1. **Prepare a model for glass classification using KNN**

**Data Description:**

**RI : refractive index**

**Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)**

**Mg: Magnesium**

**AI: Aluminum**

**Si: Silicon**

**K:Potassium**

**Ca: Calcium**

**Ba: Barium**

**Fe: Iron**

**Type: Type of glass: (class attribute)**

**1 -- building\_windows\_float\_processed**

**2 --building\_windows\_non\_float\_processed**

**3 --vehicle\_windows\_float\_processed**

**4 --vehicle\_windows\_non\_float\_processed (none in this database)**

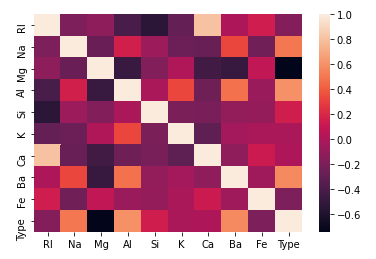
**5 --containers**

**6 --tableware**

**7 --headlamps**

**Soln. :**

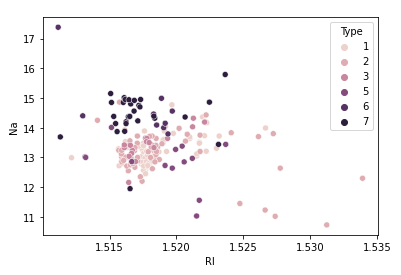
# Data exploration and visualizaion



We can notice that Ca and K values don't affect Type that much.

Also Ca and RI are highly correlated, this means using only RI is enough.

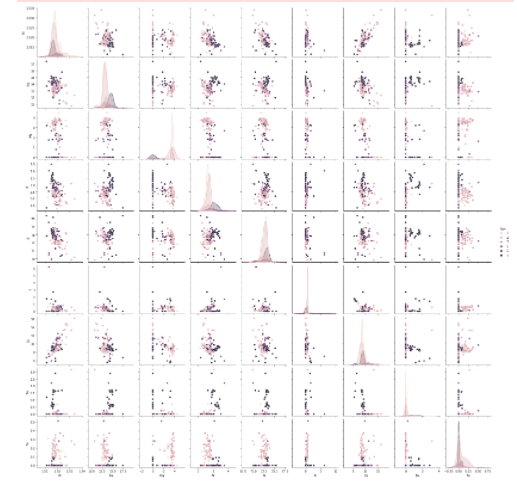
So we can go ahead and drop Ca, and also K.



Suppose we consider only RI, and Na values for classification for glass type.

* From the above plot, We first calculate the nearest neighbors from the new data point to be calculated.
* If the majority of nearest neighbors belong to a particular class, say type 4, then we classify the data point as type 4.

But there are a lot more than two features based on which we can classify. So let us take a look at pairwise plot to capture all the features.



The pair plot shows that the data is not linear and KNN can be applied to get nearest neighbors and classify the glass types

Feature Scaling

Scaling is necessary for distance-based algorithms such as KNN. This is to avoid higher weightage being assigned to data with a higher magnitude.

Using standard scaler we can scale down to unit variance.

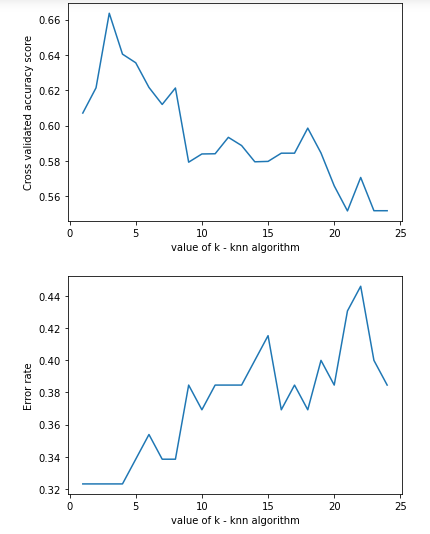
**Apply KNN**

With this setup, We found the accuracy to be 73.84%

Finding the best K value

We can do this either -

* by plotting Accuracy
* or by plotting the error rate

****

we can see that k=4 produces the most accurate results

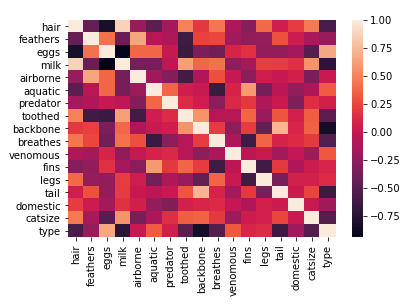
# Findings

* Manhattan distance produced better results (improved accuracy - more than 5%)
* Applying feature scaling improved accuracy by almost 5%.
* The best k value was found to be 4.
* Dropping ‘Ca’ produced better results by a bit, ‘K’ feature did not affect results in any way.
* Also, we noticed that RI and Ca are highly correlated, this makes sense as it was found that the Refractive index of glass was found to increase with the increase in Cao.

