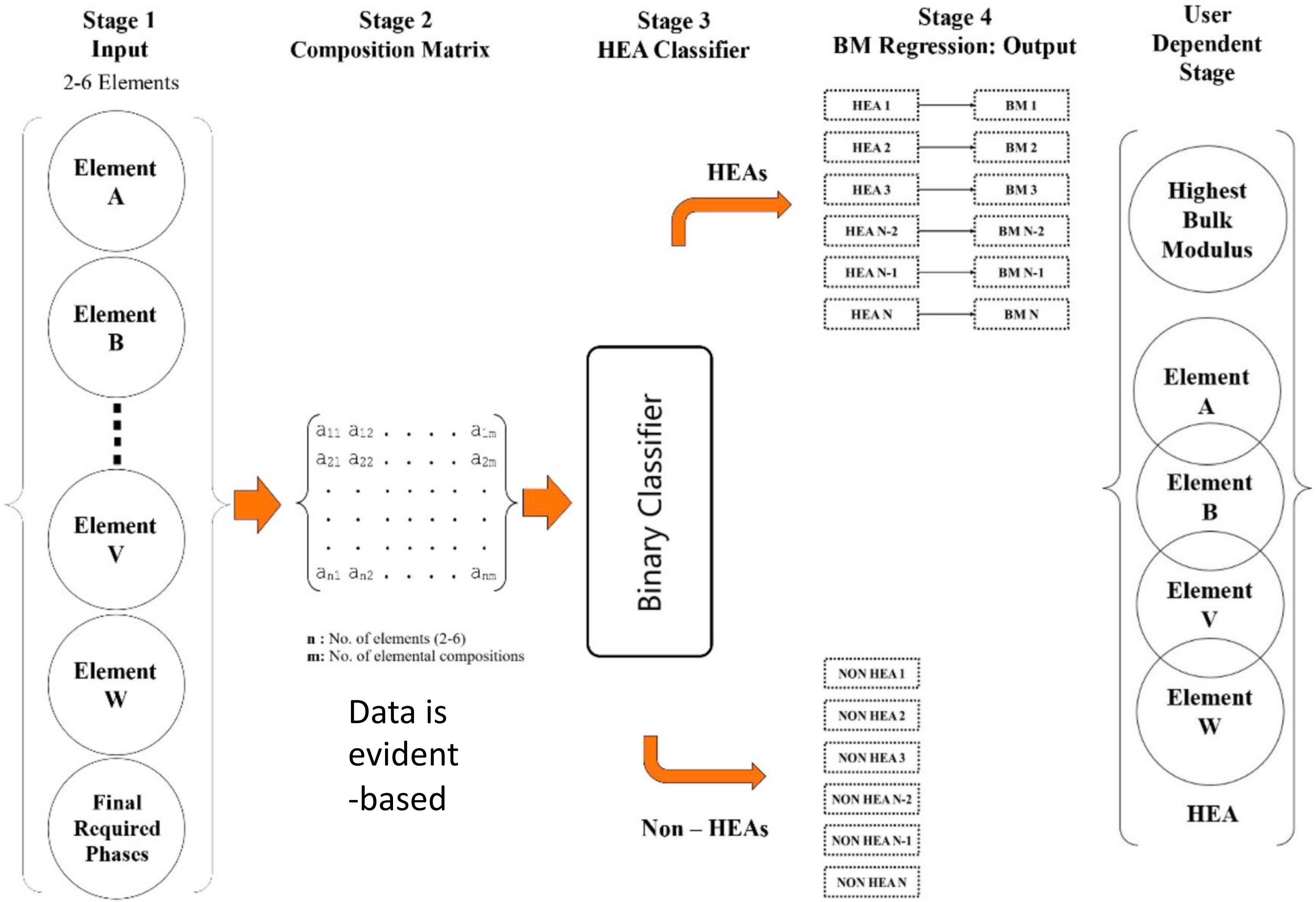
P = pressure

V = initial volume of the substance

ML has been used for optimizing
the composition of HEA to achieve enhanced bulk modulus values

George, E.P., Raabe, D. & Ritchie, R.O. High-entropy alloys. *Nat Rev Mater* **4**, 515–534 (2019). https://doi.org/10.1038/s41578-019-0121-4

Guide line



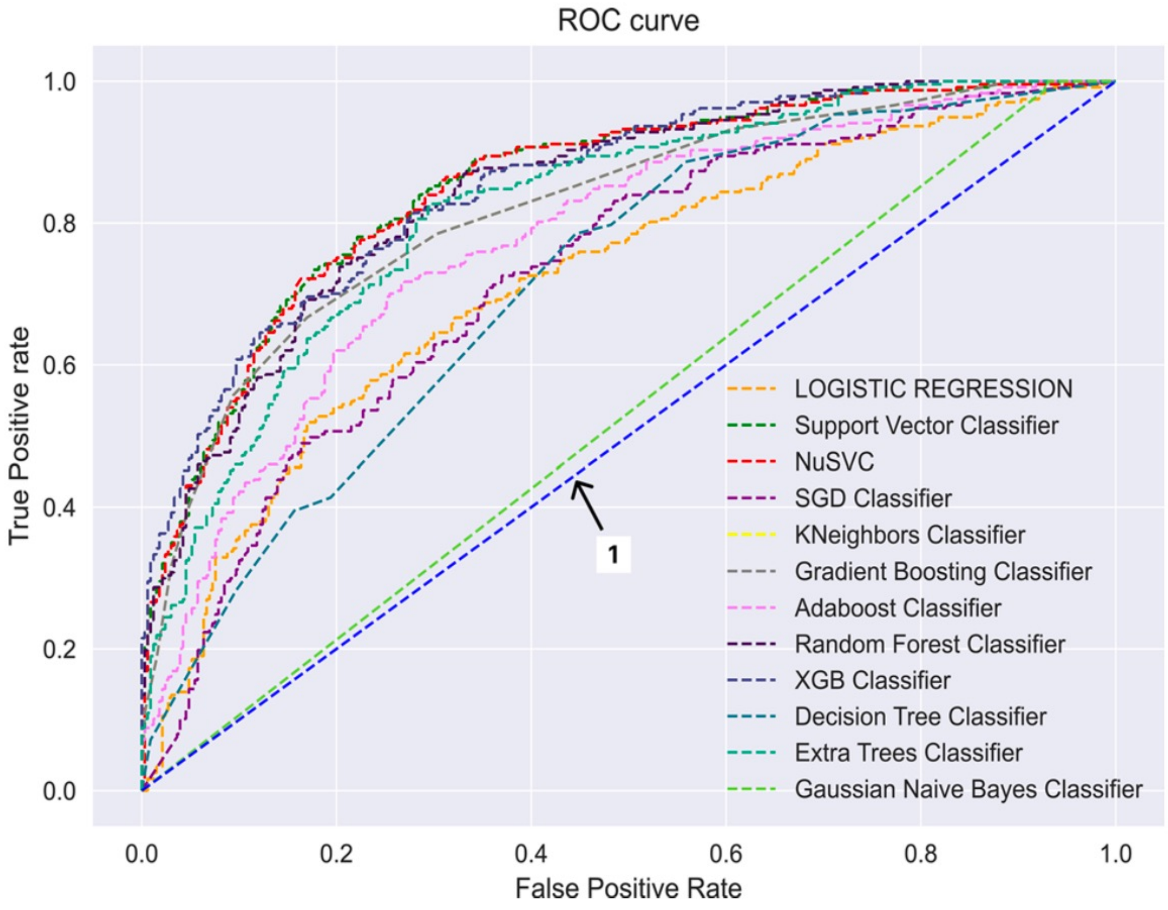
Data is evident -based

28.Dam,H.C.,Kino,H.&Ha,M.Q.High-entropyalloysdatasetsofevidence-basedrecommendersystemforcombinatorialmaterials synthesis (1.0.0). Zenodo (2021).

