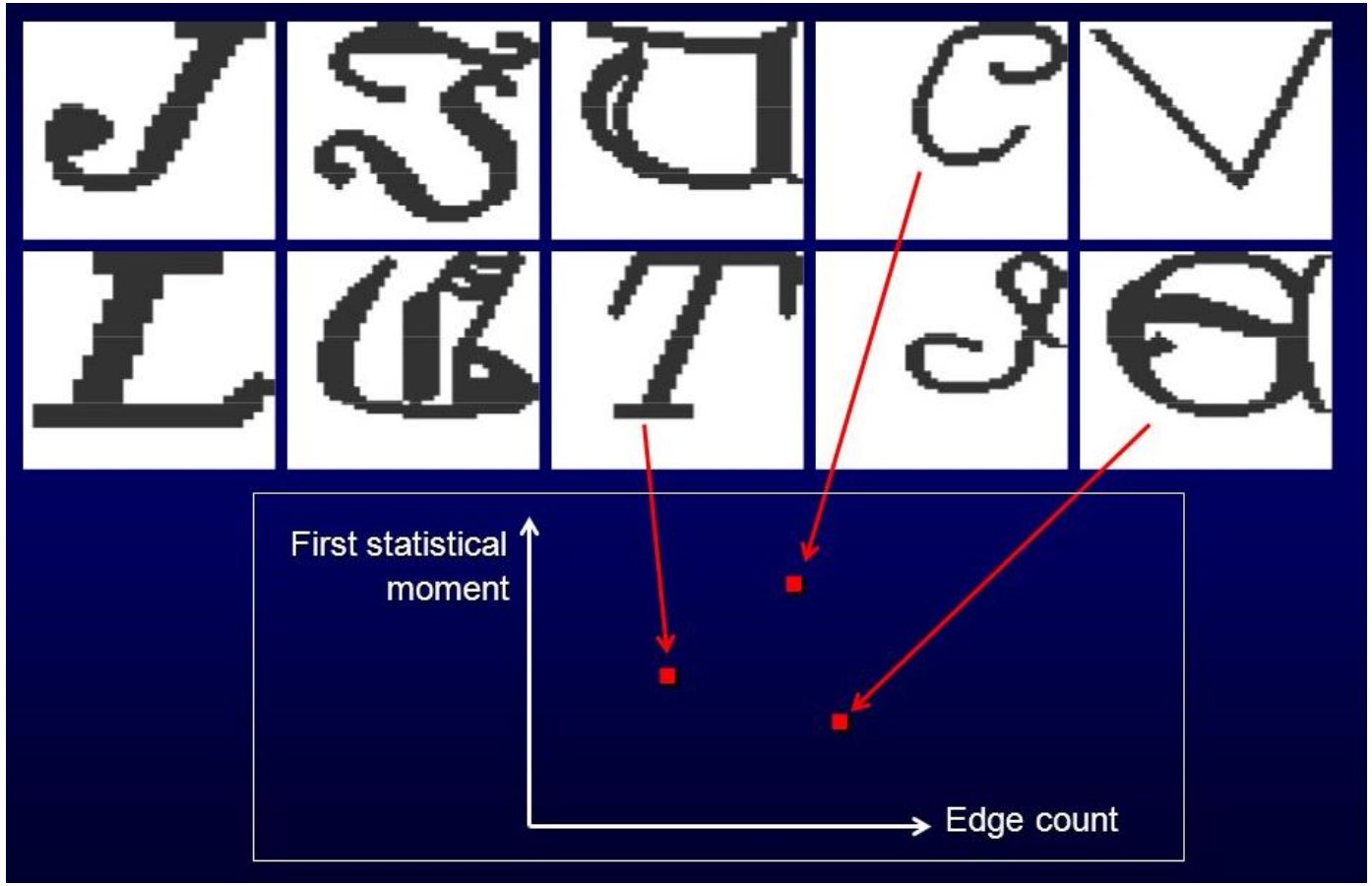
LETTER RECOGNITION : MULTICLASS CLASSIFIER

Machine Learning Intermediate

By Chiranjit Pathak

LETTER RECOGNITION

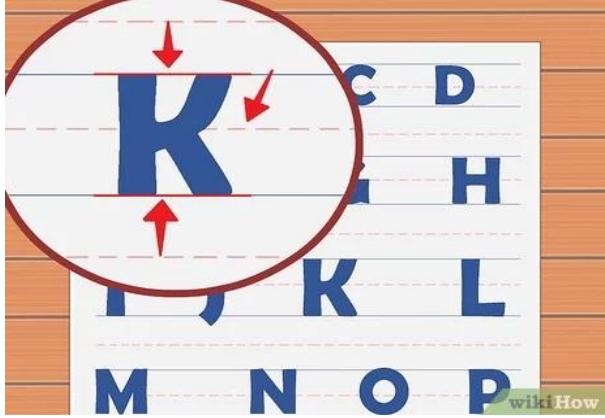
The character images were based on **20 different fonts** and each letter within these 20 fonts was randomly distorted to produce a file of **20,000 unique stimuli**. Each stimulus was converted into **16 primitive numerical attributes (statistical moments and edge counts)** which were then scaled to fit into a range of integer values from 0 through 15.



