Beans

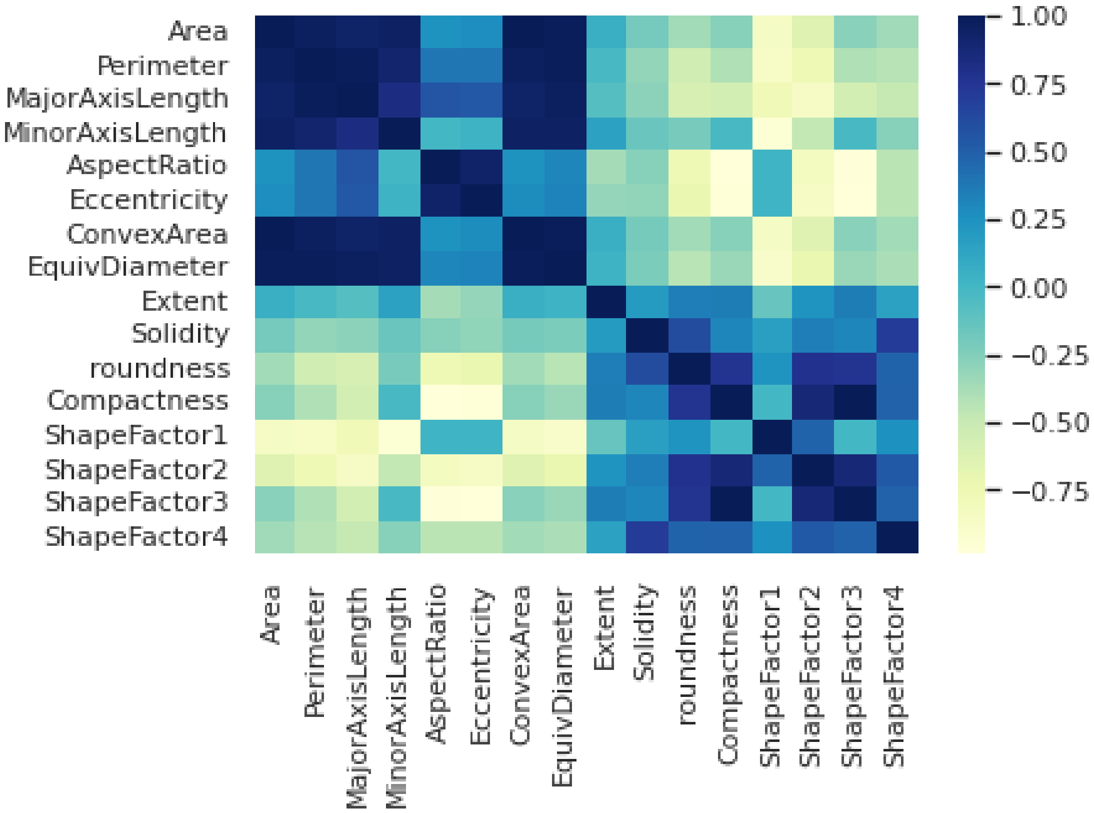
December 16, 2022

```
In [65]: import pandas as pd
         from sklearn.model_selection import train_test_split
         import seaborn as sns
         from sklearn.ensemble import RandomForestClassifier
         import matplotlib.pyplot as plt

         df = pd.read_csv('/public/bmort/python/beans.csv')
         print(df.isnull().sum())
         print("There is missing data in the ShapeFactor3 column so lets fix this")

         imputed_value = df['ShapeFactor3'].median()
         df['ShapeFactor3'].fillna(imputed_value)
         df['ShapeFactor3'] = df['ShapeFactor3'].fillna(imputed_value)

         print("")

         print(df.describe())

         print("The maginitudes between the columns are much different, some are very small and
         print("The range between the columns varies a fair amount as well with some exceptions
         print("The range of area is large, the range of the convex area is also very large")
         print("It is clear to see there are outliers here because there is the max is so far o

         sns.heatmap(df.corr(), cmap="YlGnBu")
         plt.show()
         print(df.corr())

         print("")

         train_x = df[['Area','Perimeter','AspectRatio','Eccentricity','roundness','Compactness
         train_y = df['Class'].values
         print("the decision was taken to drop some features with correlations to area over .9
         print("the benefits of this decision should be a significant reduction of the computat
         print("I chose to use major and minor axis length")

         le = preprocessing.LabelEncoder()
         le.fit(df['Class'])
         le.transform(df['Class'])
```

```
df['le_class'] = le.transform(df['Class'])

train_x = df[['MajorAxisLength','MinorAxisLength','AspectRatio','Extent','Solidity','
train_y = df['le_class'].values
X_train, X_test, y_train, y_test = train_test_split(train_x, train_y, test_size=0.2)


rf = RandomForestClassifier(n_estimators = 50)
rf.fit(X_train, y_train);

y_pred = rf.predict(X_test)
print(rf.score(X_train, y_train))
print(rf.score(X_test, y_test))

from sklearn.metrics import confusion_matrix
confusion_matrix(y_test, y_pred)

from sklearn.metrics import ConfusionMatrixDisplay
cm = confusion_matrix(y_test, y_pred)
disp = ConfusionMatrixDisplay(confusion_matrix=cm)
disp.plot()


bean = pd.read_csv('/public/bmort/python/beans-unknown.csv')
print(bean)
test_bean = bean[['MajorAxisLength','MinorAxisLength','AspectRatio','Extent','Solidity
bean_pred = rf.predict(test_bean)
print("This is the bean prediction of the sample data")
print(bean_pred)
```

```
Area               0
Perimeter          0
MajorAxisLength    0
MinorAxisLength    0
AspectRatio        0
Eccentricity       0
ConvexArea         0
EquivDiameter      0
Extent             0
Solidity           0
roundness          0
Compactness        0
ShapeFactor1       0
ShapeFactor2       0
ShapeFactor3       1
ShapeFactor4       0
Class              0
dtype: int64
```

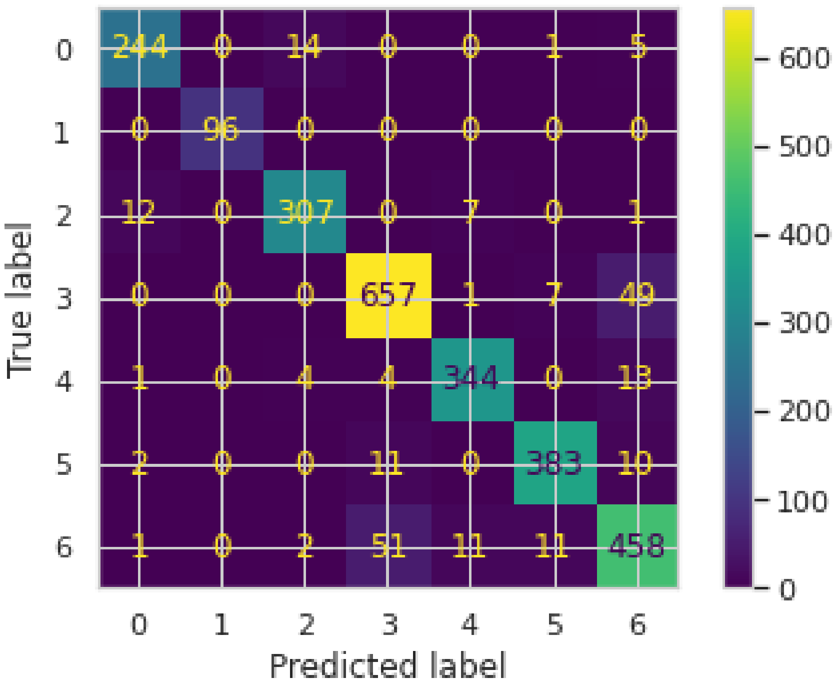
```
