Feature Selection and Classification in Materials Science datasets

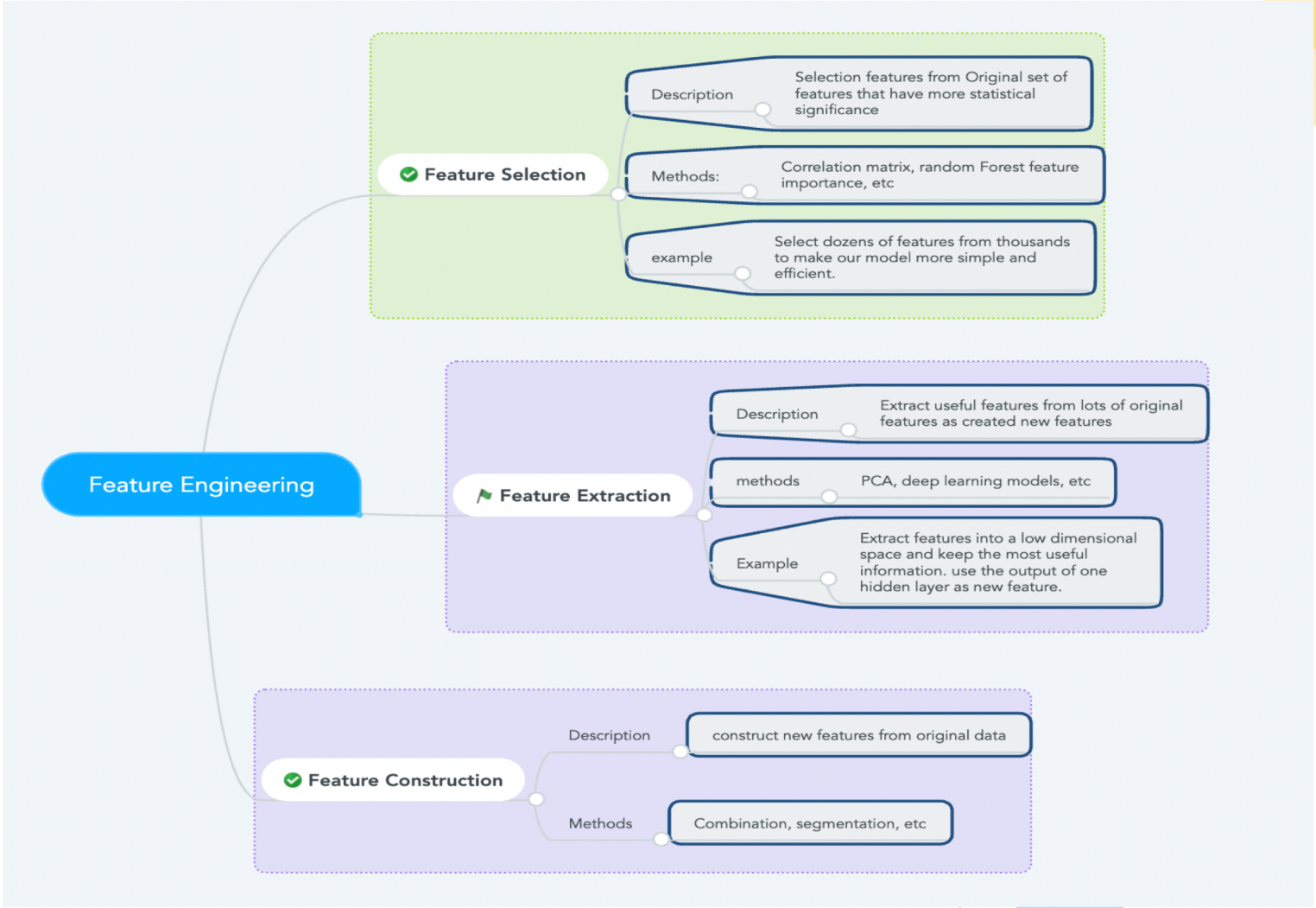
Formability Dataset

28/11/2021

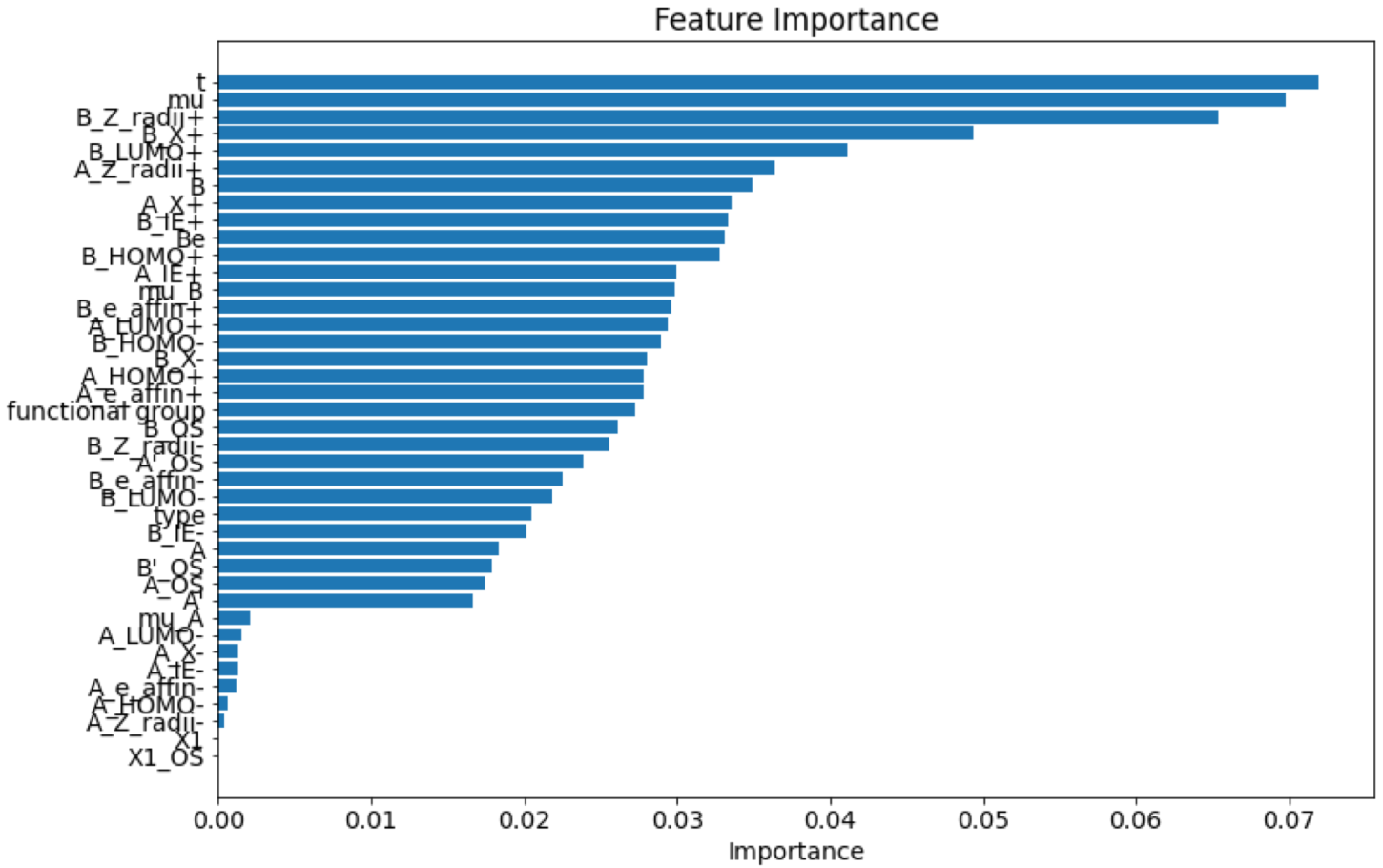
Masrat Rasool, Samir Belhouari, Fadwa El Mellouhi