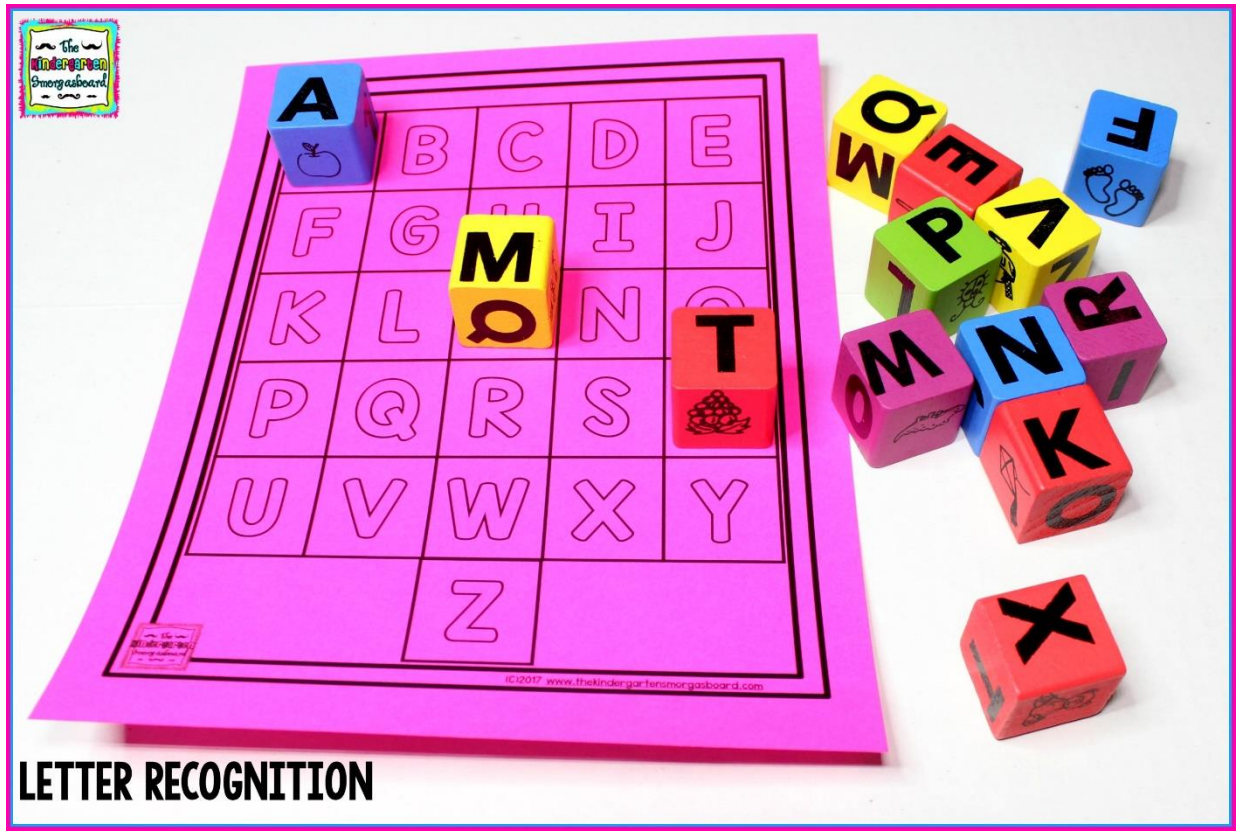
PROBLEM STATEMENT

The objective is to identify each of a large number of **black-and-white rectangular pixel** displays as one of the 26 capital letters in the English alphabet

DATA LOADING AND DESCRIPTION



Source of the data:
https://archive.ics.uci.edu/ml/datasets/letter+recognition



Column Name	Description
letter	capital letter (26 values from A to Z) interval
x-box	horizontal position of box(integer)
y-box	vertical position of box (integer)
width	width of box (integer)
high	height of box (integer)
onpix	total # on pixels (integer)
x-bar	mean x of on pixels in box (integer)
y-bar	mean y of on pixels in box (integer)
x2bar	mean x variance (integer)
y2bar	mean y variance (integer)
xybar	mean x y correlation (integer)
x2ybar	mean of x * x * y (integer)
xy2bar	mean of x * y * y (integer)
xedge	mean edge count left to right (integer)
xedgey	correlation of xedge with y (integer)
yedge	mean edge count bottom to top (integer)
yedgex	correlation of yedge with x (integer)