There is missing data in the ShapeFactor3 column so lets fix this

                Area     Perimeter  MajorAxisLength  MinorAxisLength  \
count   13533.000000  13533.000000     13533.000000     13533.000000
mean    53057.388384    855.066479       319.922981       202.378468
std     29401.235132    214.781993        85.833897        45.064896
min     20420.000000    524.736000       183.601165       122.512653
25%     36269.000000    703.180000       253.059398       175.884179
50%     44581.000000    793.897000       296.441382       192.500481
75%     61422.000000    977.266000       376.352986       217.263451
max    254616.000000   1985.370000       738.860154       460.198497

        AspectRatio  Eccentricity    ConvexArea  EquivDiameter        Extent  \
count  13533.000000  13533.000000   13533.000000   13533.000000  13533.000000
mean       1.581111      0.750310   53777.120890     253.052530      0.749827
std        0.245328      0.091890   29853.152141      59.324886      0.048939
min        1.024868      0.218951   20684.000000     161.243764      0.555315
25%        1.430641      0.715134   36669.000000     214.893288      0.718753
50%        1.549898      0.764011   45123.000000     238.248383      0.759903
75%        1.704026      0.809699   62388.000000     279.651425      0.786847
max        2.430306      0.911423  263261.000000     569.374358      0.866195

           Solidity     roundness   Compactness  ShapeFactor1  ShapeFactor2  \
count  13533.000000  13533.000000  13533.000000  13533.000000  13533.000000
mean       0.987150      0.873653      0.800347      0.006561      0.001719
std        0.004651      0.059410      0.061485      0.001130      0.000596
min        0.919246      0.489618      0.640577      0.002778      0.000564
25%        0.985675      0.833360      0.763181      0.005890      0.001158
50%        0.988287      0.883447      0.801505      0.006643      0.001700
75%        0.990018      0.917039      0.834520      0.007271      0.002173
max        0.994677      0.990685      0.987303      0.010451      0.003665

       ShapeFactor3  ShapeFactor4
count  13533.000000  13533.000000
mean       0.644336      0.995077
std        0.098687      0.004348
min        0.410339      0.947687
25%        0.582445      0.993717
50%        0.642410      0.996393
75%        0.696423      0.997891
max        0.974767      0.999733
The maginitudes between the columns are much different, some are very small and some are large
The range between the columns varies a fair amount as well with some exceptions where they are
The range of area is large, the range of the convex area is also very large
It is clear to see there are outliers here because there is the max is so far off of the mean.
```



```
                     Area  Perimeter  MajorAxisLength  MinorAxisLength  \
Area             1.000000   0.966904         0.932615         0.952038
Perimeter        0.966904   1.000000         0.977558         0.914326
MajorAxisLength  0.932615   0.977558         1.000000         0.828341
MinorAxisLength  0.952038   0.914326         0.828341         1.000000