Hamad Bin Khalifa University

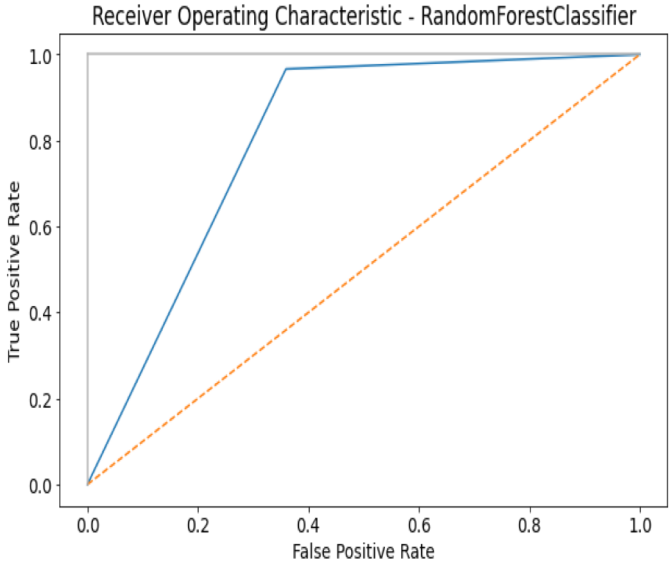
Ontology of the working

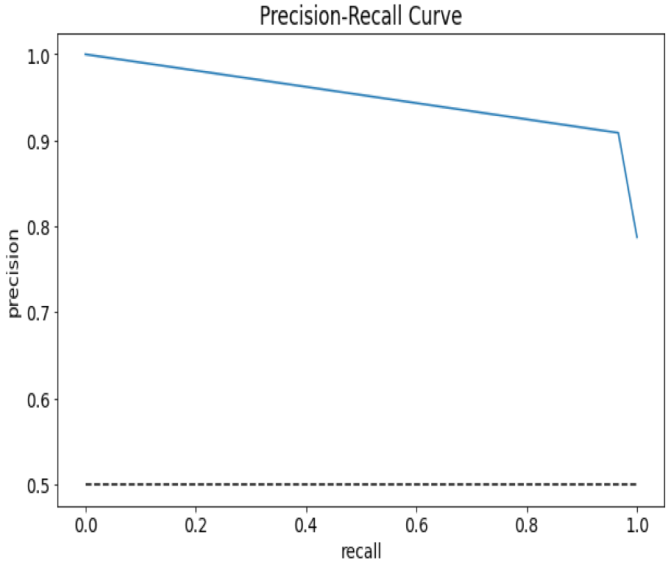


EXTRA TREES CLASSIFIER RESULTS FOR PERVOSKITE FORMABILITY
Feature importance plot



ROC and Precision-Recall Curves of the Cross-validated Extra-Trees classification of test data

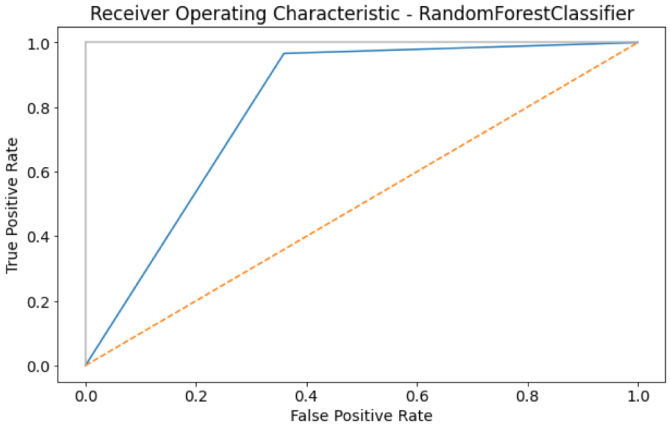


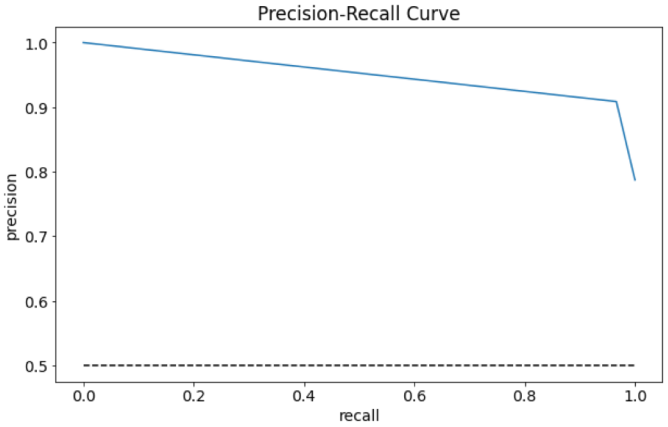


classification report

	precision	recall	f1-score	support
0	0.88	0.67	0.76	64
1	0.92	0.97	0.94	237
accuracy			0.91	301
macro avg	0.90	0.82	0.85	301
weighted avg	0.91	0.91	0.91	301

ROC and Precision-Recall Curves of the Cross-validated Random-Forest classification of test data

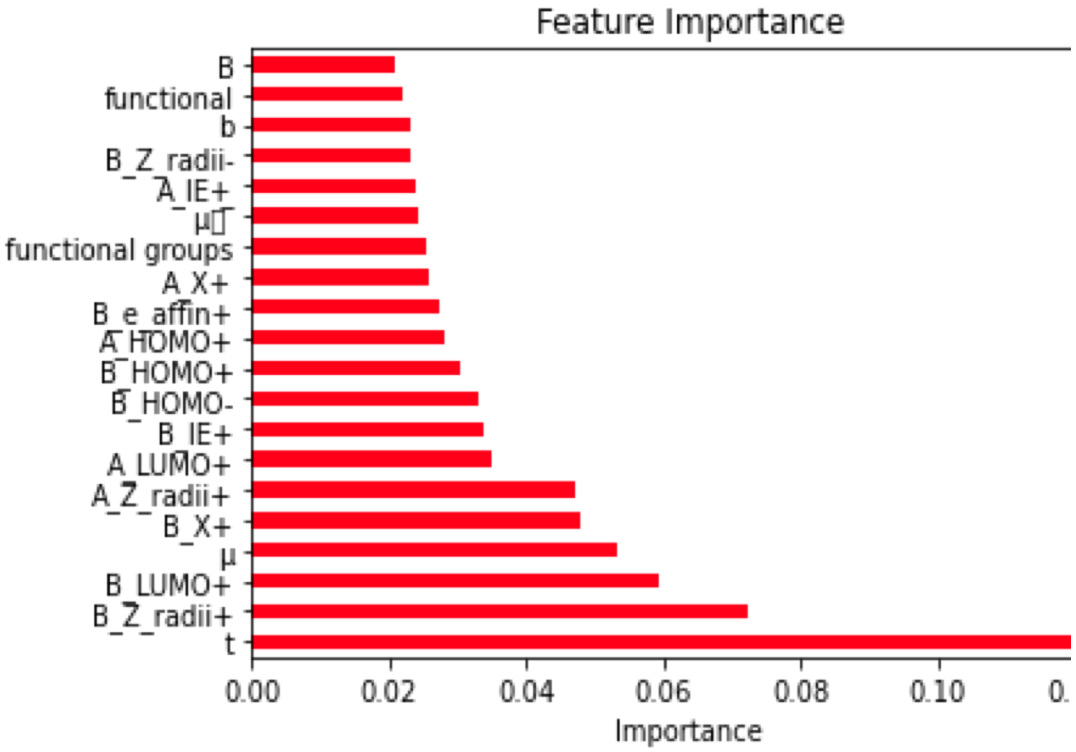




Classification Report

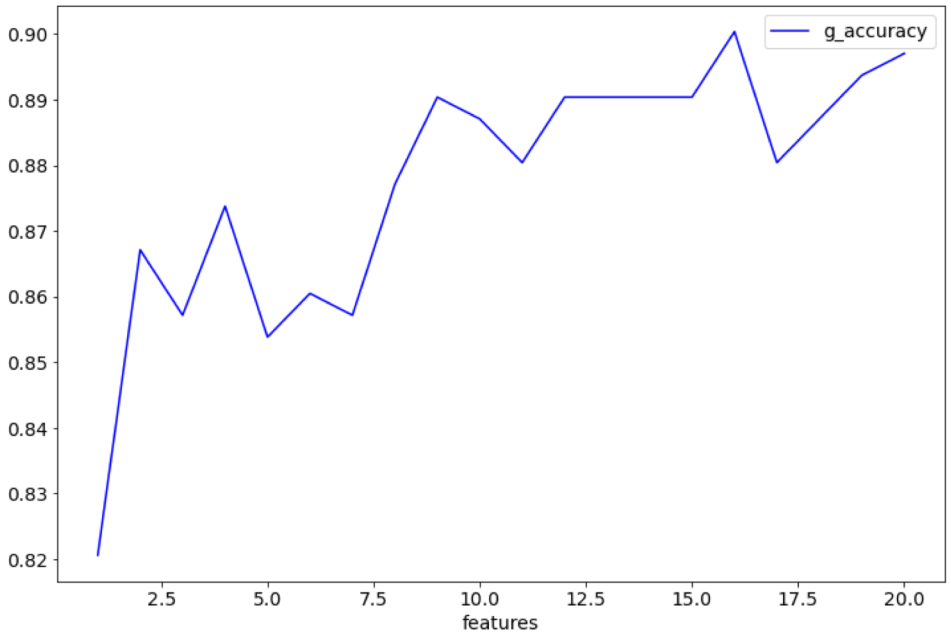
	precision	recall	f1-score	support
0	0.84	0.64	0.73	64
1	0.91	0.97	0.94	237
accuracy			0.90	301
macro avg	0.87	0.80	0.83	301
weighted avg	0.89	0.90	0.89	301