EXPLANATORY DATA ANALYSIS

Processing, Profiling and Analysis

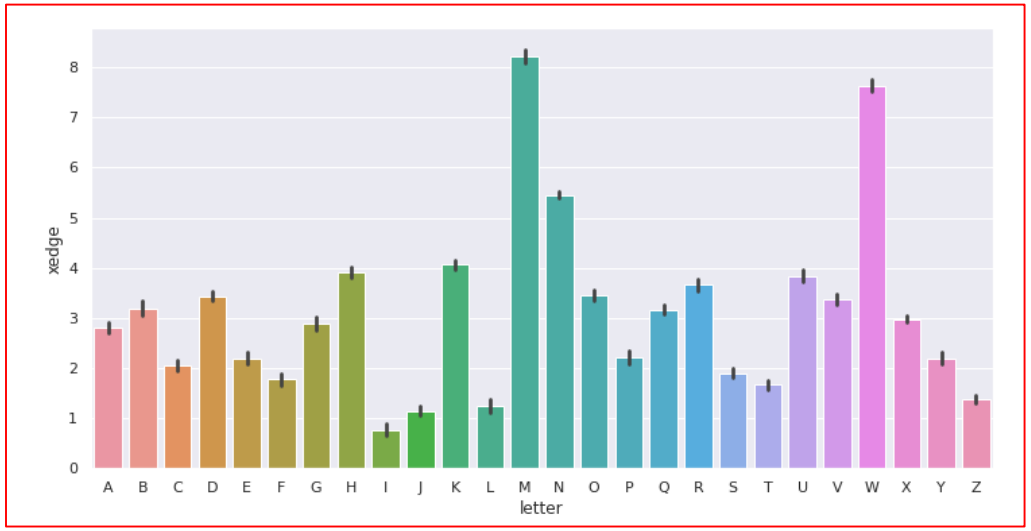
PROCESSING AND EDA

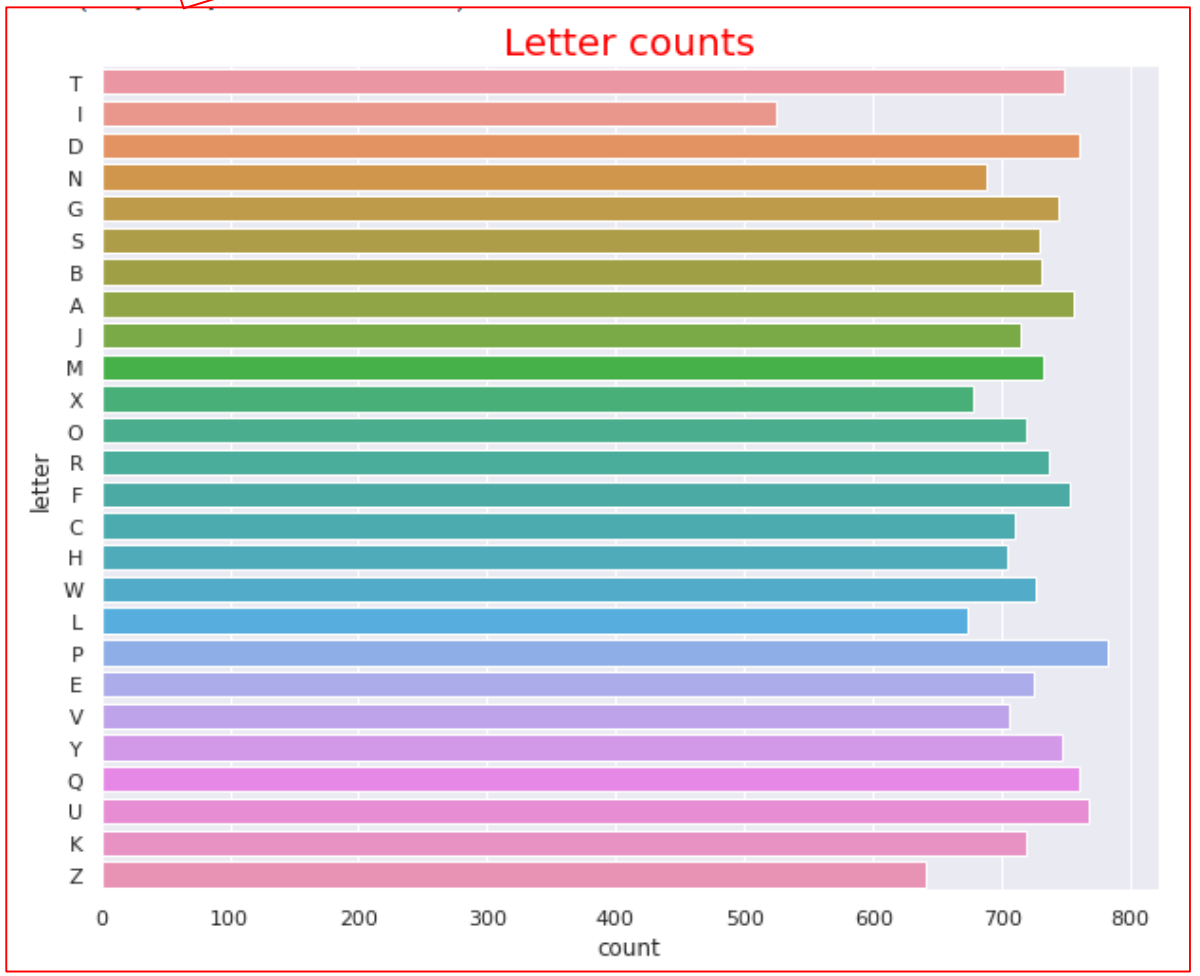
Balanced among all classes

	letter	xbox	ybox	width	height	onpix	xbar	ybar	x2bar	y2bar	xybar	x2ybar	xy2bar	xedge	xedgey	yedge	yedgex
0	T	2	8	3	5	1	8	13	0	6	6	10	8	0	8	0	8
1	I	5	12	3	7	2	10	5	5	4	13	3	9	2	8	4	10
2	D	4	11	6	8	6	10	6	2	6	10	3	7	3	7	3	9

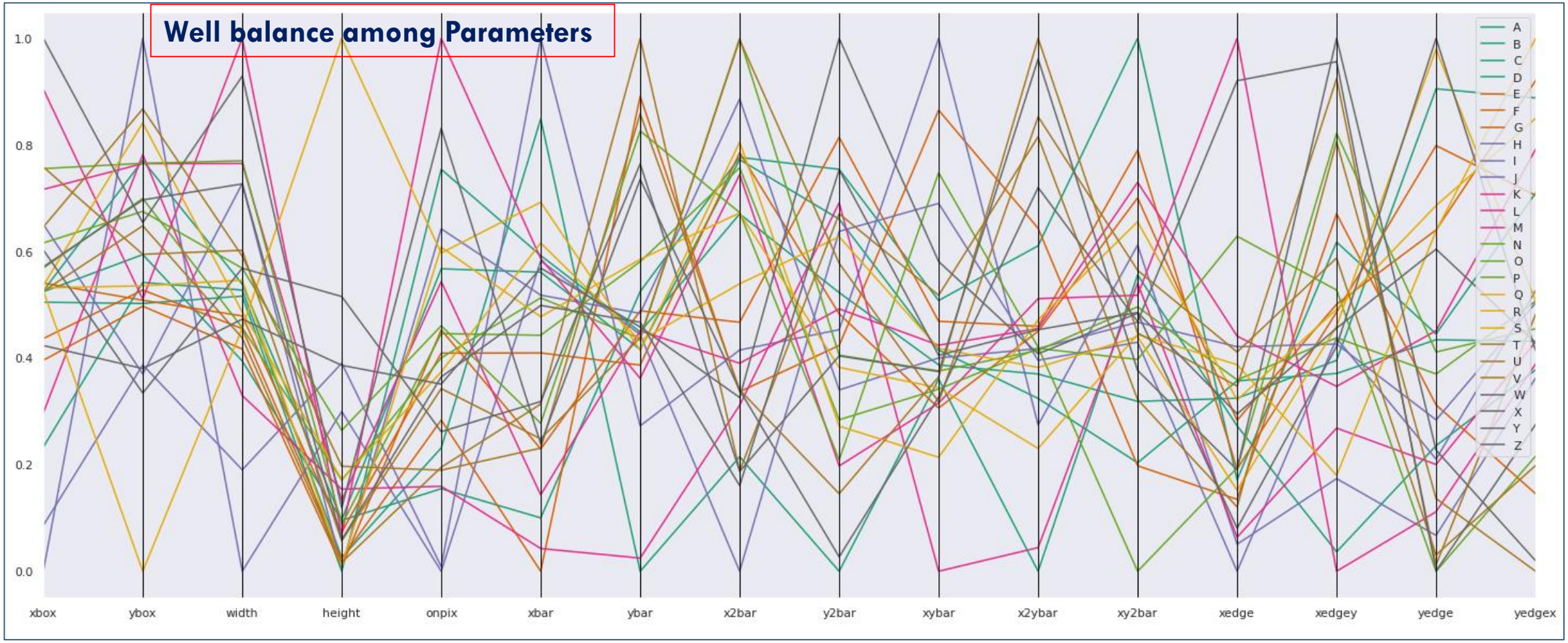
Duplicate removed

```
Data Shape [Before]: (20000, 17)
Data Shape [After]: (18668, 17)
Drop Ratio: 6.660000000000001 %
```