AspectRatio      0.243660   0.386073         0.550062        -0.005404
Eccentricity     0.268623   0.391125         0.541075         0.022423
ConvexArea       0.999940   0.967868         0.933384         0.951777
EquivDiameter    0.984997   0.991452         0.962267         0.949208
Extent           0.054675  -0.020630        -0.077350         0.146002
Solidity        -0.197216  -0.304551        -0.284758        -0.156688
roundness       -0.358979  -0.548265        -0.595651        -0.213982
Compactness     -0.269787  -0.407432        -0.567913        -0.018598
ShapeFactor1    -0.848382  -0.865748        -0.775824        -0.947191
ShapeFactor2    -0.641205  -0.768603        -0.859415        -0.475313
ShapeFactor3    -0.273756  -0.408907        -0.567630        -0.022736
ShapeFactor4    -0.357928  -0.431119        -0.484385        -0.266295

                 AspectRatio  Eccentricity  ConvexArea  EquivDiameter  \
Area                0.243660      0.268623    0.999940       0.984997
Perimeter           0.386073      0.391125    0.967868       0.991452
```

```
MajorAxisLength     0.550062      0.541075    0.933384       0.962267
MinorAxisLength    -0.005404      0.022423    0.951777       0.949208
AspectRatio         1.000000      0.924207    0.245229       0.305206
Eccentricity        0.924207      1.000000    0.270393       0.319410
ConvexArea          0.245229      0.270393    1.000000       0.985254
EquivDiameter       0.305206      0.319410    0.985254       1.000000
Extent             -0.371479     -0.319910    0.052892       0.028773
Solidity           -0.269104     -0.298372   -0.206784      -0.232230
roundness          -0.764988     -0.720220   -0.363531      -0.437107
Compactness        -0.987647     -0.970317   -0.271641      -0.328977
ShapeFactor1        0.020914      0.017238   -0.848374      -0.893397
ShapeFactor2       -0.837337     -0.859269   -0.642770      -0.714696
ShapeFactor3       -0.978534     -0.981064   -0.275634      -0.331603
ShapeFactor4       -0.451580     -0.450671   -0.364211      -0.394600

                   Extent  Solidity  roundness  Compactness  ShapeFactor1  \
Area             0.054675 -0.197216  -0.358979    -0.269787     -0.848382
Perimeter       -0.020630 -0.304551  -0.548265    -0.407432     -0.865748
MajorAxisLength -0.077350 -0.284758  -0.595651    -0.567913     -0.775824
MinorAxisLength  0.146002 -0.156688  -0.213982    -0.018598     -0.947191
AspectRatio     -0.371479 -0.269104  -0.764988    -0.987647      0.020914
Eccentricity    -0.319910 -0.298372  -0.720220    -0.970317      0.017238
ConvexArea       0.052892 -0.206784  -0.363531    -0.271641     -0.848374
EquivDiameter    0.028773 -0.232230  -0.437107    -0.328977     -0.893397
Extent           1.000000  0.192236   0.344658     0.355158     -0.141615
Solidity         0.192236  1.000000   0.609621     0.304833      0.154229
roundness        0.344658  0.609621   1.000000     0.766030      0.234064
Compactness      0.355158  0.304833   0.766030     1.000000     -0.005994