Feature Importance

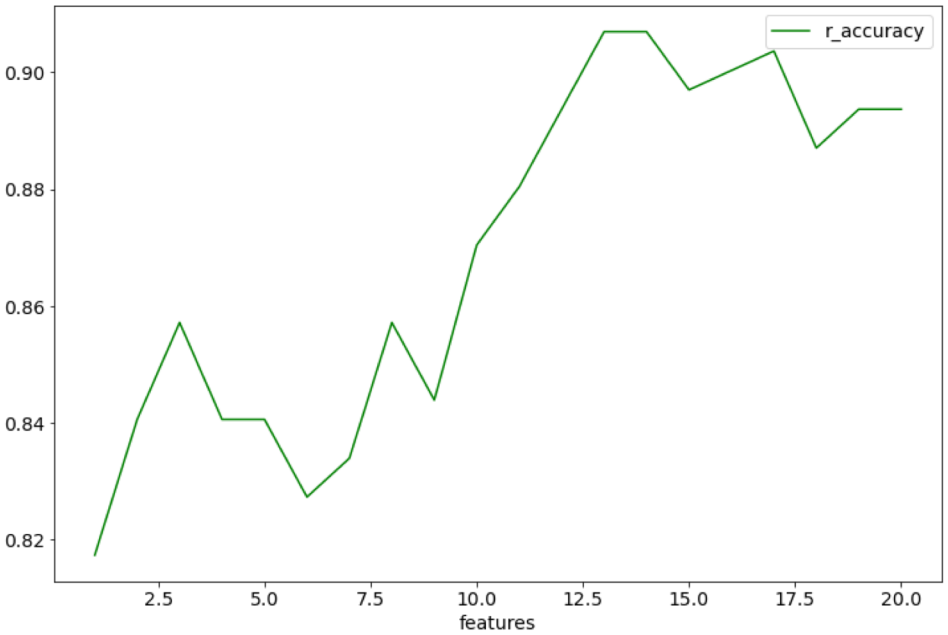


Number of features and accuracy for random forest classifier

GradientBoosting



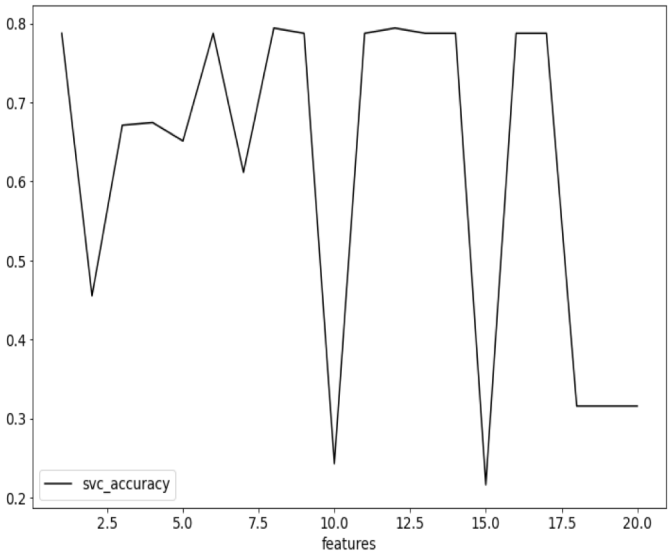
RandomForest



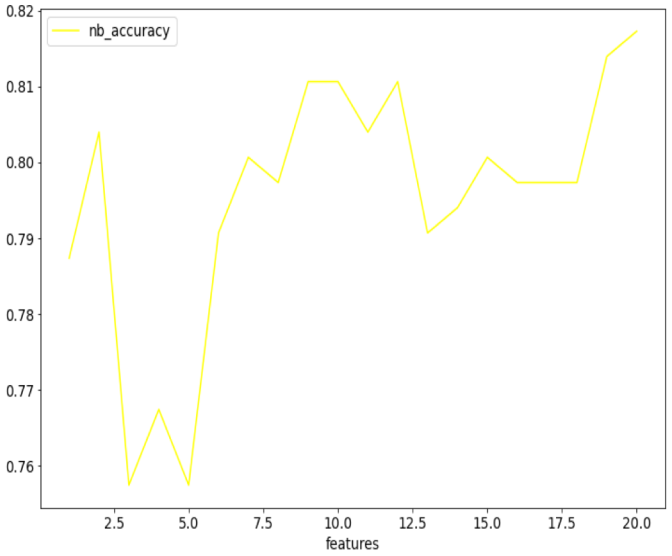
Number of features and accuracy for random forest classifier