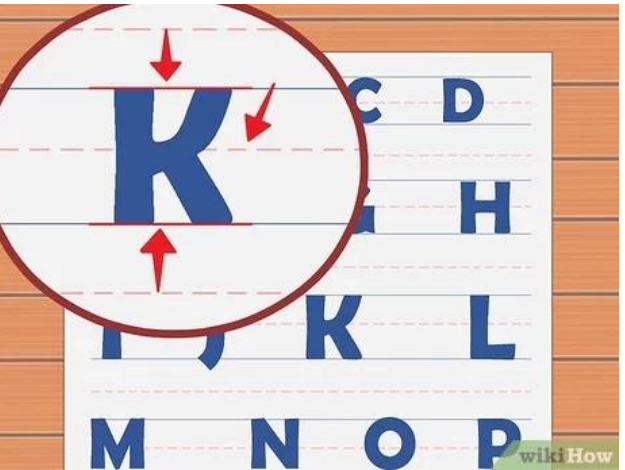
MEAN STIMULI PARAMETERS OF CLASSES



CORRELATION AMONG FEATURES

Feature Selection to be done

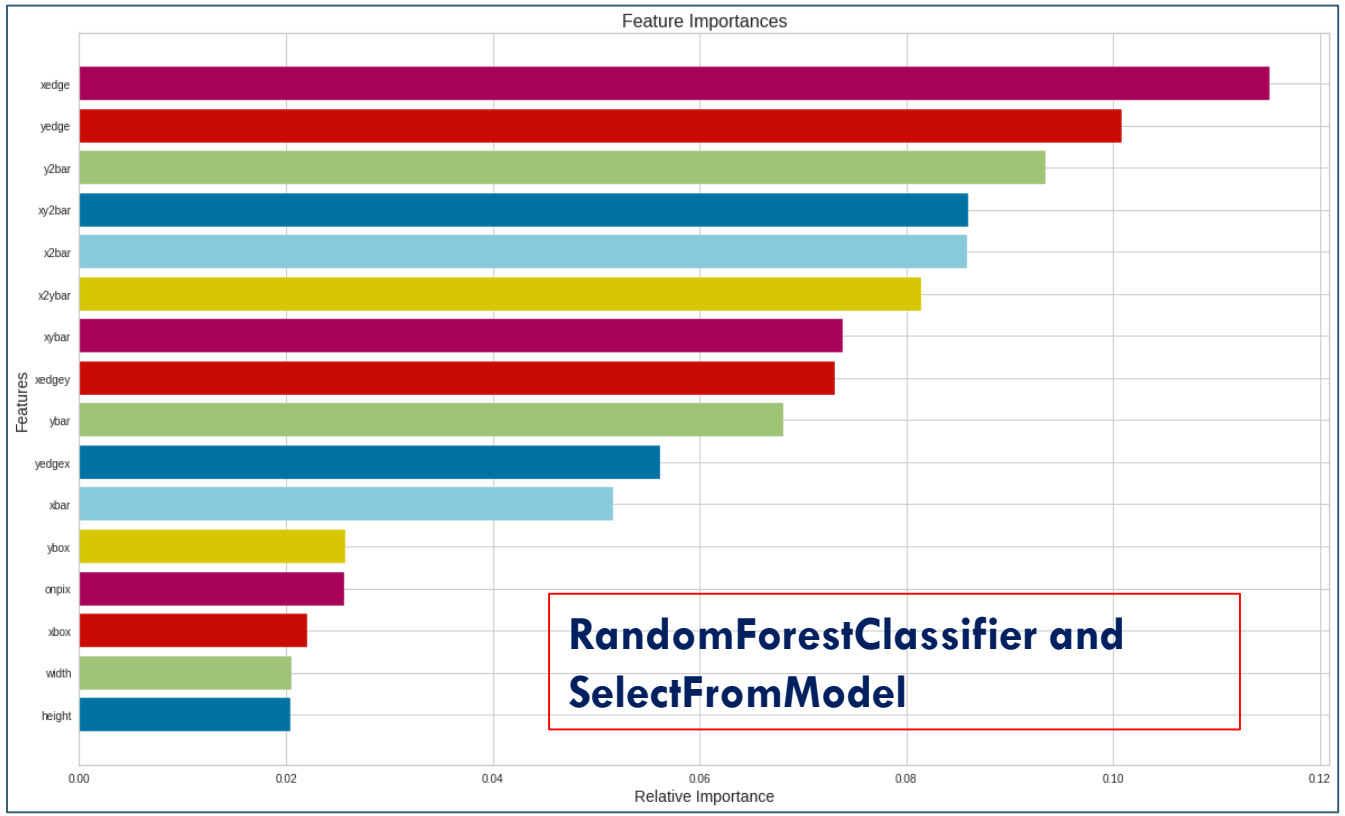
	xbox	ybox	width	height	onpix	xbar	ybar	x2bar	y2bar	xybar	x2ybar	xy2bar	xedge	xedgey	yedge	yedgex
xbox	1.00	0.75	0.84	0.66	0.59	-0.04	0.04	0.01	0.04	0.14	0.03	-0.05	0.47	0.09	0.25	-0.10
ybox	0.75	1.00	0.66	0.80	0.53	0.04	-0.06	-0.02	0.08	0.15	-0.06	-0.00	0.26	-0.01	0.21	-0.03
width	0.84	0.66	1.00	0.65	0.75	0.06	0.01	-0.11	0.04	0.09	0.00	-0.04	0.55	0.04	0.23	-0.11
height	0.66	0.80	0.65	1.00	0.63	0.04	-0.03	0.09	0.03	-0.01	-0.01	0.03	0.25	0.02	0.29	-0.01
onpix	0.59	0.53	0.75	0.63	1.00	0.14	-0.05	-0.01	-0.10	-0.11	-0.08	-0.04	0.63	0.01	0.48	-0.05
xbar	-0.04	0.04	0.06	0.04	0.14	1.00	-0.39	-0.05	-0.13	0.06	-0.38	-0.04	0.14	-0.26	0.13	0.26
ybar	0.04	-0.06	0.01	-0.03	-0.05	-0.39	1.00	-0.11	-0.05	0.16	0.60	-0.28	-0.05	0.56	-0.10	-0.21
x2bar	0.01	-0.02	-0.11	0.09	-0.01	-0.05	-0.11	1.00	-0.16	-0.31	0.04	0.09	0.11	-0.08	0.03	0.19
y2bar	0.04	0.08	0.04	0.03	-0.10	-0.13	-0.05	-0.16	1.00	0.12	-0.05	0.12	-0.39	-0.06	0.24	-0.06
xybar	0.14	0.15	0.09	-0.01	-0.11	0.06	0.16	-0.31	0.12	1.00	0.04	-0.12	-0.20	0.02	-0.12	-0.11
x2ybar	0.03	-0.06	0.00	-0.01	-0.08	-0.38	0.60	0.04	-0.05	0.04	1.00	0.06	0.04	0.53	-0.23	-0.24
xy2bar	-0.05	-0.00	-0.04	0.03	-0.04	-0.04	-0.28	0.09	0.12	-0.12	0.06	1.00	-0.01	-0.19	0.05	0.24
xedge	0.47	0.26	0.55	0.25	0.63	0.14	-0.05	0.11	-0.39	-0.20	0.04	-0.01	1.00	0.00	0.11	-0.05
xedgey	0.09	-0.01	0.04	0.02	0.01	-0.26	0.56	-0.08	-0.06	0.02	0.53	-0.19	0.00	1.00	-0.07	-0.19
yedge	0.25	0.21	0.23	0.29	0.48	0.13	-0.10	0.03	0.24	-0.12	-0.23	0.05	0.11	-0.07	1.00	0.15
yedgex	-0.10	-0.03	-0.11	-0.01	-0.05	0.26	-0.21	0.19	-0.06	-0.11	-0.24	0.24	-0.05	-0.19	0.15	1.00