Binary Classification of HEA & Non-HEA

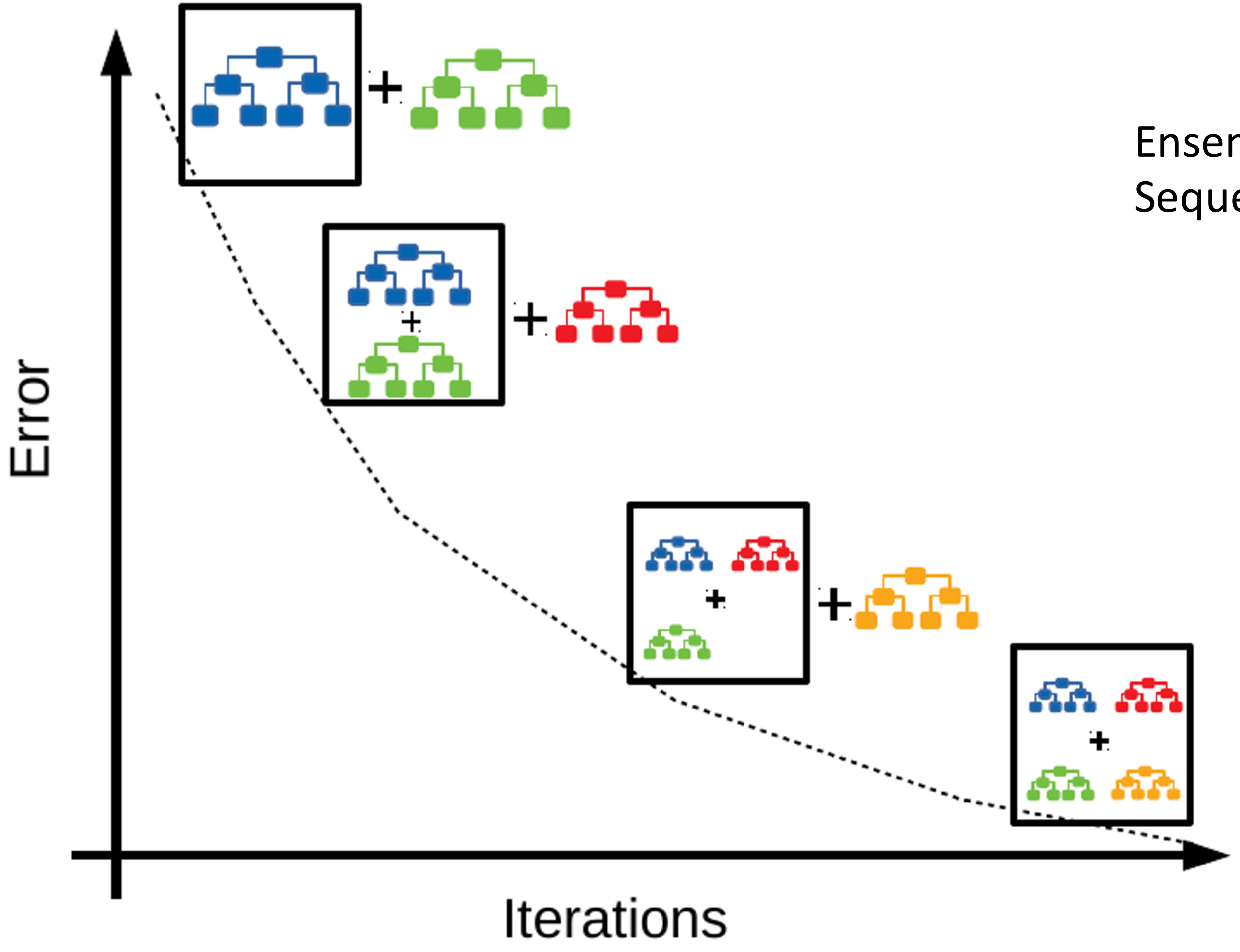
12 ML algorithms

Linear Regression
SVC(linear)
NuSVC
SGD
KNC
Gradient Boosting Classification
AC
RFC
XGB
Decision Three Classification
ETC
Gaussian Naive B



ROC Curves of Twelve HEA classifiers

Gradient Boosted Classification



Ensemble ML algorithm
Sequential Decision Tree Classifications

Anshul Saini,(2024) **Gradient Boosting Algorithm: A Complete Guide for Beginners,**Data Science Blogathon

Regression

seven regression models were trained to predict the bulk modulus of HEAs

RF
XGB
Linear
Lasso
Ridge
ElasticNet
KNN

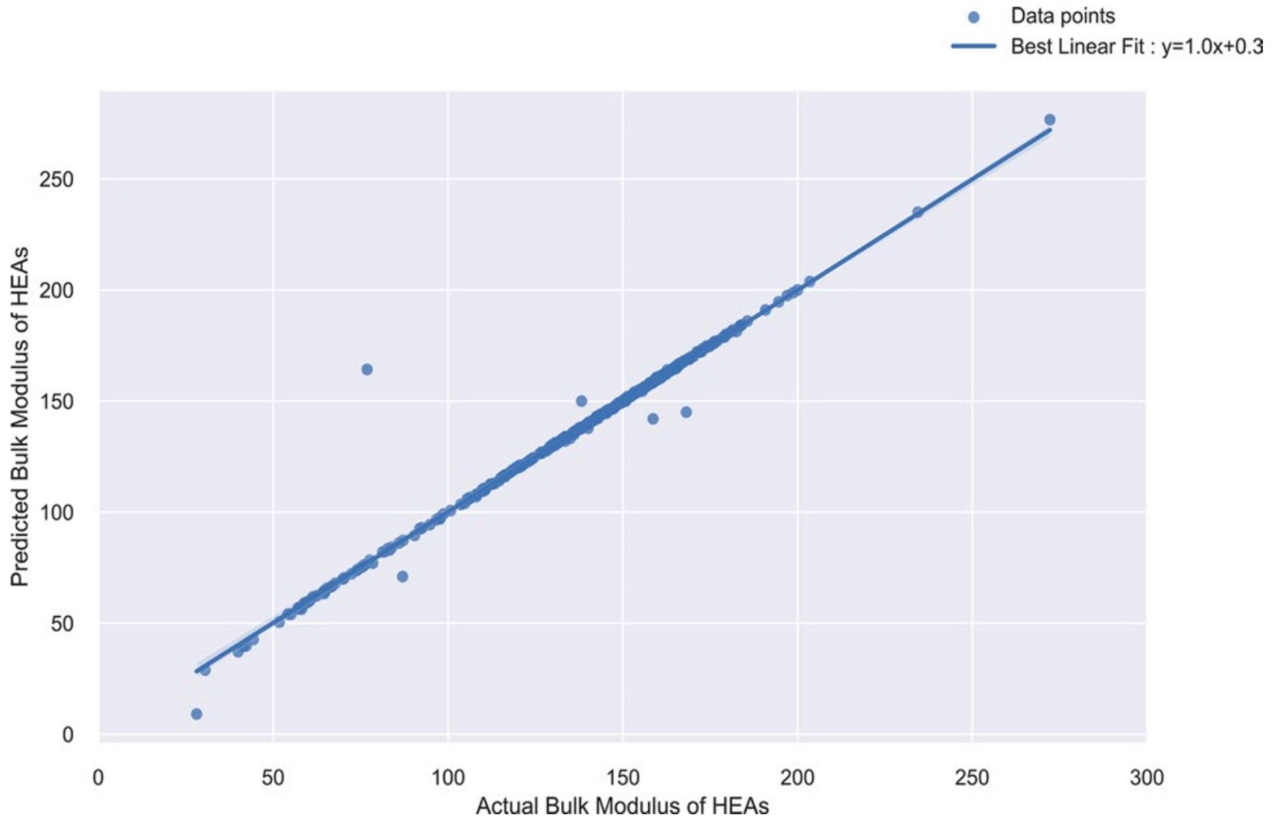


Figure 7. Predicted vs. actual bulk moduli for HEAs.