ShapeFactor1    -0.141615  0.154229   0.234064    -0.005994      1.000000
ShapeFactor2     0.237762  0.344337   0.781478     0.868350      0.473225
ShapeFactor3     0.348469  0.308660   0.761057     0.998685     -0.005055
ShapeFactor4     0.148651  0.700132   0.472688     0.486344      0.251063

                 ShapeFactor2  ShapeFactor3  ShapeFactor4
Area                -0.641205     -0.273756     -0.357928
Perimeter           -0.768603     -0.408907     -0.431119
MajorAxisLength     -0.859415     -0.567630     -0.484385
MinorAxisLength     -0.475313     -0.022736     -0.266295
AspectRatio         -0.837337     -0.978534     -0.451580
Eccentricity        -0.859269     -0.981064     -0.450671
ConvexArea          -0.642770     -0.275634     -0.364211
EquivDiameter       -0.714696     -0.331603     -0.394600
Extent               0.237762      0.348469      0.148651
Solidity             0.344337      0.308660      0.700132
roundness            0.781478      0.761057      0.472688
Compactness          0.868350      0.998685      0.486344
ShapeFactor1         0.473225     -0.005055      0.251063
ShapeFactor2         1.000000      0.872323      0.531714
```

```
ShapeFactor3         0.872323      1.000000      0.486049
ShapeFactor4         0.531714      0.486049      1.000000

the decision was taken to drop some features with correlations to area over .9,
the benefits of this decision should be a significant reduction of the computational complexity
I chose to use major and minor axis length
0.9995381489007944
0.9194680458071666
    Area  Perimeter  MajorAxisLength  MinorAxisLength  AspectRatio  \
0  37500    728.191       275.840463       173.818266     1.586948
1  37500    715.578       272.171813       175.668301     1.549351
2  37511    718.350       267.039757       179.141937     1.490660
3  37513    720.028       269.589608       177.510928     1.518721
4  37514    725.847       269.881174       177.418223     1.521158

   Eccentricity  ConvexArea  EquivDiameter    Extent  Solidity  roundness  \
0      0.776481       37944     218.509686  0.703406  0.988299   0.888690
1      0.763818       37797     218.509686  0.786229  0.992142   0.920295
2      0.741599       37868     218.541732  0.717365  0.990573   0.913474
3      0.752626       37981     218.547558  0.780545  0.987678   0.909270
4      0.753547       37920     218.550471  0.793309  0.989293   0.894773

   Compactness  ShapeFactor1  ShapeFactor2  ShapeFactor3  ShapeFactor4
0     0.792160      0.007356      0.001787      0.627517      0.995836
1     0.802837      0.007258      0.001860      0.644548      0.998631
2     0.818387      0.007119      0.001970      0.669756      0.998379
3     0.810668      0.007187      0.001915      0.657182      0.998076
4     0.809803      0.007194      0.001908      0.655780      0.997545
This is the bean prediction of the sample data
[3 3 3 3 3]
```


True label
Predicted label
0
1
2
3
4
5
6
244
0
14
0
0
1
5
0
96
0
0
0
0
0
12
0
307
0
7
0
1
0
0
0
657
1
7
49
1
0
4
4
344
0
13
2
0
0
11
0
383
10
1
0
2
51
11
11
458
600
500
400
300
200
100
0