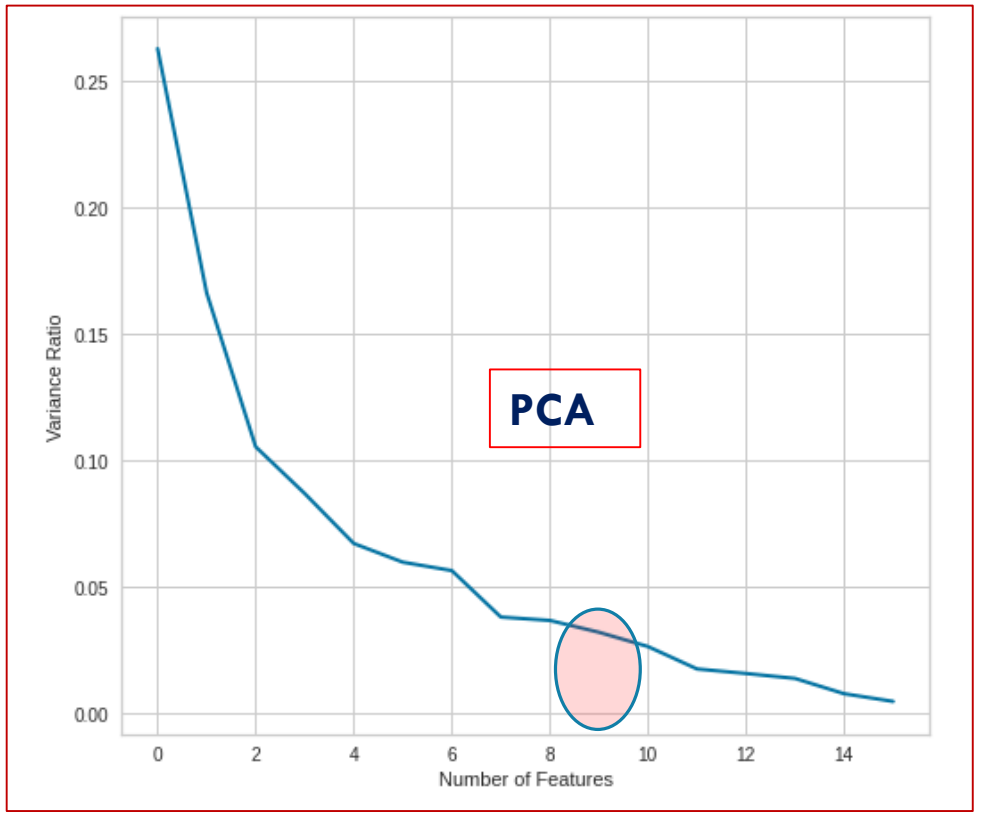


FEATURE SELECTION

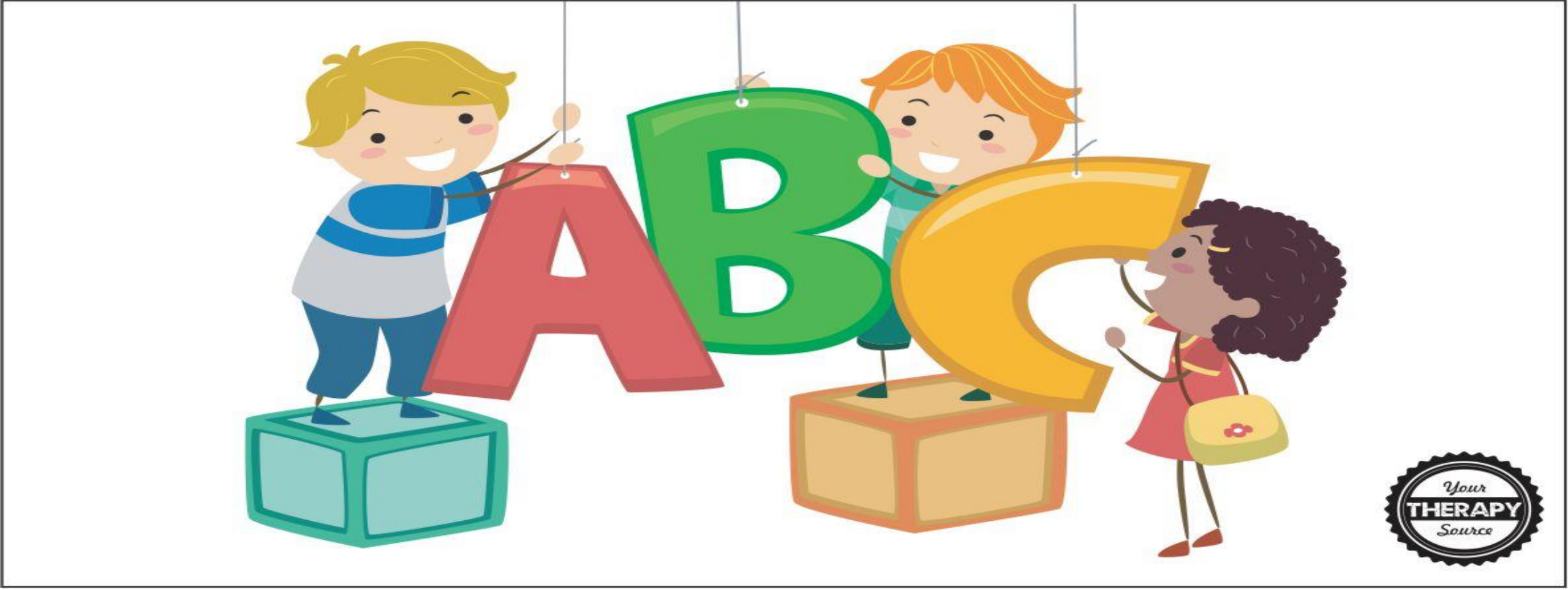
Select important features and PCA

BEST FEATURES AND PCA






```
Total Features Selected are 9
Threshold set by Model: 0.06
Features: ['ybar', 'x2bar', 'y2bar', 'xybar', 'x2ybar', 'xy2bar', 'xedge', 'xedgey', 'yedge']
```



MODEL FORMULATION

Multiclass Classification Models and Evaluation

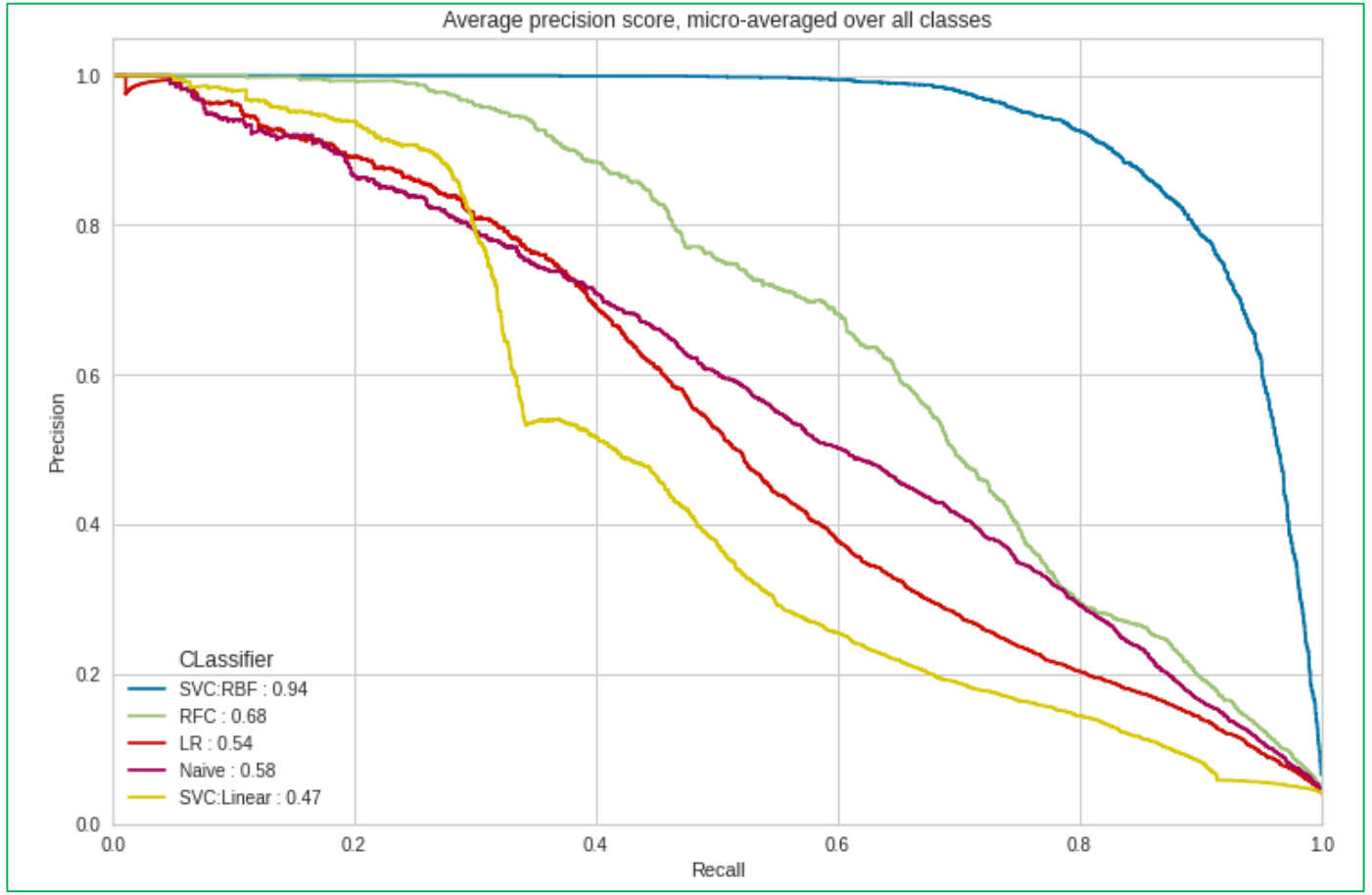
MODEL BUILDING AND EVALUATION


```
from sklearn import preprocessing
lb = preprocessing.LabelBinarizer()
lb.fit(order)
Y = lb.transform(y)
n_classes = Y.shape[1]
```

```
n_classes
```

```
26
```