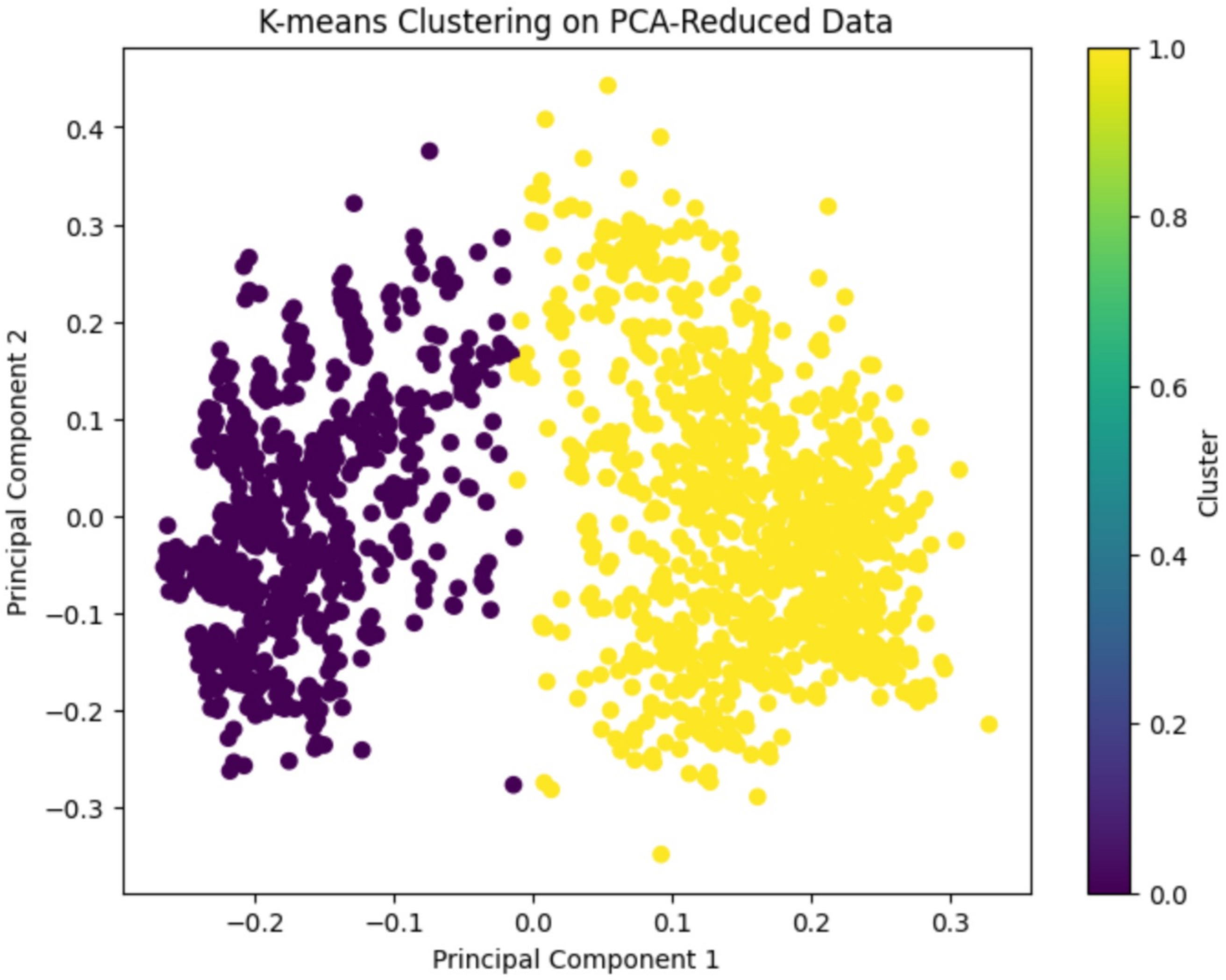
Our Analysis

Classification

Model	Accuracy					Precision					Recall				
	train		Test		cv	Train		test		cv	train		test		cv
	article	us	article	us		article	us	article	us		article	us	article	us	
LR	0.71	0.70	0.69	0.70	0.70	0.73	0.69	0.66	0.68	0.68	0.56	0.60	0.54	0.60	0.58
SVC(linear)	0.87	0.70	0.78	0.68	0.70	0.87	0.88	0.76	0.92	0.69	0.83	0.88	0.70	0.83	0.58
NuSVC	0.86	0.90	0.78	0.79	0.78	0.87	0.92	0.76	0.78	0.76	0.82	0.86	0.69	0.72	0.71
SGD	0.70	0.68	0.66	0.68	0.69	0.66	0.76	0.58	0.78	0.67	0.69	0.38	0.65	0.4	0.62
KNC	0.80	0.85	0.76	0.72	0.73	0.76	0.82	0.74	0.70	0.69	0.73	0.83	0.67	0.69	0.70
GBC(Opt)	0.99	0.84	0.77	0.76	0.77	0.77	0.83	0.73	0.73	0.75	0.98	0.79	0.71	0.74	0.69
AC	0.76	0.72	0.72	0.73	0.72	0.72	0.68	0.67	0.70	0.68	0.68	0.69	0.65	0.73	0.66
RFC	0.93	1.0	0.76	0.78	0.79	0.76	1.0	0.75	0.79	0.78	0.89	1.0	0.65	0.70	0.71
XGB	0.99	0.94	0.77	0.78	0.79	0.77	0.94	0.75	0.77	0.78	0.99	0.92	0.70	0.71	0.73
DTC	0.69	1.0	0.65	0.75	0.75	0.65	1.0	0.56	0.71	0.71	0.80	1.0	0.78	0.73	0.71
ETC	0.87	1.0	0.74	0.79	0.79	0.74	1.0	0.74	0.77	0.78	0.74	1.0	0.57	0.77	0.72
GNB	0.48	0.47	0.46	0.47	0.47	0.46	0.45	0.43	0.45	0.45	0.43	1.0	0.99	1.0	1.0
LDA	-	0.70	-	0.70	0.70	-	0.69	-	0.68	0.66	-	0.63	-	0.63	0.62
SVM(non-linear)	-	0.86	-	0.78	0.78	-	0.86	-	0.78	0.76	-	0.82	-	0.71	0.70

Model	F1_score					AUC_ROC				
	train		test		cv	train		test		cv
	article	us	article	us		article	us	article	us	
LR	0.63	0.63	0.59	0.63	0.62	0.70	0.75	0.67	0.72	0.74
SVC	0.85	0.88	0.73	0.88	0.62	0.87	0.94	0.77	0.92	0.74
NuSVC	0.84	0.87	0.72	0.75	0.74	0.86	0.97	0.76	0.87	0.84
SGD	0.67	0.50	0.61	0.52	0.63	0.79	0.76	0.66	0.73	0.74
KNC	0.77	0.83	0.70	0.69	0.70	0.79	0.92	0.75	0.80	0.80
GBC	0.99	0.80	0.72	0.73	0.72	0.99	0.91	0.76	0.83	0.84
AC	0.71	0.68	0.66	0.71	0.67	0.75	0.82	0.71	0.80	0.79
RFC	0.92	1.0	0.70	0.74	0.74	0.93	1.0	0.75	0.85	0.87
XGB	0.99	0.93	0.72	0.72	0.74	0.99	0.99	0.76	0.86	0.87
DTC	0.70	1.0	0.65	0.72	0.71	0.70	1.0	0.67	0.75	0.74
ETC	0.84	1.0	0.64	0.77	0.75	0.86	1.0	0.72	0.86	0.87
GNB	0.63	0.62	0.60	0.62	0.62	0.53	0.52	0.53	0.52	0.52
LDA	-	0.65	-	0.65	0.64	-	0.77	-	0.75	0.75
SVM	-	0.84	-	0.74	0.73	-	0.84	-	0.87	0.86

PCA and k-means



Gradient Booster Classification model optimization by hyper tunning

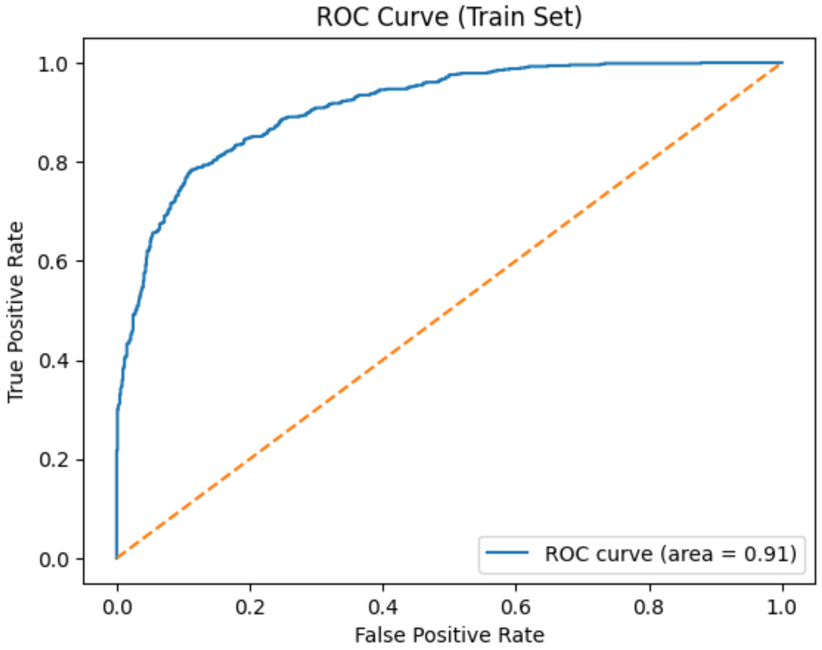
The optimization is performed by hyper-parameter tuning using an open-source Python library Hypopt
Hypopt uses a Grid Search algorithm to choose the best combination of hyperparameters after exploring various combinations of hyperparameters