Gradient Boosting

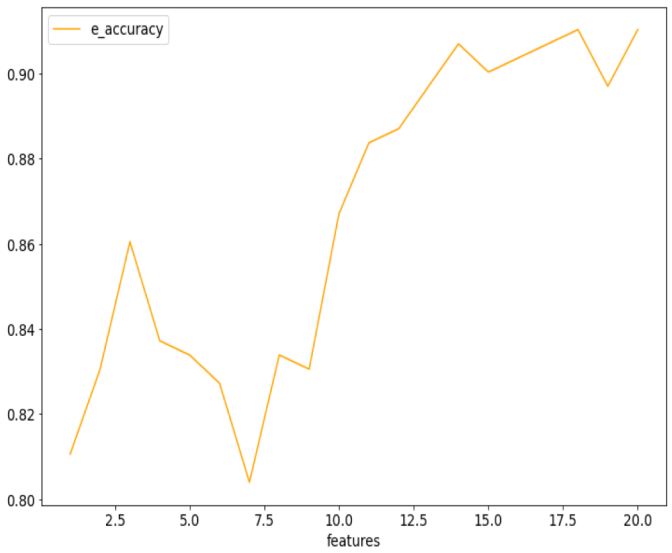
SVC



Gaussian NB

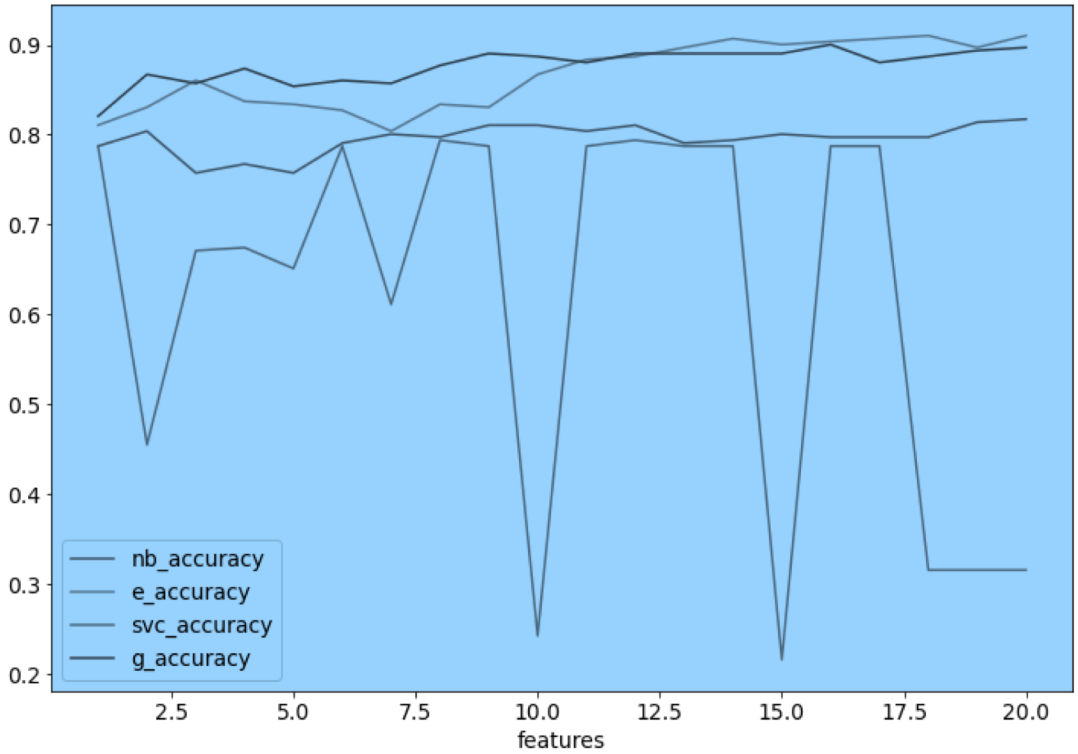


ExtraTrees

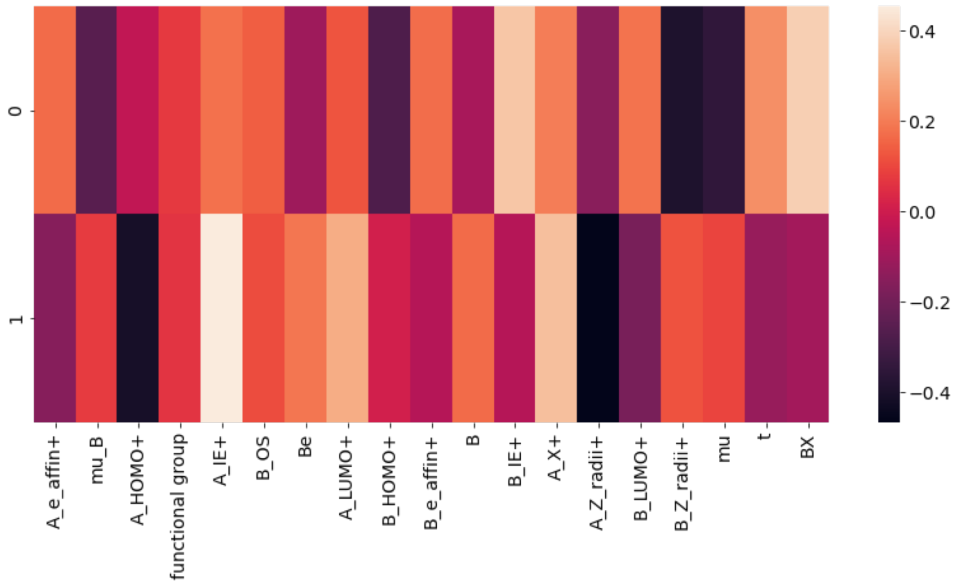


Accuracy with 20 features

Classification Algorithm	Best Accuracy
ExtraTrees	0.910
RandomForest	0.897
GradientBoosting	0.897
SVC	0.807
GaussianNB	0.817



New or Combined Features-19



Classification Algorithm	Best Accuracy
ExtraTrees	0.897
RandomForest	0.90
GradientBoosting	0.893
SVC	0.807
GaussianNB	0.817

Generating New Features

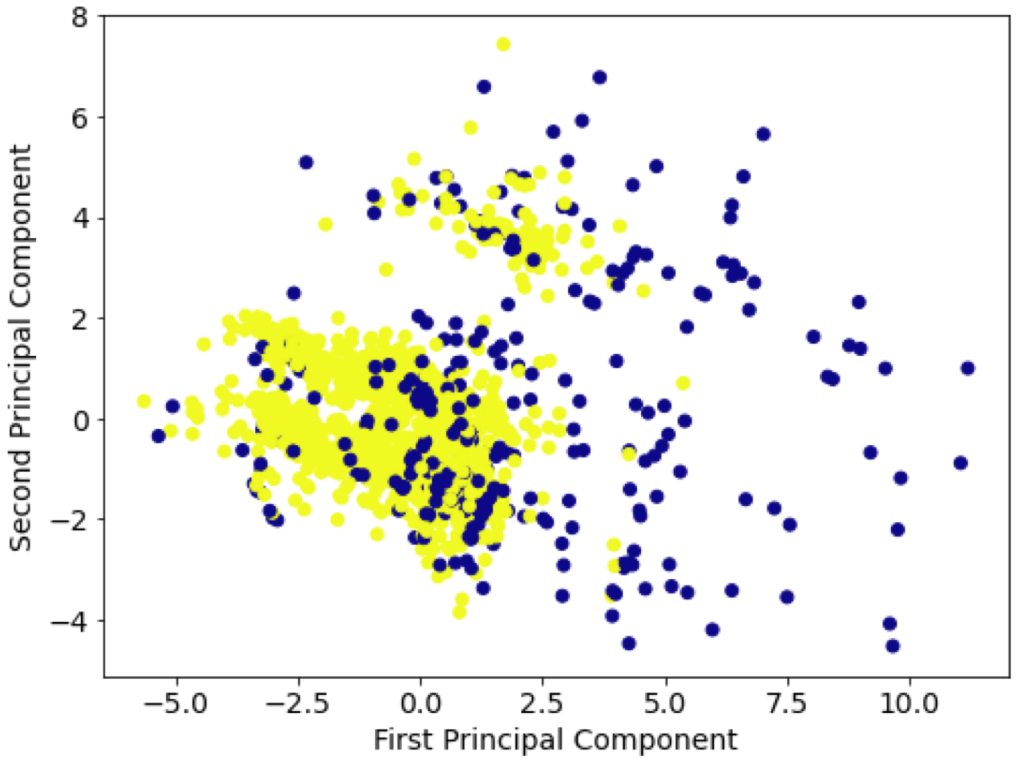
1. A_HUMO+ and B_HUMO+ → HUMO
2. A_LUMO+ and B_LUMO+. → LUMO
3. A_z_radii+ and B_z_Radii+. →Radii

- Reduced set is 16

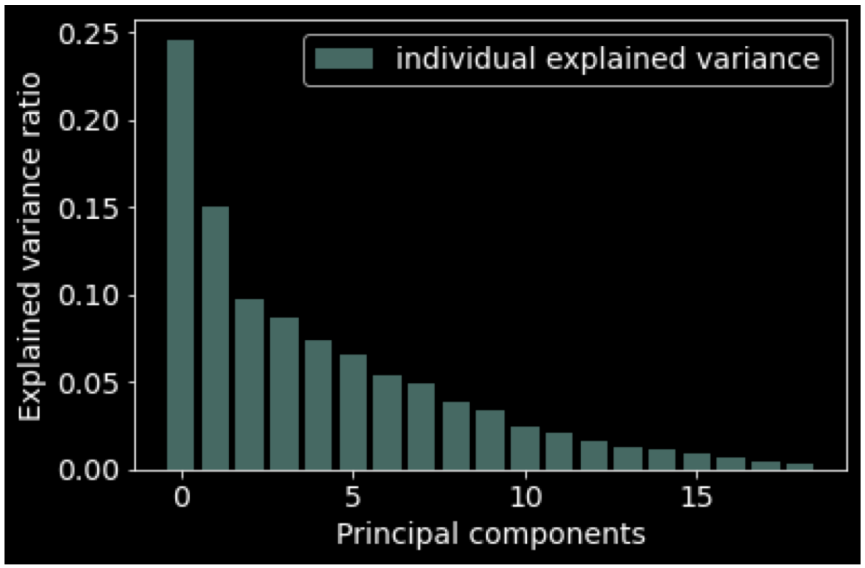
	A_e_affin+	mu_B	functional group	A_IE+	B_OS	Be	B_e_affin+	B	B_IE+	A_X+	mu	t	BX	HOMO	LUMO	radii
0	93.8	0.0000	1095	806.0	5	43	28.0	55	1522.6	1.64	0.4571	1.0815	3.00	-11.796	-8.560	13.780
1	-92.0	0.0804	194	1005.8	4	19	201.0	42	1401.0	1.78	0.5268	0.9957	2.26	-15.184	-7.964	13.912
2	-92.0	0.0786	196	1005.8	4	28	156.0	42	1362.4	1.78	0.5286	0.9946	2.26	-15.566	-8.225	13.934
3	-92.0	0.0196	215	1005.8	4	19	201.0	52	1421.2	1.78	0.4661	1.0370	2.34	-14.926	-7.705	13.572
4	-92.0	0.0482	232	1005.8	4	19	201.0	56	1442.4	1.78	0.4946	1.0172	2.20	-14.612	-7.384	13.322