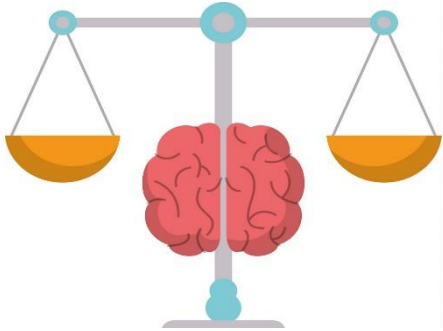
LabelBinarizer for multiclass encoding



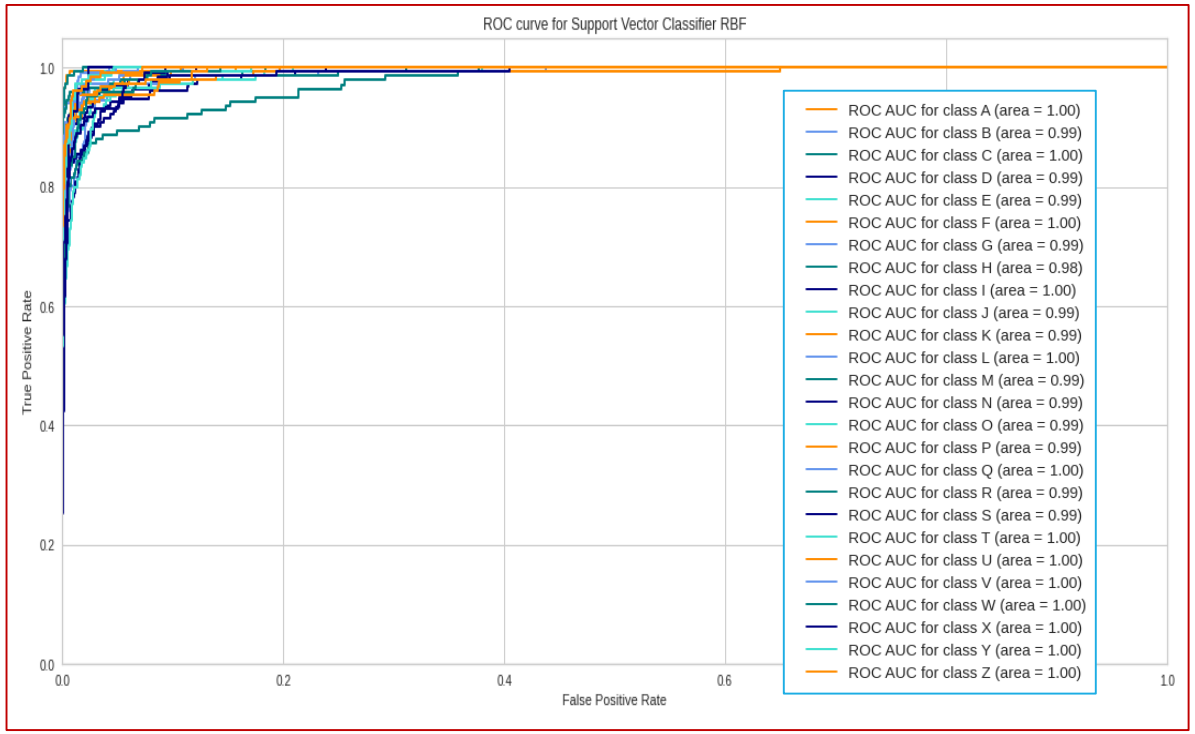
OneVsRestClassification technique has been employed using below algorithms,

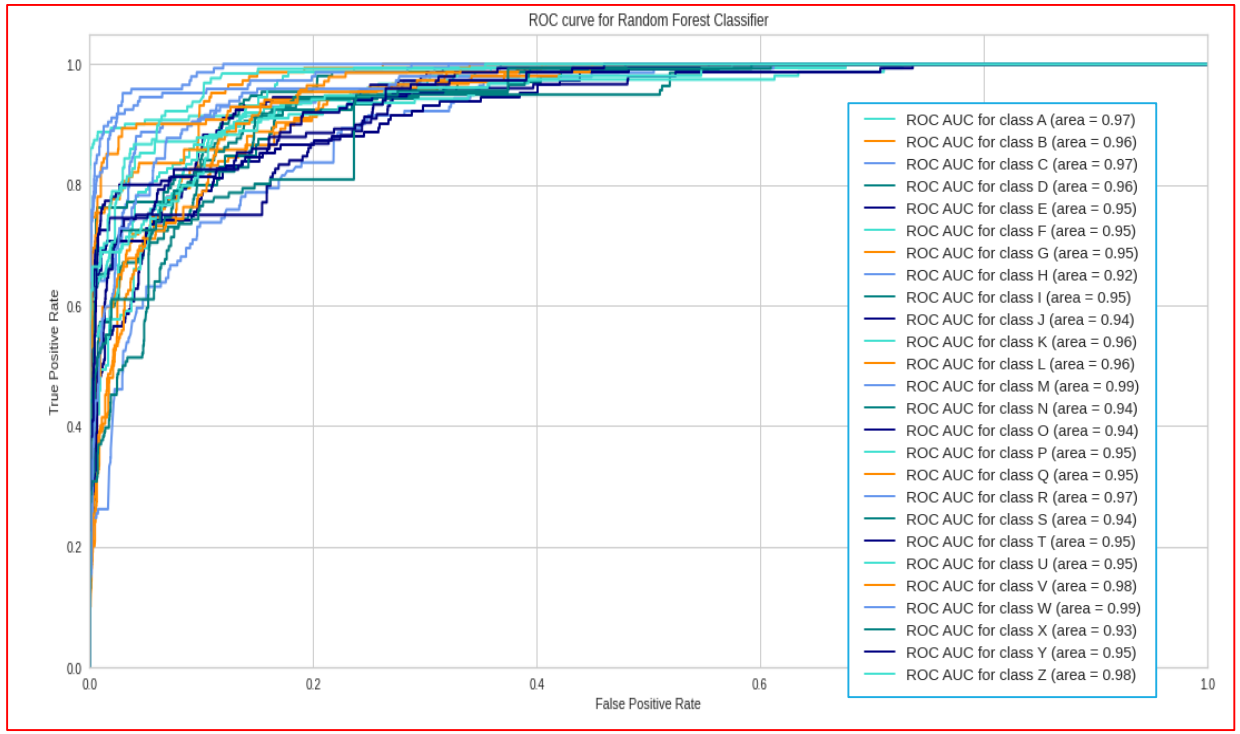
- **Support Vector Machine with kernel 'rbf' and 'Linear'**
- **Logistic regression**
- **Naïve Bayes**
- **Random Forest**