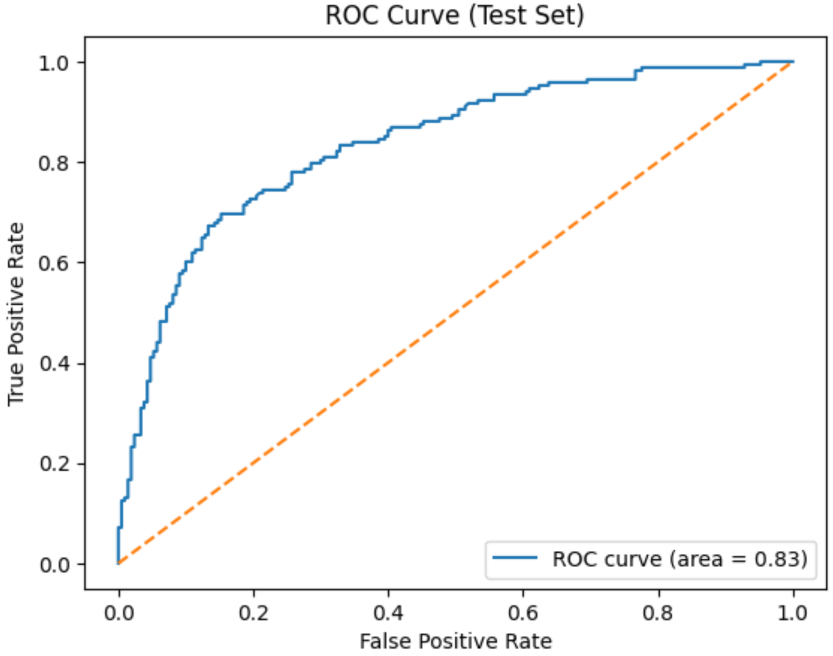
Parameters	Value
Loss	Deviance
n_estimators	100
max_depth	8
random_state	0

Northcutt,C.G.Hypopt1.0.9.(2019)https://pypi.org/project/Hypopt.

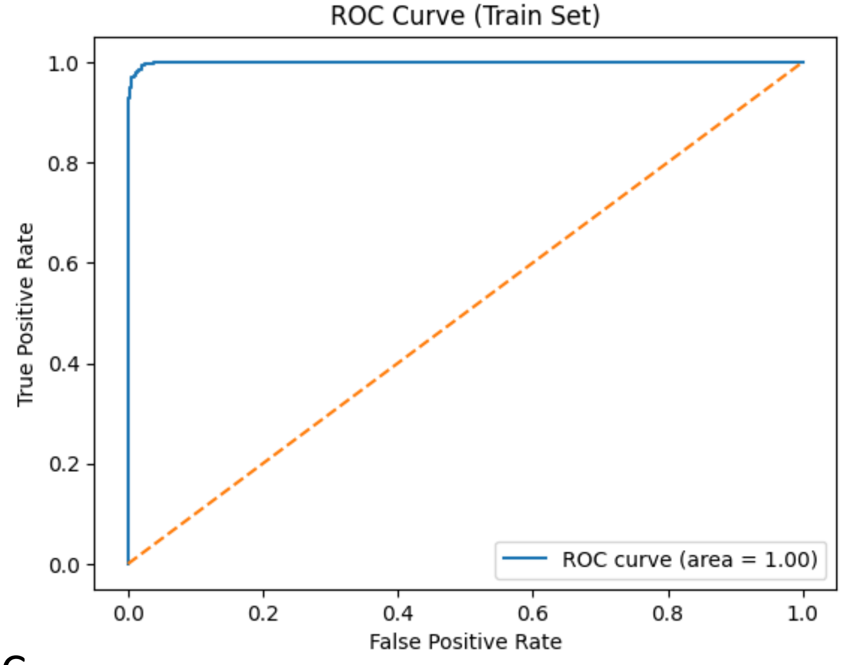
Gradient Booster Classification optimization with hyper tunning

