EDA

February 13, 2025

I mainly focused on three parts of the data.

First thing is relationship between features of images of plankton. For example, there are similar features so that I can group them. Grouping features based on their covariances can help to make model more efficiently. I can have enough reasons to remove redundant features or run PCA. Or I can derive new composite features. For example, this is not finalized version but maybe I can combine circulairtyu and convexity. Also, if we use linear regression like logistic regression or SVM in further study, multicollinearity due to highly correlated value can hinder to interpret what feature is critical for classifying. Also in neural network like CNN, redundant information can cause weight redundancy since redundant features will learn same underlying pattern which will cause inefficient learning. Also redundant features can make model overly complex which will cause overfitting.

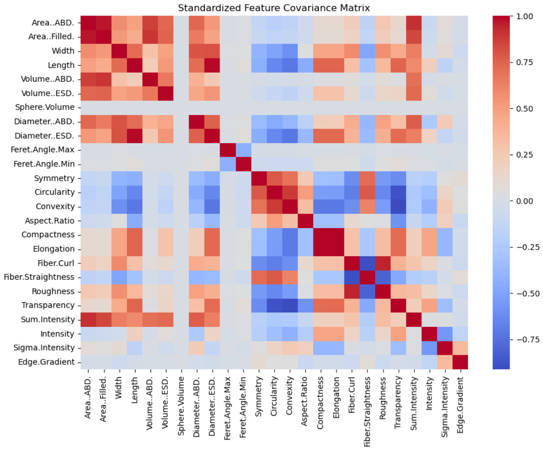
Second thing is unbalanced label. Since in data explanation it says data is unbalanced and we have specific classes want to focus on, I'll discover how unbalanced are they.

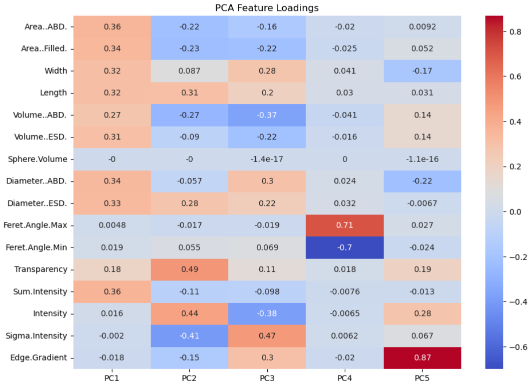
Third thing is variations and unusual values.

Before I do EDA, I used SIMC_OverlapTiffsWithPP data. Also, for better understanding, I merged all data into one and proceeded EDA. In here, notice that 9 files out of 252 files have three extra columns which are {'Biovolume..P..Spheroid.', 'Biovolume..Sphere.', 'Biovolume..Cylinder.'}, so I deleted them.

Before I use covariance, I grouped features into three based on my knowledge. Those are based on shape and size, structural and shape, and optical. This is what they are and what columns are there. You can find the definition of columns in GlossaryParticleProperties.pdf. shape_and_size_cols = ["Area..ABD.", "Area..Filled.", "Width", "Length", "Volume..ABD.", "Volume..ESD.", "Sphere.Volume", "Diameter..ABD.", "Diameter..ESD.", "Feret.Angle.Max", "Feret.Angle.Min"] structural_and_shape_cols = ["Symmetry", "Circularity", "Convexity", "Aspect.Ratio", "Compactness", "Elongation", "Fiber.Curl", "Fiber.Straightness", "Roughness"] optical_cols = ["Transparency", "Sum.Intensity", "Intensity", "Sigma.Intensity", "Edge.Gradient"].

Now I draw heatmap based on each features covariance. Before draw heatmap, I standardize the data since the range of data is really big and each features have different ranges. As we can see in the heatmap, my first intuition regarding grouping features are almost right. We can mainly see three groups of red. It means features inside of each group are strongly related to each others. This is understandable since when the area of plankton inside of the photo is big, then it means the plankton has higher possibility to have long width and length and volume. Similar as other groups, like structural and shape group since how symmetry they are literally means how circular they are. However we can also find that shape and size features group have low relationship with structural. Since we understand that shape and size of planktons are independent with structure, this is reasonable. Interesting part is that dimension and size groups is somewhat related with optical features. We can think of some reasons. First, larger plankton may block more light and make them appear darker and this is the reason why we have relation between width, length, diameter with transparency and sum intensity. Also, since Transparency = 1-(ABD Diameter/ESD Diameter) this also proves why we have correlation between dimension and size group with optical group. Also as we can see in PCA with 5 PCs, PC1 shows


Standardized Feature Covariance Matrix
Area..ABD.
Area..Filled.
Width
Length
Volume..ABD.
Volume..ESD.
Sphere.Volume
Diameter..ABD.
Diameter..ESD.
Feret.Angle.Max
Feret.Angle.Min
Symmetry
Circularity
Convexity
Aspect.Ratio
Compactness
Elongation
Fiber.Curl
Fiber.Straightness
Roughness
Transparency
Sum.Intensity
Intensity
Sigma.Intensity
Edge.Gradient


PCA Feature Loadings
PC1
PC2
PC3
PC4
PC5
Area..ABD. 0.36 -0.22 -0.16 -0.02 0.0092
Area..Filled. 0.34 -0.23 -0.22 -0.025 0.052
Width 0.32 0.087 0.28 0.041 -0.17
Length 0.32 0.31 0.2 0.03 0.031
Volume..ABD. 0.27 -0.27 -0.37 -0.041 0.14
Volume..ESD. 0.31 -0.09 -0.22 -0.016 0.14
Sphere.Volume -0 -0 -1.4e-17 0 -1.1e-16
Diameter..ABD. 0.34 -0.057 0.3 0.024 -0.22
Diameter..ESD. 0.33 0.28 0.22 0.032 -0.0067
Feret.Angle.Max 0.0048 -0.017 -0.019 0.71 0.027
Feret.Angle.Min 0.019 0.055 0.069 -0.7 -0.024
Transparency 0.18 0.49 0.11 0.018 0.19
Sum.Intensity 0.36 -0.11 -0.098 -0.0076 -0.013
Intensity 0.016 0.44 -0.38 -0.0065 0.28
Sigma.Intensity -0.002 -0.41 0.47 0.0062 0.067
Edge.Gradient -0.018 -0.15 0.3 -0.02 0.87