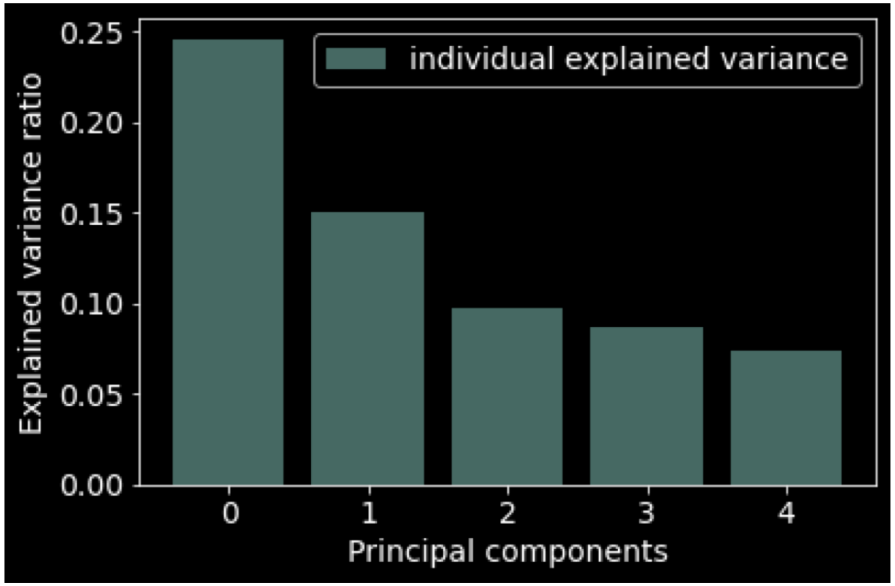
Classification Algorithm	Best Accuracy
ExtraTrees	0.9069
RandomForest	0.9069
GradientBoosting	0.8903
SVC	0.8107
GaussianNB	0.8107

Applying PCA-1

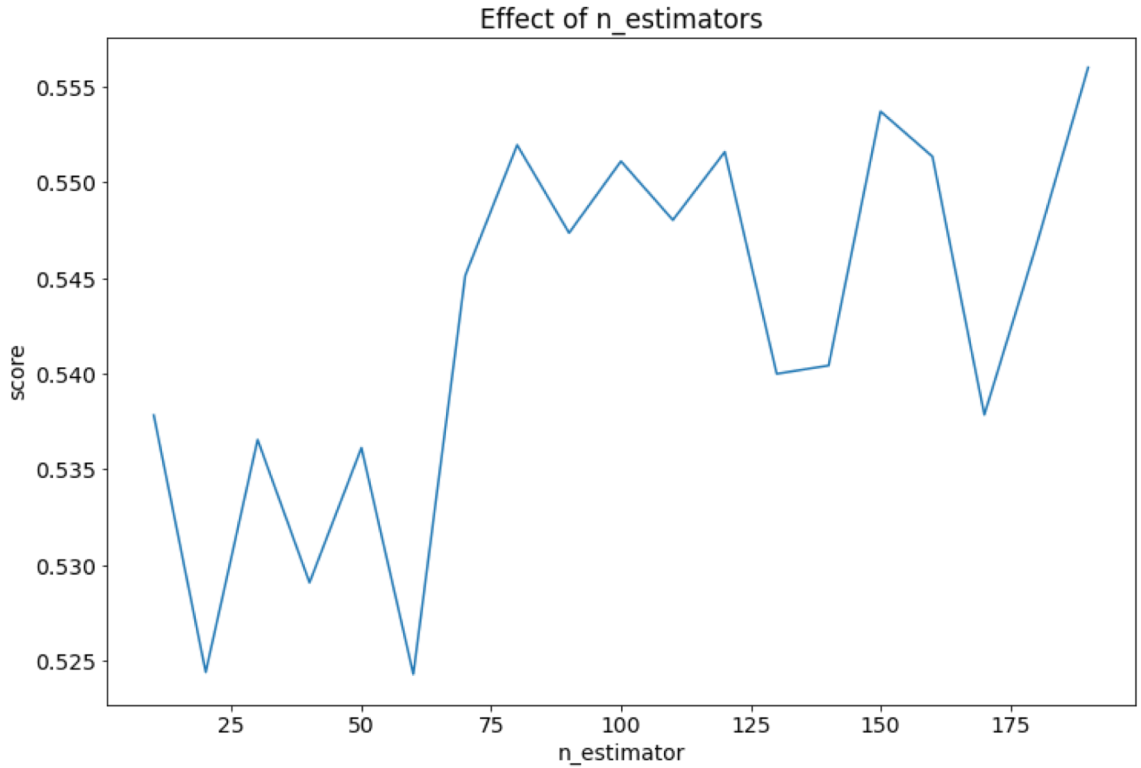


PCA-2

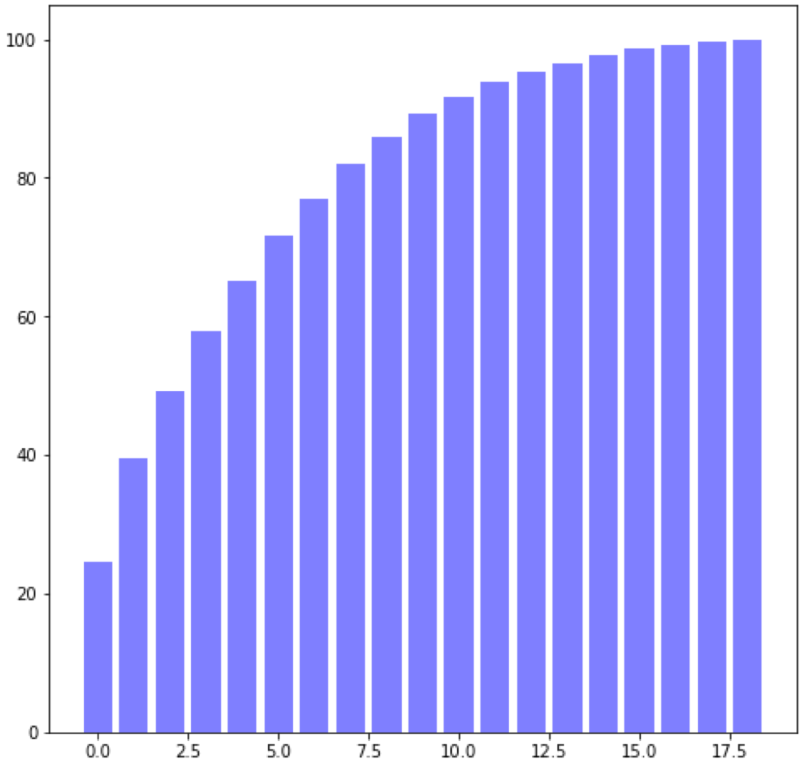

individual explained variance
Explained variance ratio
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
Principal components

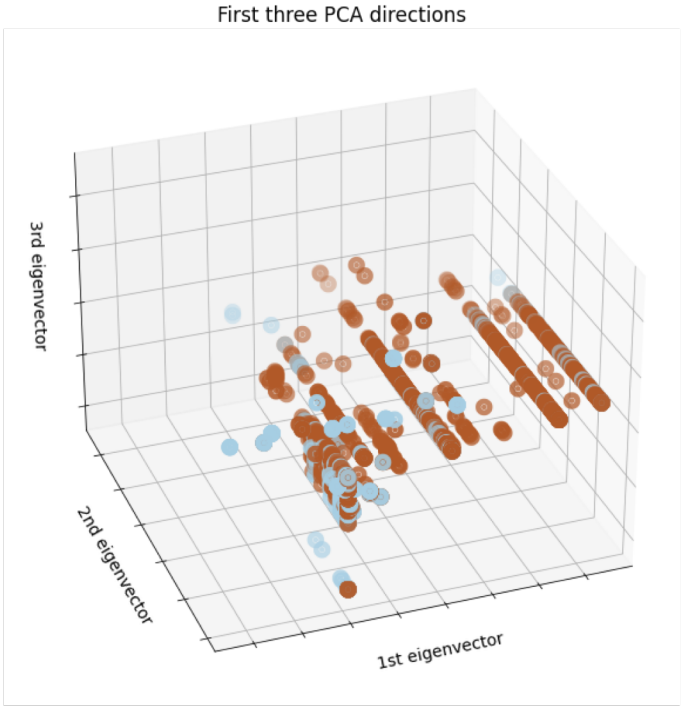

individual explained variance
Explained variance ratio
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
Principal components