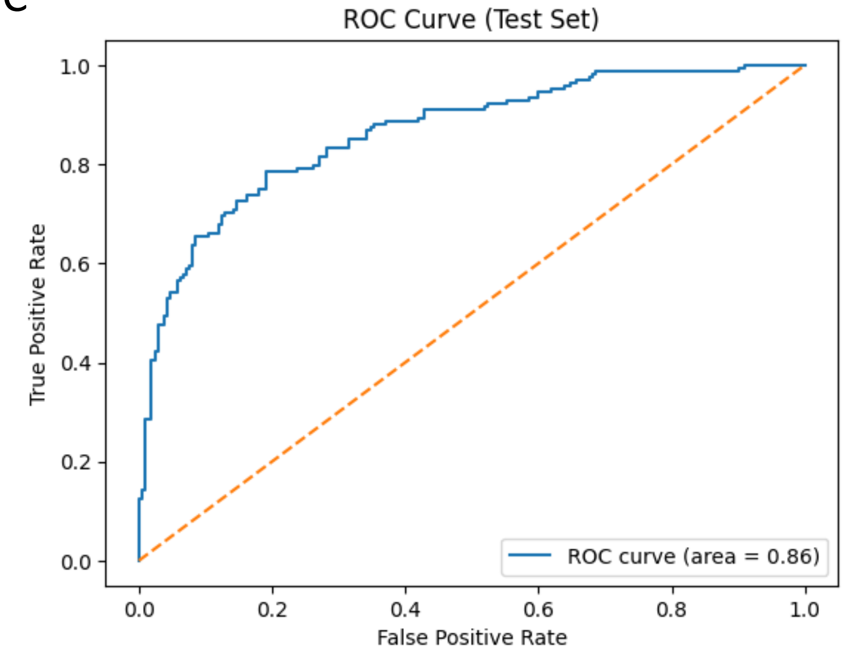
GBC





Optimized GBC

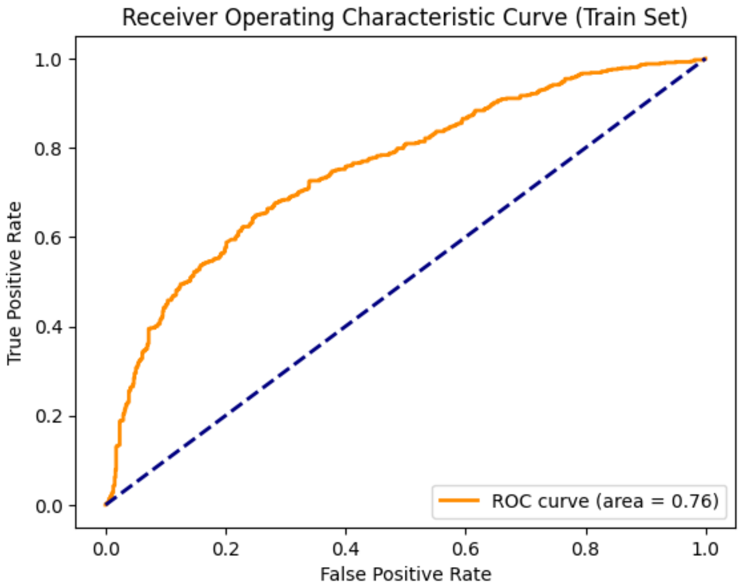




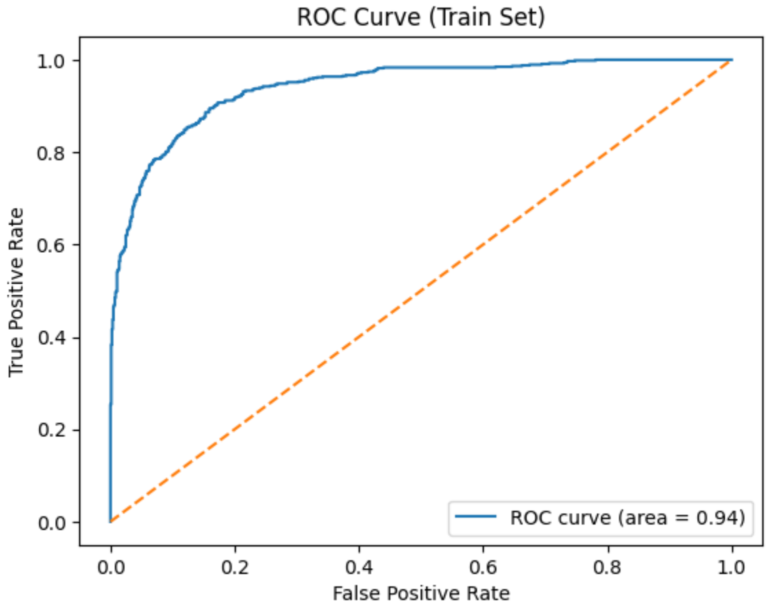
Gradient Booster Classification optimization with hyper tunning

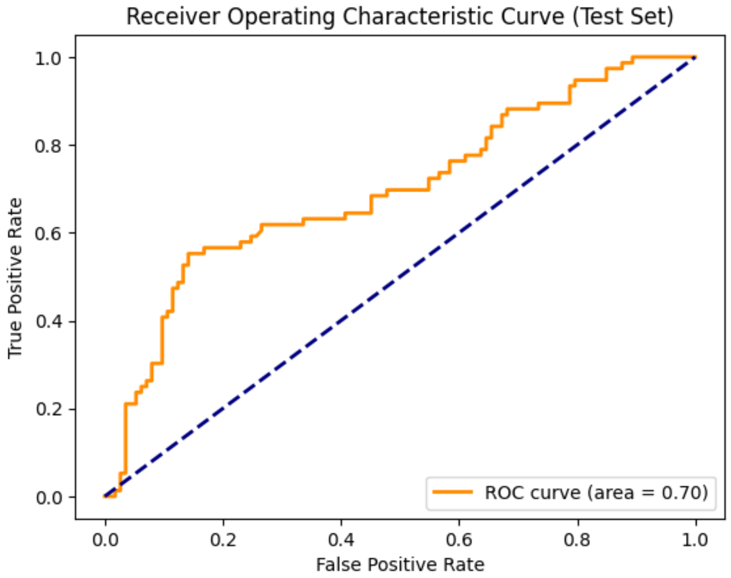
model	accuracy				precision				recall				F1_score				AUC_ROC			
	train		test		train		test		train		test		train		test		train		test	
	article	us	article	us	article	us	article	us	article	us	article	us	article	us	article	us	article	us	article	us
GBC	0.99	0.84	0.77	0.76	0.77	0.83	0.73	0.73	0.98	0.79	0.71	0.74	0.99	0.80	0.72	0.73	0.99	0.91	0.76	0.83
GBC opt	0.98	0.98	0.77	0.98		0.98	0.73	0.77		0.98	0.71	0.78		0.98	0.72	0.78		0.99	0.76	0.86

Model Cross validation	accuracy	Precision	Recall	F1_score	AUC_ROC
GCB	0.77	0.75	0.70	0.72	0.84
GCB OPT	0.78	0.76	0.71	0.73	0.86

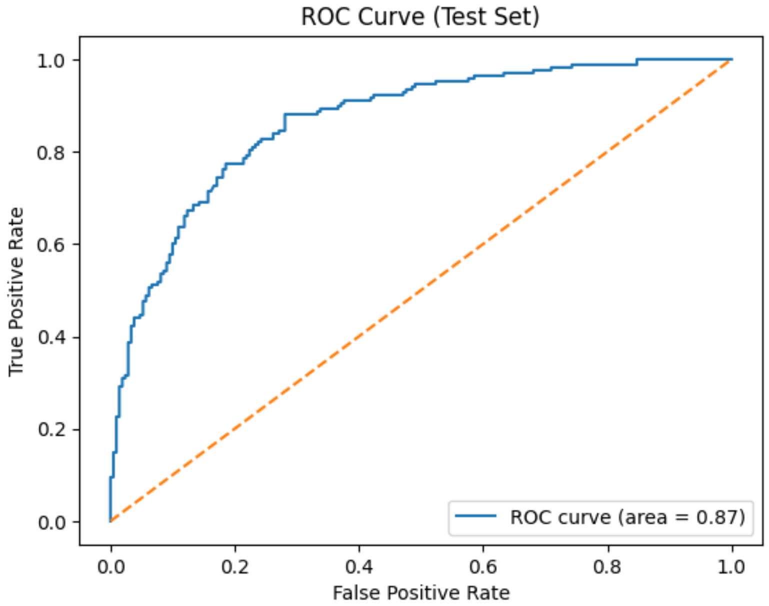


SVC vs SVM





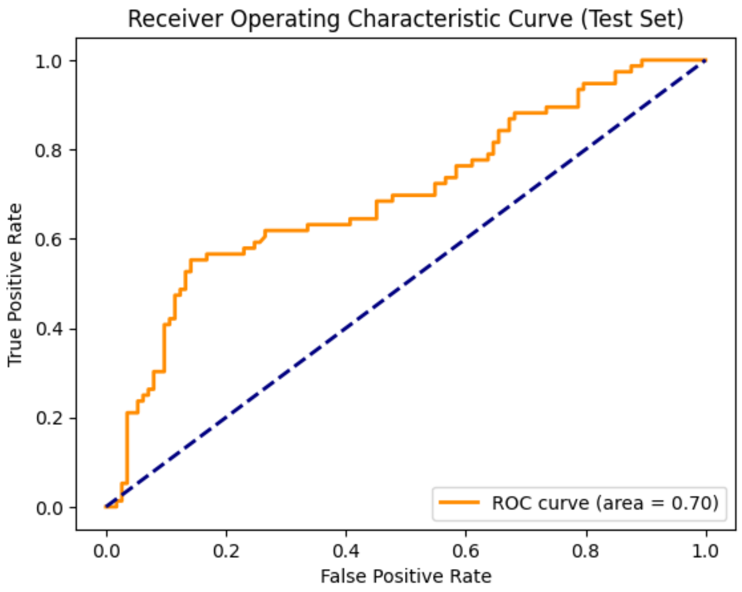
If the hyperplane classifies the dataset linearly then the algorithm we call it as SVC and the algorithm that separates the dataset by non-linear approach then we call it as SVM