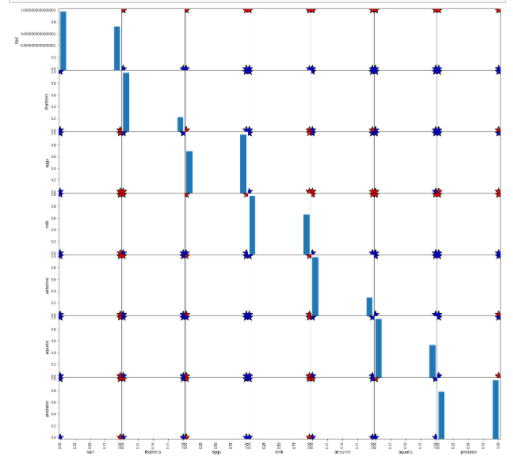
1. **Implement a KNN model to classify the animals in to categorie**

**Soln. :**

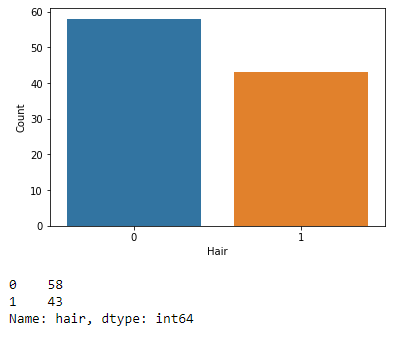
# Data exploration and visualizaion



# Plotting scatter matrix



# Visualizing has hair or not ?

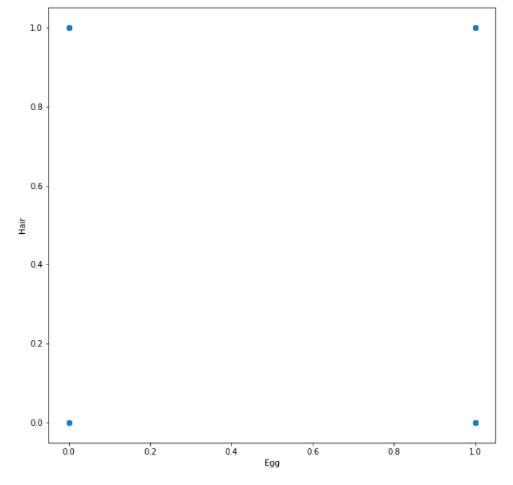


# KNN

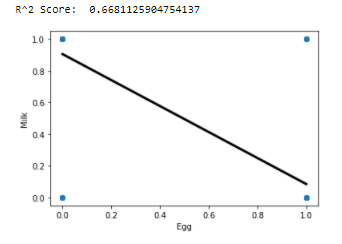
# Train Test Split



# Visualizing Eggs and Hair on Scatter

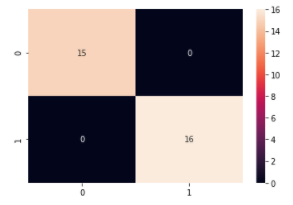


# Linear Regression

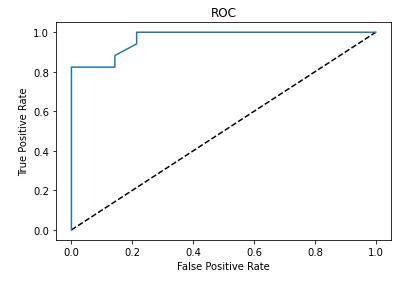


Ridge Score: 0.930239727992853

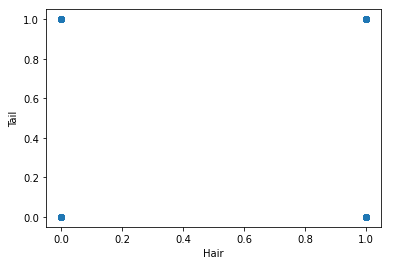
Lasso score: 0.9999970989932222



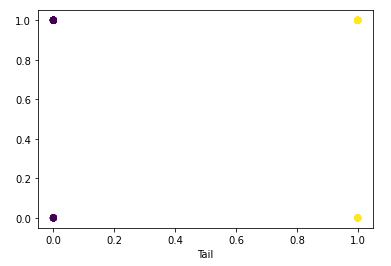
# Logistic Regression



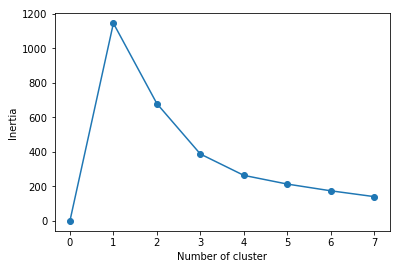
# Support Vector Machine



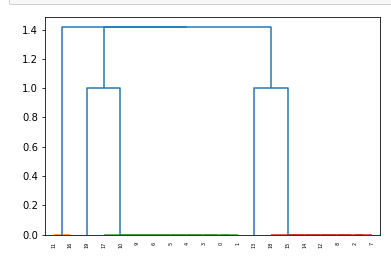
# K-Means Clustering



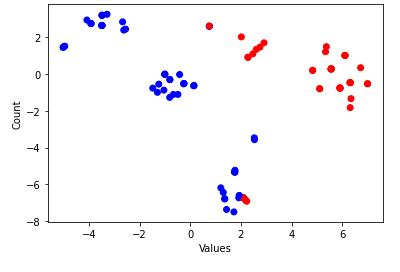
# Inertia



# Dendogram



# t-distributed Stochastic Neighbor Embedding



# PCA