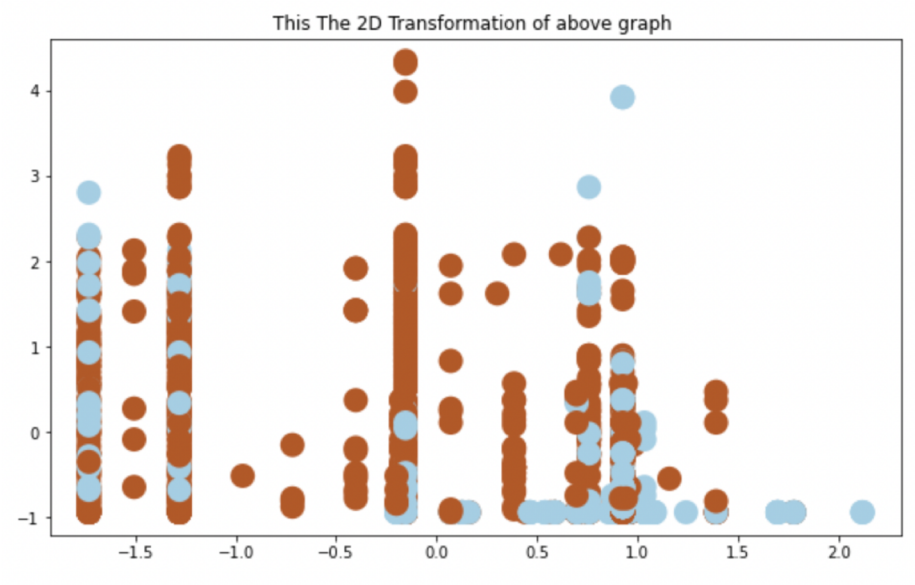
Establishing Model-randomForest Regressor



PCA-3




First three PCA directions
3rd eigenvector
2nd eigenvector
1st eigenvector


This The 2D Transformation of above graph
4
3
2
1
0
−1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

Classification

New Feature set=16

Before PCA	
Classification	**Best Accuracy**
ExtraTrees	0.9069
RandomForest	0.9069
GradientBoosting	0.8903
SVC	0.8107
GaussianNB	0.8107

After PCA	
classifier	**Accuracy**
Random Forest	0.92026
Extra Trees	0.91694
Gradient Boosting	0.91362
Support Vector	0.900332
Naïve Bayes	0.900332

classifier	Accuracy
Random Forest	0.92026
Extra Trees	0.91694
Gradient Boosting	0.91362
Support Vector	0.900332
Naïve Bayes	0.900332

Shallow neural Network

- Accuracy on test set=0.7639

MODnet

```
md = MODData(mate = df['functional group'],
             targets = df['Perovskite'].values,
             structure_ids = df.index,
             target_names = ['Perovskite_index']
            )
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-12-394af5e232af> in <module>()
      2              targets = df['Perovskite'].values,
      3              structure_ids = df.index,
----> 4              target_names = ['Perovskite_index']
      5             )

TypeError: __init__() got an unexpected keyword argument 'mate'
```

SEARCH STACK OVERFLOW

Double-click (or enter) to edit

```
[ ] data = MODData(materials = df['functional group'],
                   targets = df['Perovskite'],
                  )
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-13-8adfe4f42432> in <module>()
      1 data = MODData(materials = df['functional group'],
----> 2                targets = df['Perovskite'],
      3               )
```

```
[ ]              structure_ids = df.index,
                 target_names = ['Perovskite_index']
                )
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-12-394af5e232af> in <module>()
      2              targets = df['Perovskite'].values,
      3              structure_ids = df.index,
----> 4              target_names = ['Perovskite_index']
      5             )

TypeError: __init__() got an unexpected keyword argument 'mate'
```

SEARCH STACK OVERFLOW

Double-click (or enter) to edit

```
[ ] data = MODData(materials = df['functional group'],
                   targets = df['Perovskite'],
                  )
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
<ipython-input-13-8adfe4f42432> in <module>()
      1 data = MODData(materials = df['functional group'],
----> 2                targets = df['Perovskite'],
      3               )

TypeError: __init__() got an unexpected keyword argument 'materials'
```

Stability dataset

Data Preprocessing

- Removed Null values
- Conversion of variables from categorical to numerical

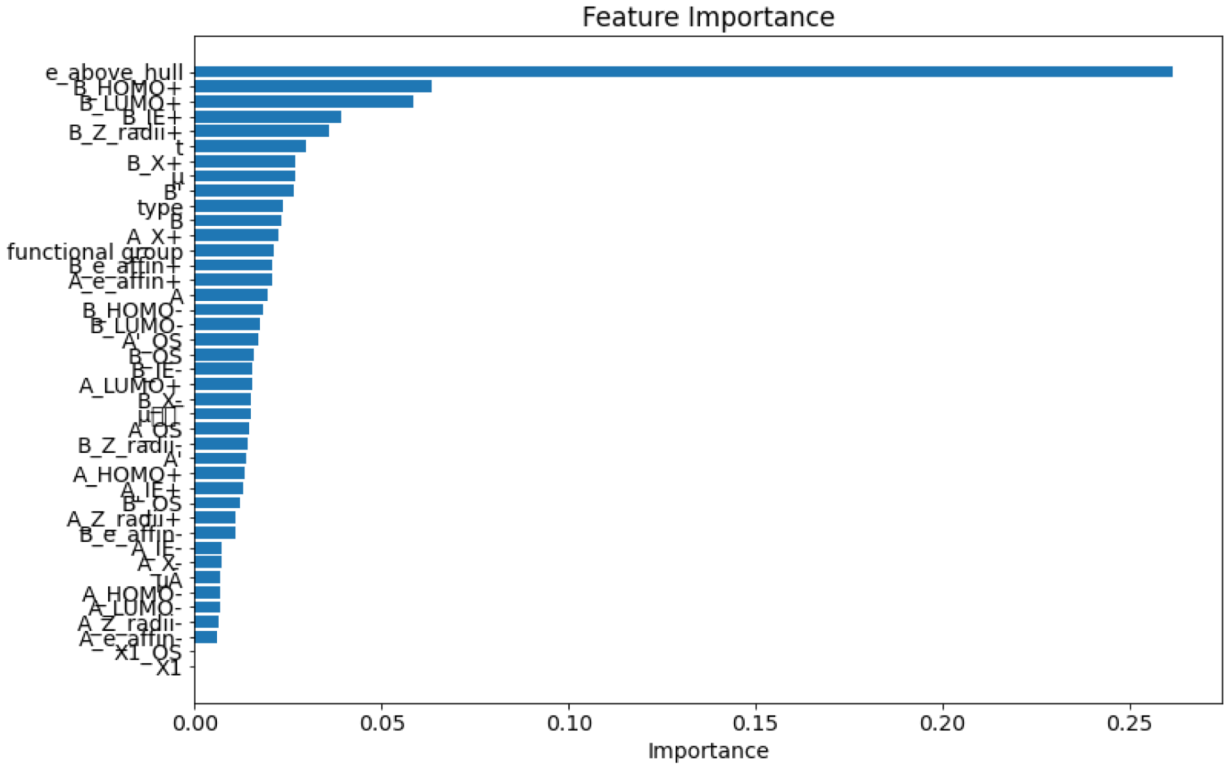
Dataset

	functional group	A	A'	B	B'	X1	type
0	BaSrSn2O6	Ba	Sr	Sn	Sn	O	AA'B2O6
1	BaSrZr2O6	Ba	Sr	Zr	Zr	O	AA'B2O6
2	Ca2ZrMnO6	Ca	Ca	Zr	Mn	O	A2BB'O6
3	Sr2HfNiO6	Sr	Sr	Hf	Ni	O	A2BB'O6
4	BaRbTaZrO6	Ba	Rb	Ta	Zr	O	AA'BB'O6
...	...	...	...	...	...	...	...
3464	Sr2YSbO6	Sr	Sr	Y	Sb	O	A2BB'O6
3465	Sr2YTaO6	Sr	Sr	Y	Ta	O	A2BB'O6
3466	Sr2ZnMoO6	Sr	Sr	Zn	Mo	O	A2BB'O6
3467	Sr2ZrMnO6	Sr	Sr	Zr	Mn	O	A2BB'O6
3468	Sr2ZrMoO6	Sr	Sr	Zr	Mo	O	A2BB'O6