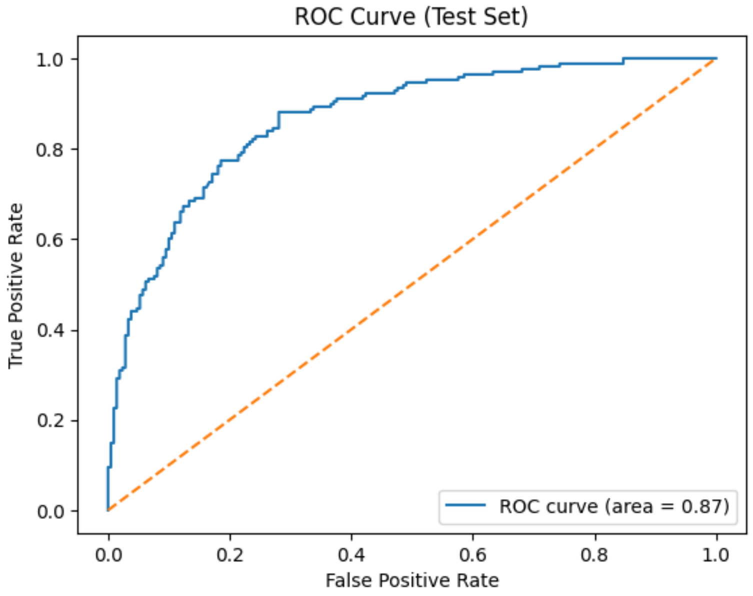


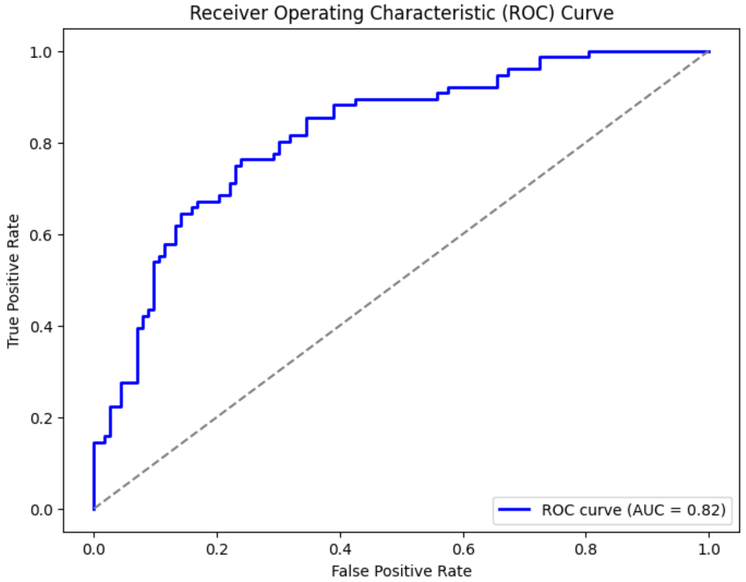
Premanand S,(2023) **The A-Z guide to Support Vector Machine,**Data Science Blogathon

SVC vs SVCopt vs SVM

Model Cross validation	accuracy	Precision	Recall	F1_score	AUC_ROC
SVC	0.70	0.68	0.58	0.62	0.73
SVC opt	0.75	0.66	0.76	0.71	0.82
SVM	0.78	0.76	0.70	0.73	0.86







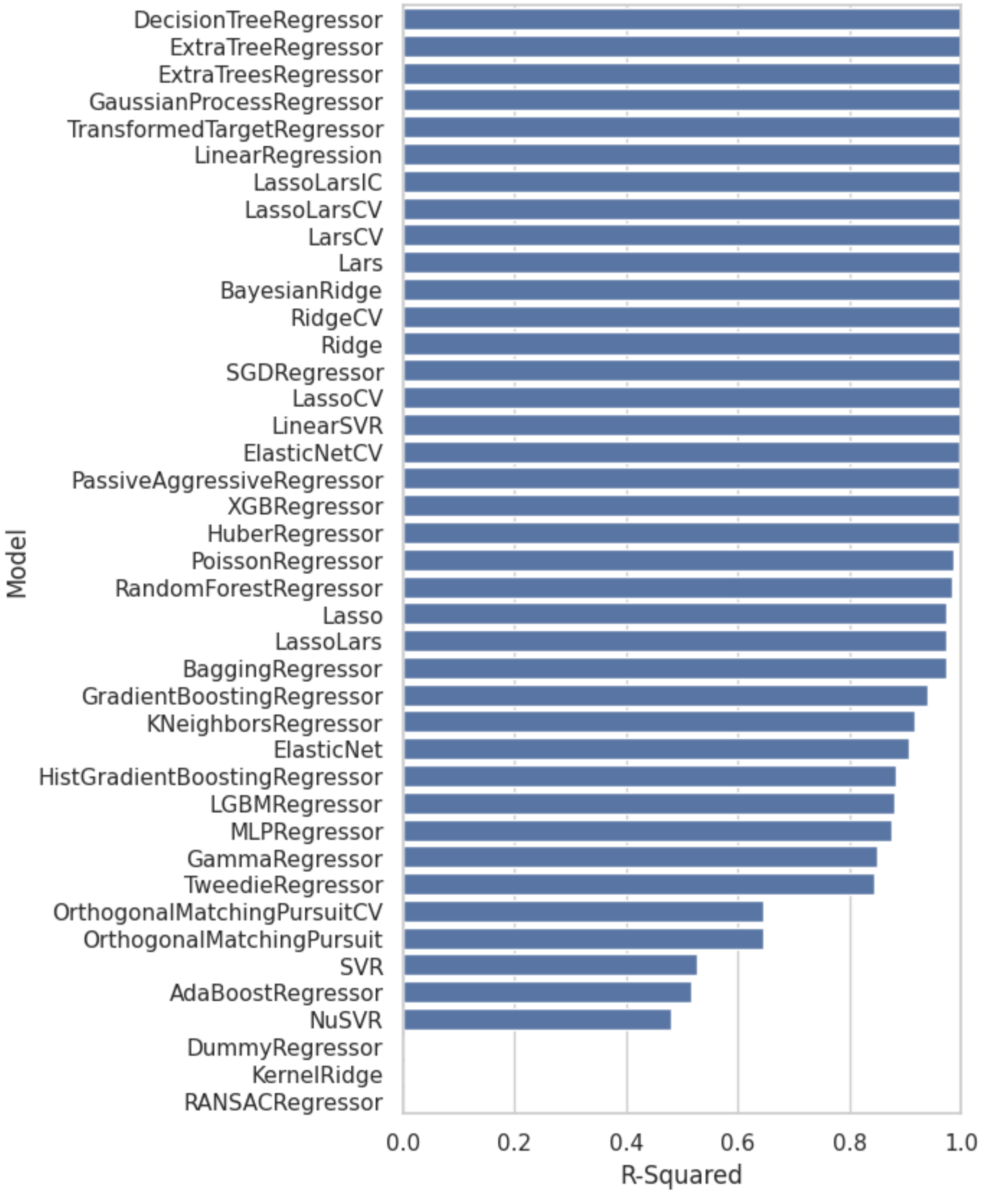
BM regression

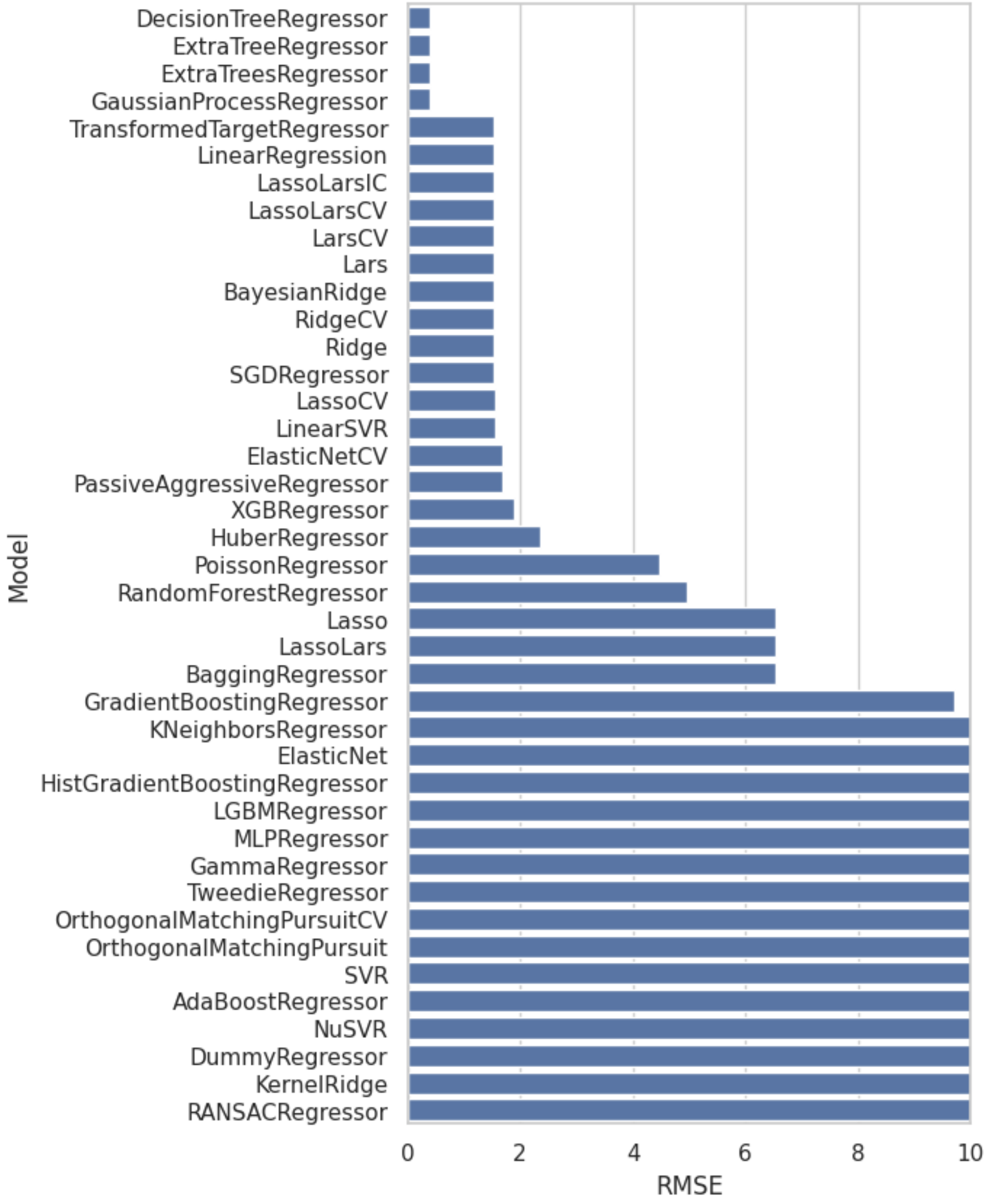
Model	R-square				Adj. R-square			
	Train		Test		Train		Test	
	article	us	article	us	article	us	article	us
RF	0.99	0.987	0.86	0.888	0.97	0.982	0.83	0.827
XGB	0.99	0.997	0.90	0.997	0.99	0.920	0.88	0.876
Linear	0.99	0.998	-4e10	-1.8e10	0.99	0.998	-5e10	-2.8e10
Lasso	0.99	0.412	0.98	0.372	0.99	0.358	0.97	0.0482
Ridge	0.99	0.936	0.97	0.917	0.99	0.930	0.97	0.872
ElasticNet	0.99	0.163	0.96	0.082	0.99	0.082	0.95	-0.316
KNN	0.96	0.901	0.86	0.837	0.96	0.892	0.82	0.749
Extra Tree's Regressor	-	0.999977	-	0.94506	-	0.99997	-	0.9150
Decision Tree Regressor	-	0.999977	-	0.84066	-	0.99997	-	0.7534