SVM(rbf) outperform other models



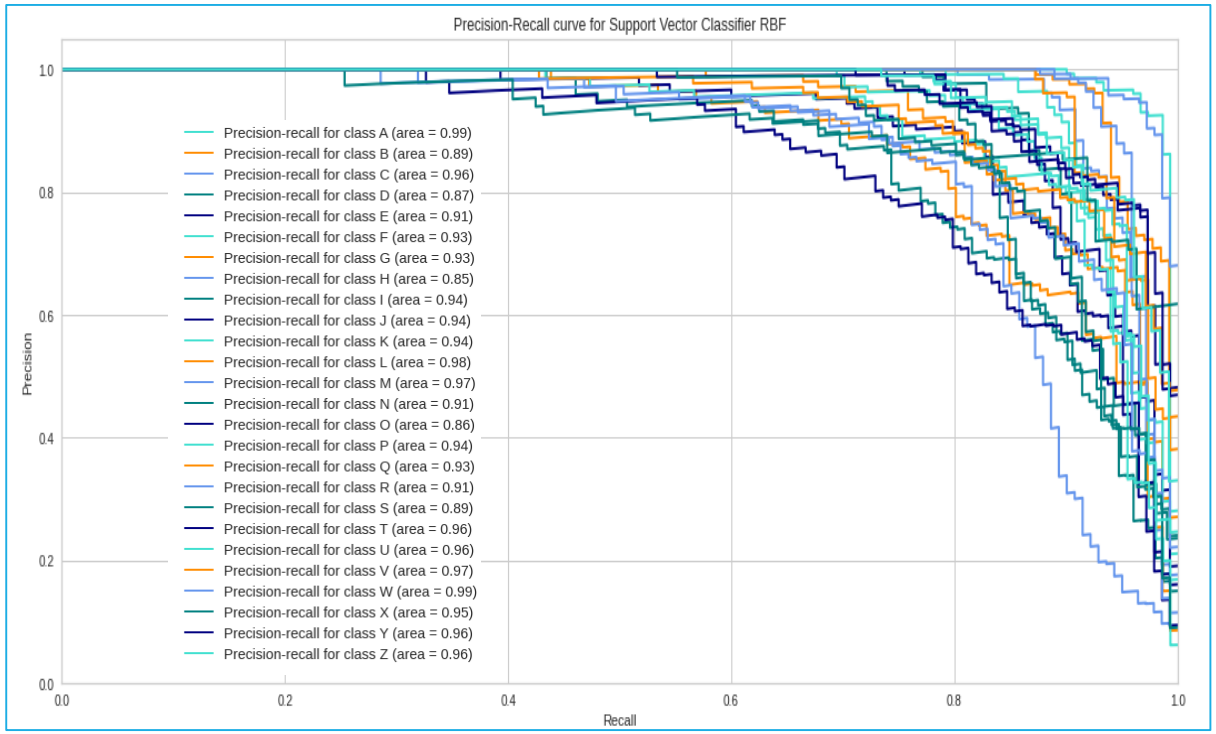
PERFORMANCE FOR BEST MODELS :ROCAUC

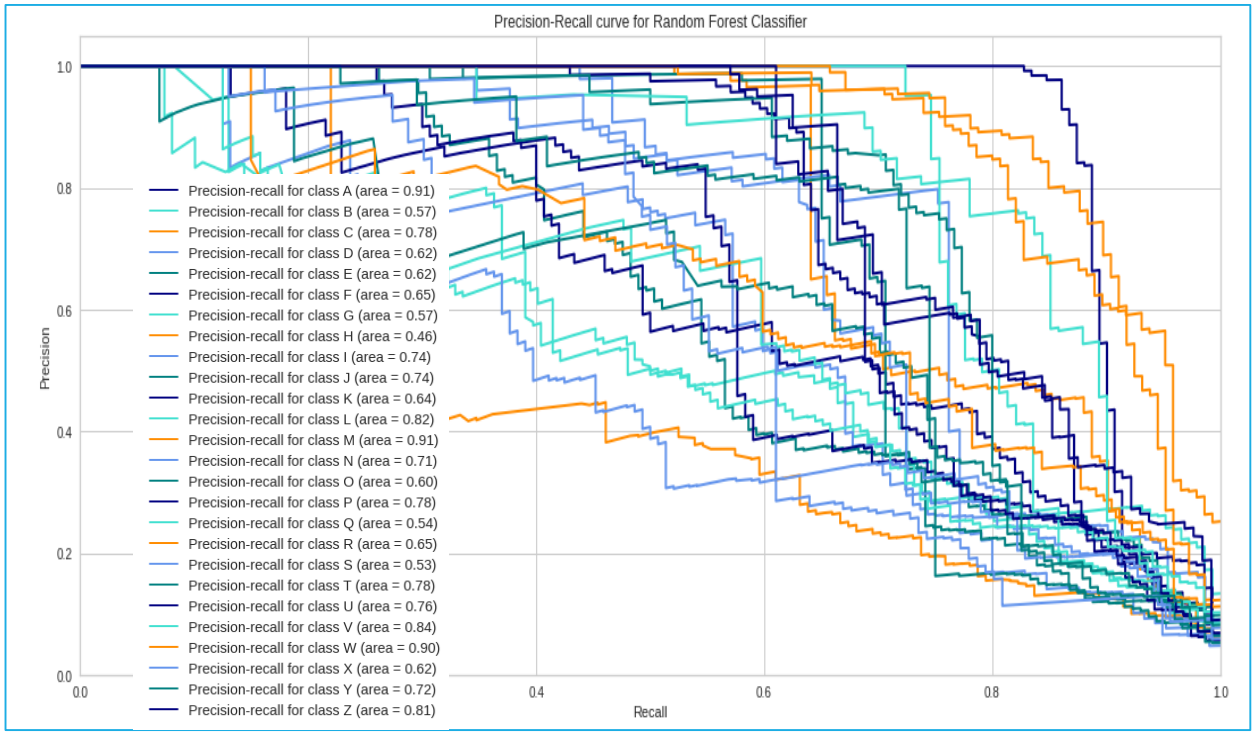




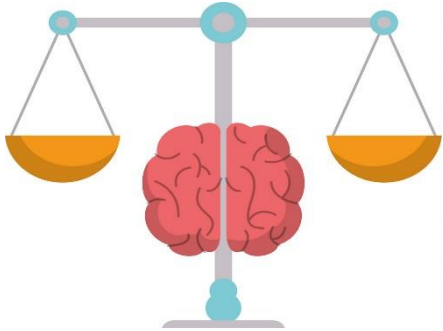
ROC AUC wise best two models are SVC (rbf) and Random Forest Classifier