	functional group	A	A_OS	A'	A'_OS	A_HOMO-	A_HOMO+	A_IE-	A_IE+	A_LUMO-	...	B_e_affin-	B_e_affin+	X1	X1_OS	e_above_hull	μ	μĀ	μB−	type	
0	502	5	2	43	2	0.293	-6.763	-46.6	1052.4	-0.739	...	0.0	232.0	0	-2	0.0288	0.4929	0.0607	0.0000	1	0
1	505	5	2	43	2	0.293	-6.763	-46.6	1052.4	-0.739	...	0.0	82.2	0	-2	0.0306	0.5143	0.0607	0.0000	1	0
2	648	8	2	6	2	0.000	-7.536	0.0	1179.6	0.000	...	91.1	-8.9	0	-2	0.0000	0.4464	0.0000	0.0679	0	0
3	2870	52	2	43	2	0.000	-7.056	0.0	1099.0	0.000	...	-139.0	173.0	0	-2	0.0000	0.4250	0.0000	0.0821	0	1
4	483	5	2	38	1	-0.989	-5.481	99.9	905.9	-1.292	...	-27.1	55.1	0	-2	0.0000	0.4857	0.0393	0.0286	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3464	2979	52	2	43	2	0.000	-7.056	0.0	1099.0	0.000	...	-71.4	130.6	0	-2	0.0519	0.5357	0.0000	0.1071	0	0
3465	2980	52	2	43	2	0.000	-7.056	0.0	1099.0	0.000	...	15.6	43.6	0	-2	0.0671	0.5500	0.0000	0.0929	0	0
3466	2989	52	2	43	2	0.000	-7.056	0.0	1099.0	0.000	...	-63.0	81.0	0	-2	0.0000	0.4750	0.0000	0.0536	0	0
3467	2994	52	2	43	2	0.000	-7.056	0.0	1099.0	0.000	...	91.1	-8.9	0	-2	0.0000	0.4464	0.0000	0.0679	0	0
3468	2995	52	2	43	2	0.000	-7.056	0.0	1099.0	0.000	...	-30.9	113.1	0	-2	0.0000	0.4893	0.0000	0.0250	0	0