Model	Mean Squared Error (MSE)				Mean Absolute Error (MAE)				Mean Absolute Percentage Error (MAPE)			
	Train		Test		Train		Test		Train		Test	
	article	us	article	us	article	us	article	us	article	us	article	us
RF	29.71	24.58	180.41	166.67	2.43	2.15	6.55	6.35	2.36	43.97	6.95	39.00
XGB	0.07	3.96	127.42	118.96	0.05	1.45	5.46	5.75	0.03	44.57	5.88	39.36
Linear	1.54	2.36	6e13	2.75	0.29	0.323	4e5	10938707.45	0.24	0.269	5e5	8913964.09
Lasso	1.68	925.929	26.98	936.06	0.37	21.57	0.84	22.06	0.36	37.12	1.26	32.67
Ridge	1.56	100.44	30.14	122.55	0.29	4.79	0.79	5.35	0.25	6.034	0.97	5.65
ElasticNet	7.58	1319.66	49.63	1269.38	1.12	27.95	1.93	26.84	1.19		3.22	31.47
KNN	49.11	155.22	191.70	241.87	3.60	6.66	7.65	8.83	3.37	7.89	8.31	9.34
ETR	-	0.035	-	81.93	-	0.012	-	4.47	-	44.66	-	39.42
DTR	-	0.035	-	237.65	-	0.012	-	8.01	-	44.66	-	40.15

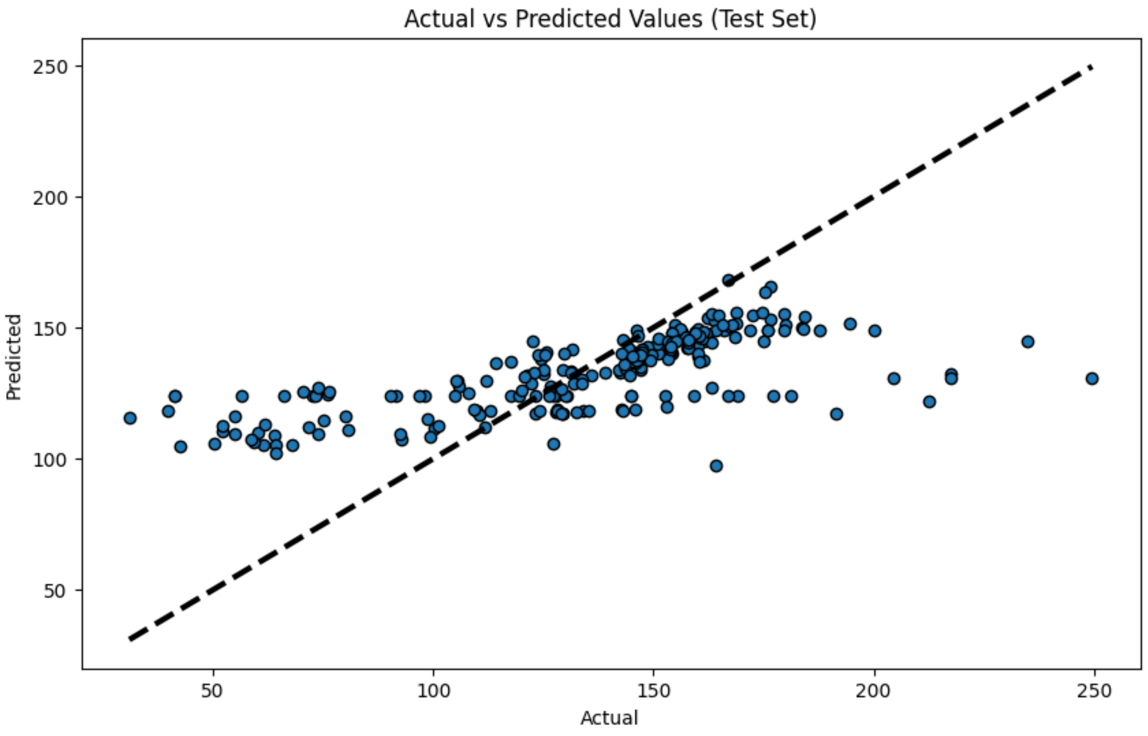
Lazy predict

```
pip install lazypredict
```

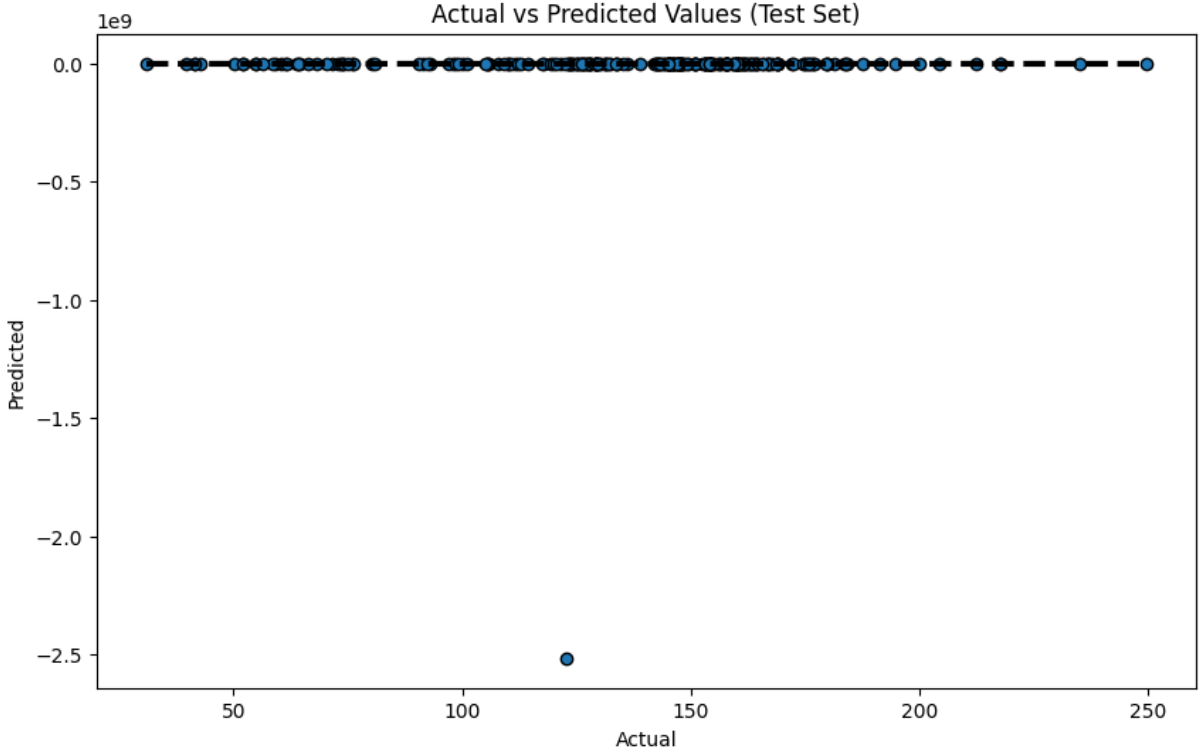




Lasso & Linear regression did not work for us!