PERFORMANCE FOR BEST MODELS :P/R AUC

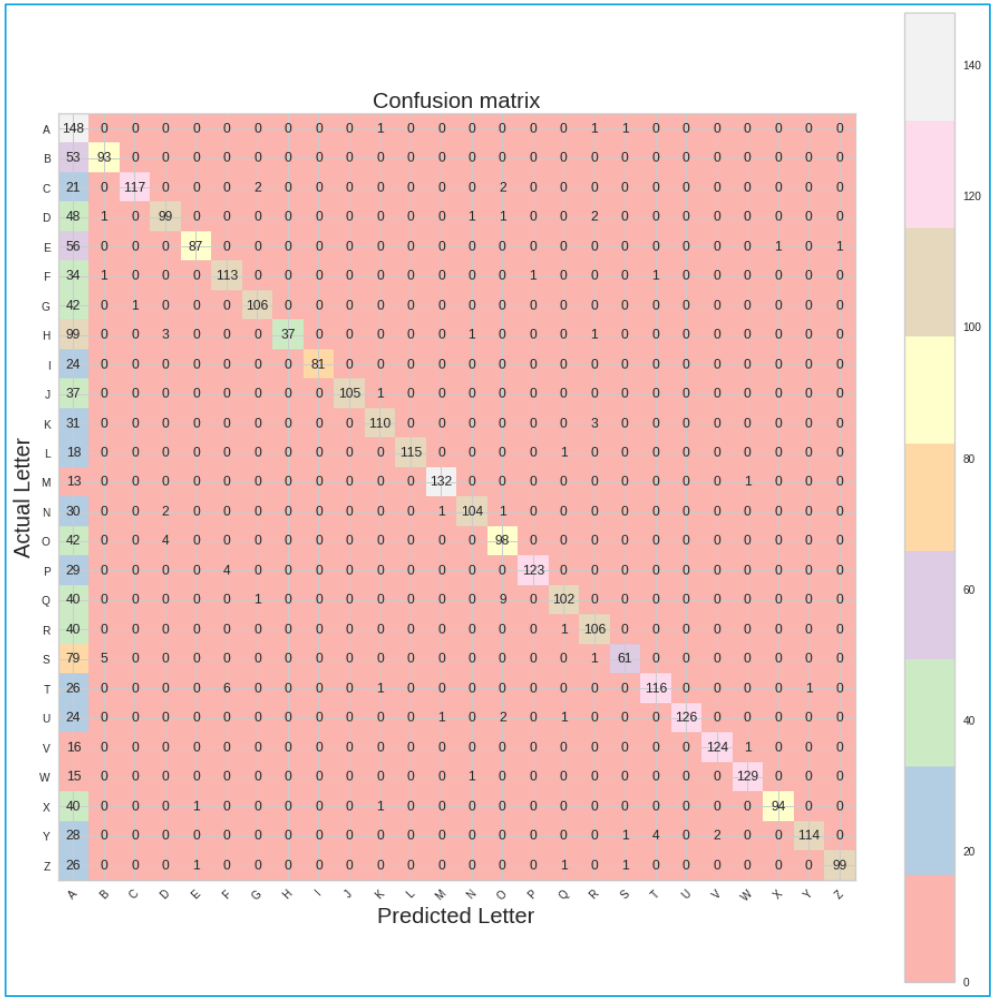




P/R AUC wise best two models are SVC (rbf) and Random Forest Classifier



CONFUSION MATRIX FOR BEST MODEL (SVC)

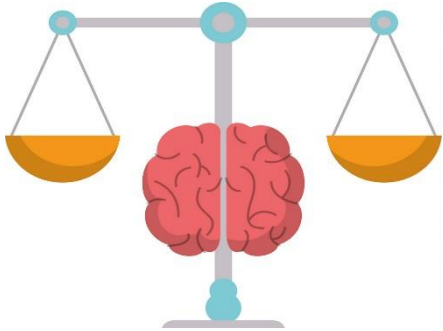


Sl No	Model	Parameters	Micro avg. Precision score
1	**Logistic Regression**	Penalty = 'l2' ; c =1.0 ; solver = 'lbfgs'	0.54
2	**Random Forest**	n_estimators=150, max_depth=3, random_state=0	0.68
3	**Support Vector Classifier (rbf)**	C=1.0, kernel='rbf', gamma='scale', shrinking=True, cache_size=200, class_weight=None, verbose=False, decision_function_shape='ovr',random_state=42	**0.94**
4	**Support Vector Classifier (linear)**	C=1.0, kernel='linear', gamma='scale', shrinking=True, cache_size=200, class_weight=None, verbose=False, decision_function_shape='ovr',random_state=42	0.47
5	**Naïve Bayes**	priors=None, var_smoothing=1e-09	0.58

CONCLUSION

k-fold CV, Final Prediction and Conclusion

CROSS VALIDATION, FINAL PREDICTION AND CONCLUSION



```
classifier
```

```
OneVsRestClassifier(estimator=SVC(C=1.0, break_ties=False, cache_size=200,
                                  class_weight=None, coef0=0.0,
                                  decision_function_shape='ovr', degree=3,
                                  gamma='scale', kernel='rbf', max_iter=-1,
                                  probability=False, random_state=42,
                                  shrinking=True, tol=0.001, verbose=False),
                    n_jobs=None)
```

```
from sklearn.model_selection import cross_validate
```

```
cv_results_svc = cross_validate(classifier, X_train,y_train, cv=5, scoring='recall_weighted',verbose = 2)
```

	ybar	x2bar	y2bar	xybar	x2ybar	xy2bar	xedge	xedgey	yedge	actual	predicted
1335	7.406165	7.415366	8.354896	6.204724	7.498532	7.057107	8.403819	8.466515	11.831519	J	A
1218	6.691736	10.571380	9.069243	6.204724	7.498532	7.057107	6.977460	7.414126	11.179085	B	B
13373	7.406165	8.046569	7.640549	6.876016	7.498532	7.843719	14.109253	8.466515	9.221785	M	M
17767	7.406165	9.308974	7.640549	8.889895	7.498532	9.416943	9.116998	7.414126	7.916918	O	O
10693	5.977306	8.677772	5.497508	6.876016	7.498532	7.843719	13.396074	9.518904	10.526652	O	A

- **Model for Multiclass classification has been built.**
- **Support Vector Classification (with RBF kernel) out-performed other models in this case study of Letter recognition from their parameters.**
- **"Recall" achieved maximum 0.73 and it is expected that this can further be improved with Neural network based models.**
- **This case study can also be referred and used for similar type multiclass classification problem.**

Thanks for reading!

Lets collaborate and happy to receive any feedback/suggestion/comment at..

pathak.chiranjit@gmail.com