3469 rows x 41 columns

Feature Importance with ExtraTrees Classifier

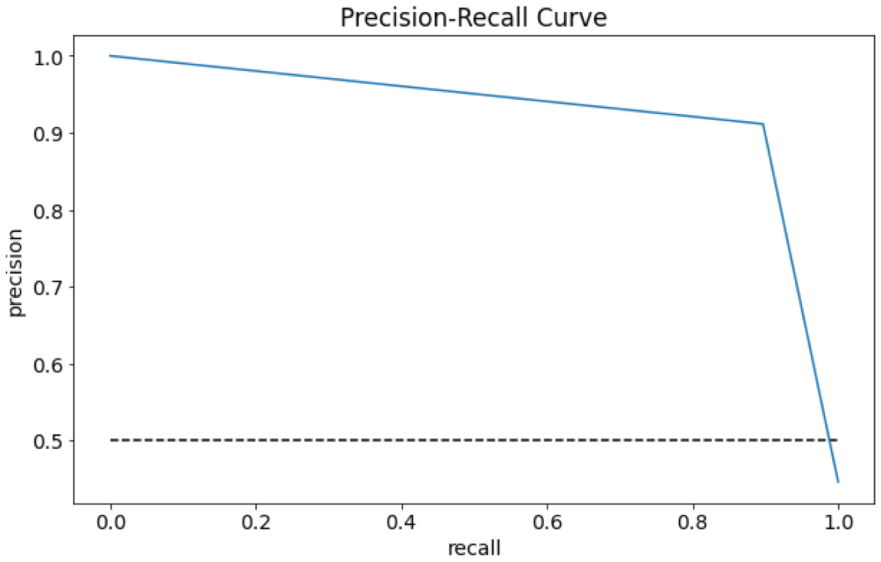


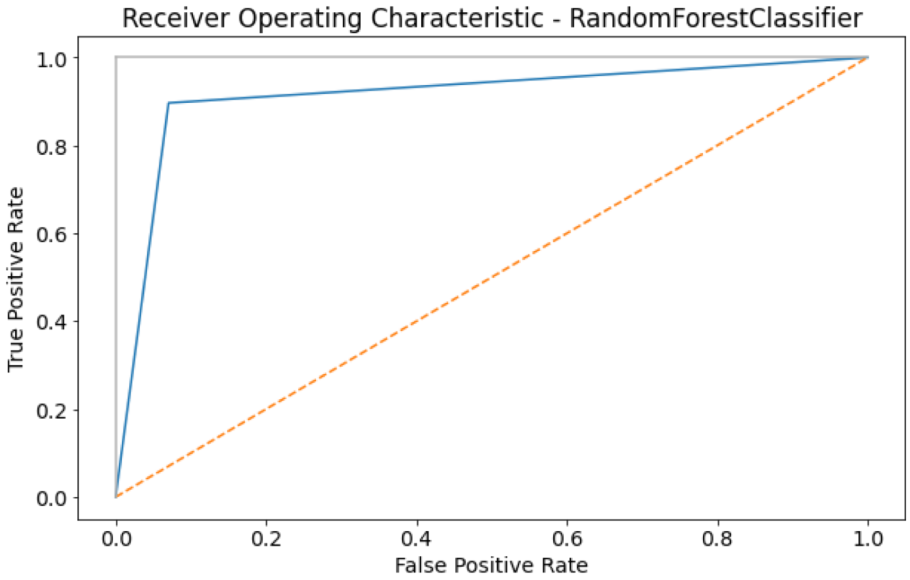
Wednesday, November 24, 2021

Random Forest Classifier-

Confusion matrix

```
array([[357,  27],
       [ 32, 278]], dtype=int64)
```

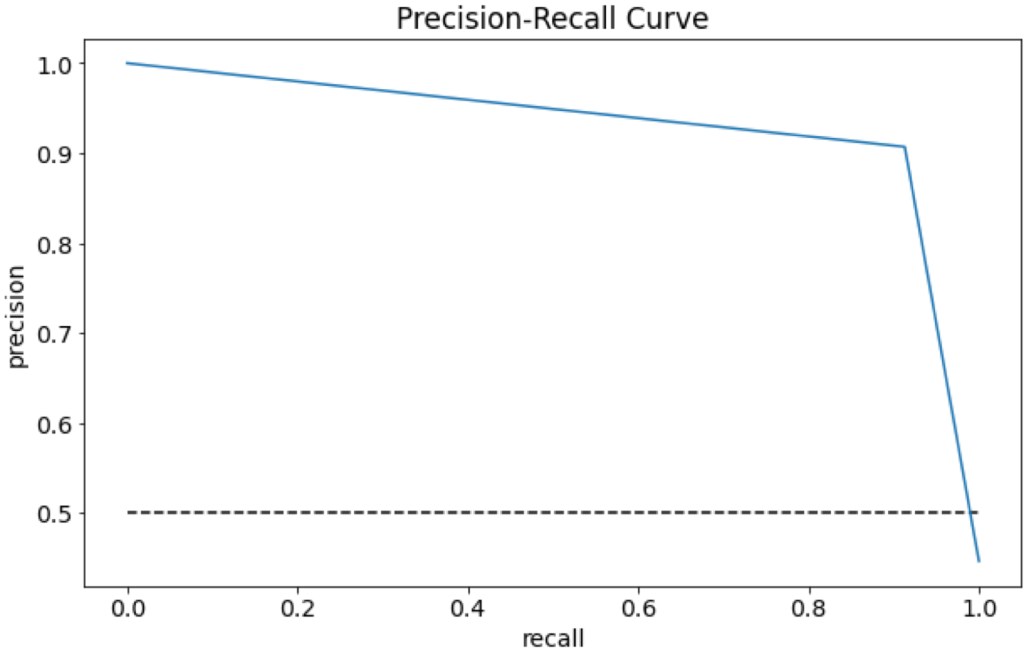


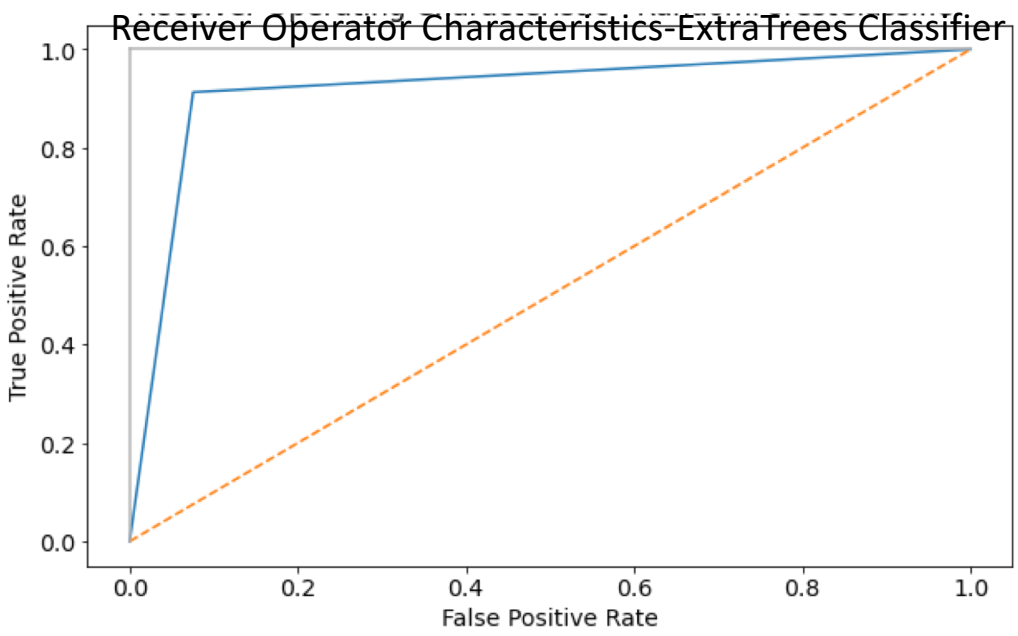


Classificaton Report

	precision	recall	f1-score	support
0	0.92	0.93	0.92	384
1	0.91	0.90	0.90	310
accuracy			0.91	694
macro avg	0.91	0.91	0.91	694
weighted avg	0.91	0.91	0.91	694

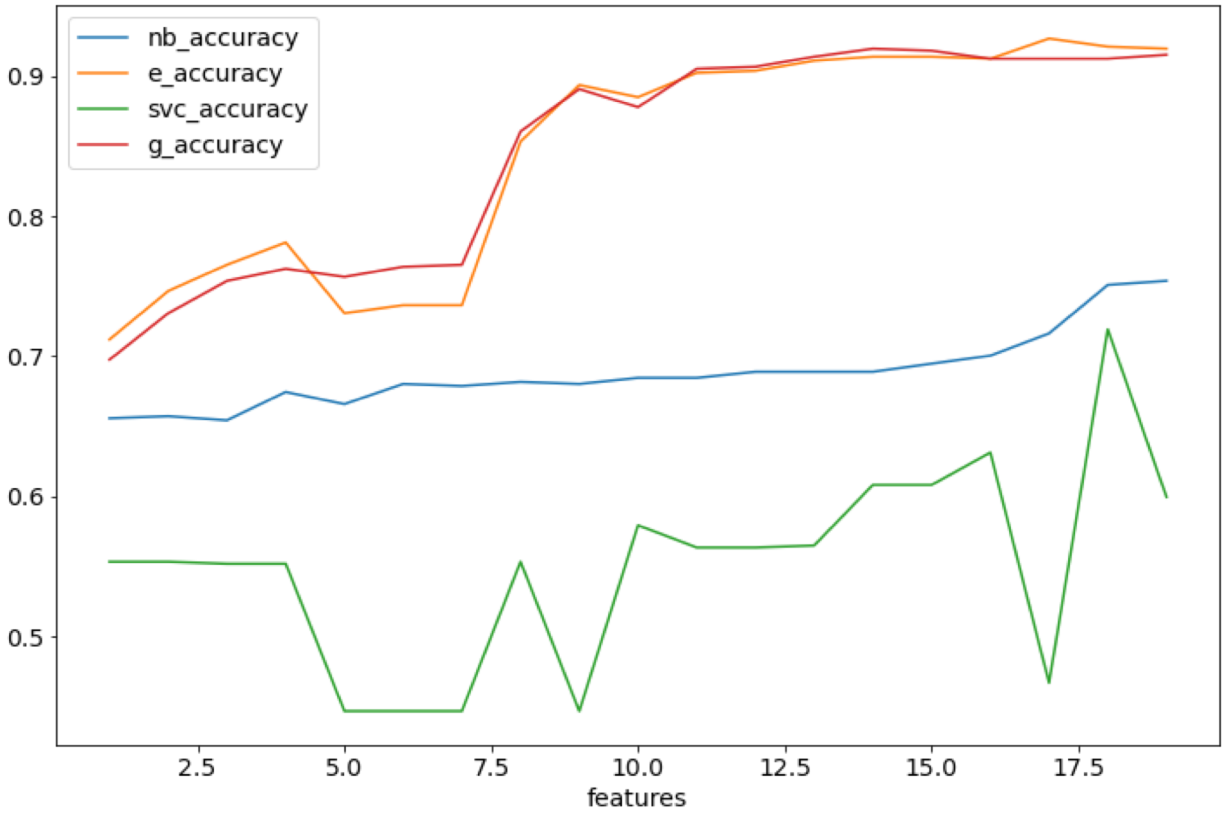
Extra Trees Classifier





Accuracy comparison of five different Classifiers

Classifier	Maximm Accuracy	Number of features
Random Forest	0.9265	18
Extra Trees	0.9265	16
Gradient Boosting	0.9193	14
Support Vector	0.7190	18
Gaussian Naïve Bayes	0.7536	19



Generating New Features

	functional group	A	A_OS	A'	A'_OS	A_HOMO-	A_HOMO+	A_IE-	A_IE+	A_LUMO-	...	B_e_affin+	X1	X1_OS	e_above_hull	μ	μĀ	μB−	type	t	stable
0	BaSrSn2O6	Ba	2	Sr	2	0.293	-6.763	-46.6	1052.4	-0.739	...	232.0	O	-2	0.0288	0.4929	0.0607	0.0000	AA'B2O6	0.9896	1
1	BaSrZr2O6	Ba	2	Sr	2	0.293	-6.763	-46.6	1052.4	-0.739	...	82.2	O	-2	0.0306	0.5143	0.0607	0.0000	AA'B2O6	0.9756	1
2	Ca2ZrMnO6	Ca	2	Ca	2	0.000	-7.536	0.0	1179.6	0.000	...	-8.9	O	-2	0.0000	0.4464	0.0000	0.0679	A2BB'O6	0.9568	1
3	Sr2HfNiO6	Sr	2	Sr	2	0.000	-7.056	0.0	1099.0	0.000	...	173.0	O	-2	0.0000	0.4250	0.0000	0.0821	A2BB'O6	1.0066	1
4	BaRbTaZrO6	Ba	2	Rb	1	-0.989	-5.481	99.9	905.9	-1.292	...	55.1	O	-2	0.0000	0.4857	0.0393	0.0286	AA'BB'O6	1.0420	1

5 rows × 42 columns

	B_HOMO-	A_IE+	B_OS	B_LUMO-	functional group	A'_OS	A_X+	B_e_affin+	B'	A_e_affin+	...	t	B_Z_radii+	B_IE+	B_LUMO+	B_HOMO+	HOMO	AFFIN	LUMO	OS	
0	0.000	1052.4	4	0.000	502	2	1.84	232.0	53	-192.0	...	0.9896	3.760	1417.2	-1.080	-7.386	-7.386	40.0	-1.080	6	246
1	0.000	1052.4	4	0.000	505	2	1.84	82.2	67	-192.0	...	0.9756	5.650	1320.0	-7.016	-7.336	-7.336	-109.8	-7.016	6	237
2	0.776	1179.6	4	-2.447	648	2	2.00	-8.9	30	-372.0	...	0.9568	5.045	1377.4	-4.569	-8.112	-7.336	-380.9	-7.016	6	255
3	1.680	1099.0	4	-2.027	2870	2	1.90	173.0	36	-292.0	...	1.0066	5.090	1412.1	-3.521	-7.236	-5.556	-119.0	-5.548	6	25
4	0.016	905.9	5	-0.038	483	1	1.71	55.1	67	0.9	...	1.0420	5.615	1421.3	-7.054	-7.320	-7.304	56.0	-7.092	6	232

5 rows × 24 columns

Reduced Data SET

New Features

- HOMO = HOMO- + B_HOMO+
- AFFIN =B_e_affin+ + A_e_affin+
- LUMO = B_LUMO+ + B_LUMO-
- OS = B_OS + A'_OS
- IE= A_IE+ + B_IE+