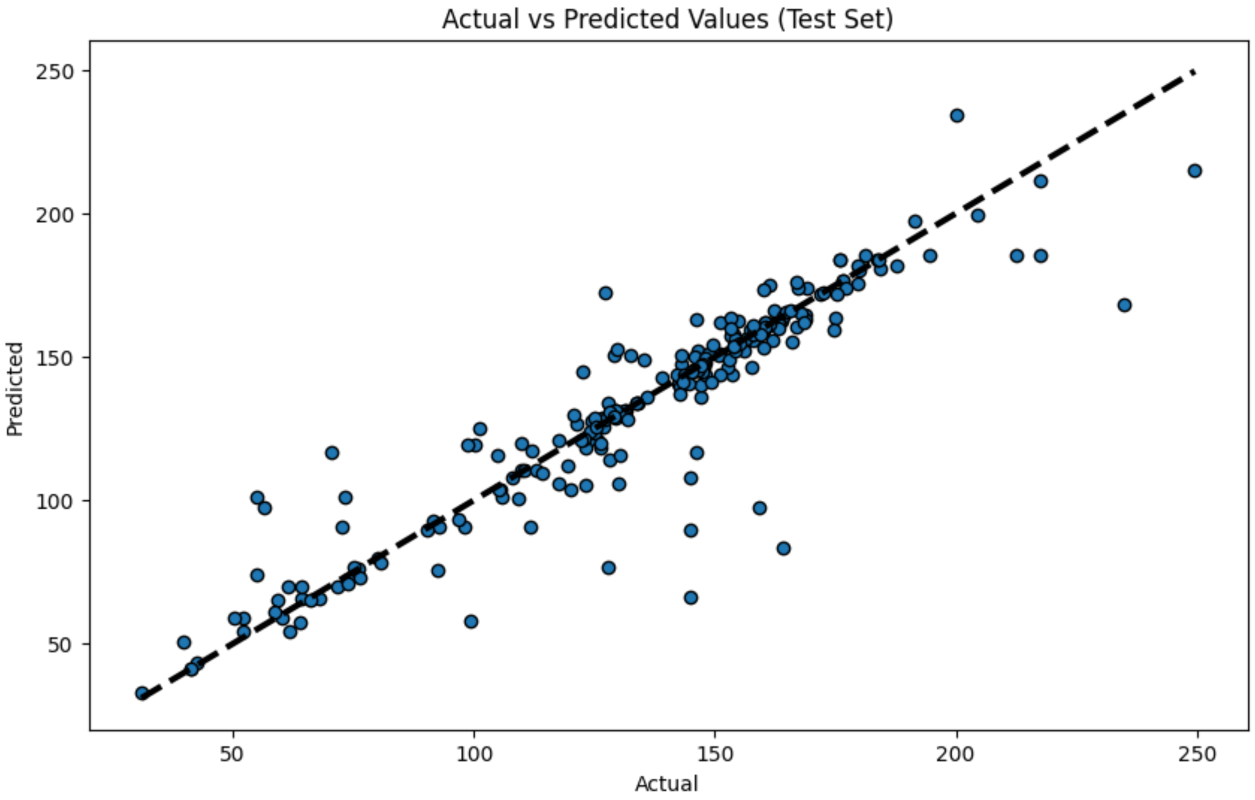


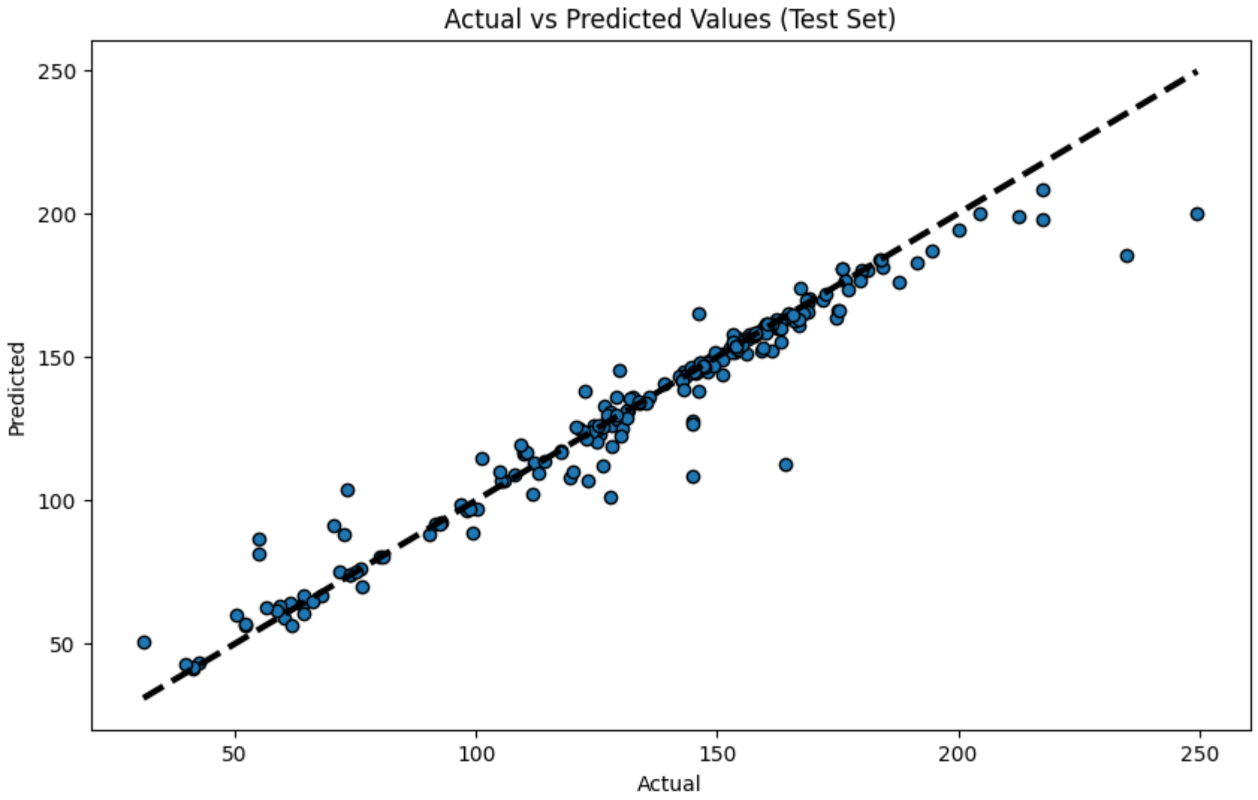
Lasso plot



Linear plot

DTR & ETR


Actual vs Predicted Values (Test Set)
Predicted
Actual
250
200
150
100
50
50
100
150
200
250


Actual vs Predicted Values (Test Set)
Predicted
Actual
250
200
150
100
50
50
100
150
200
250

Conclusion

- The Article proposed:
  - GBC as the best model for HEA classification
  - Lasso as the best BM regression model

- Our results:
  - GBC and SVM as the best models for HEA classification
  - ETR as the best BM regression model
  - Data set did not follow any linear attribute

Thank you for your attention