1.1. Let's try determining the type of disease based on the patient's Age.

Model 1:

The gradient descent algorithm can be explained in these steps:

I. Choose a starting point of initialization
II. Calculate the gradient at this point
III. Take a step towards the opposite direction in order to minimize the cost function
IV. Calculate the cost function
V. Update θ_j with adequately choosing the value of α
VI. Repeat the last three steps until convergence is reached or stop after i iterations or when the change in slope is below a threshold value.

Using the gradient algorithm, the optimal intercept came out to be **-0.09886794**
While the gradient is **-0.40586823**. The learning rate was taken as **0.001**

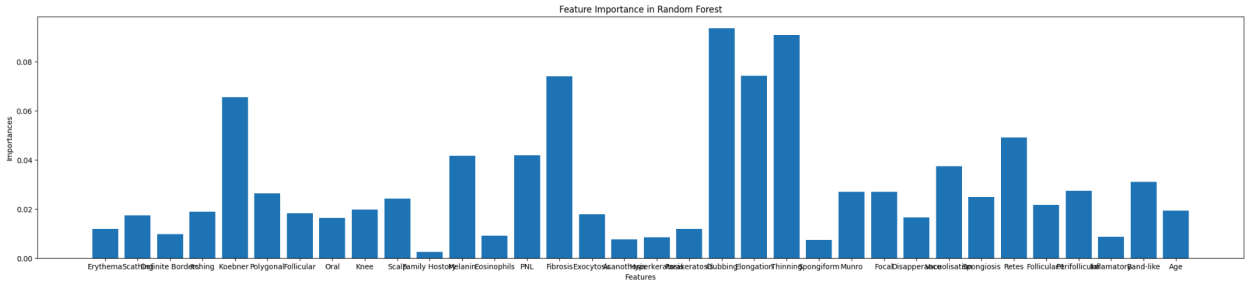
According to our equation:
y = c + m1x1
= (-0.09886794) + (-0.40586823)x1
This was reached in 35 iterations

Note: The dataset was scaled before the processing, although the gradient is negative.

1.2. **Use random forest on the clinical as well as histopathological attributes to classify the disease type (*model2*).**

Model 2:

A random forest model was built over all the attributes, i.e. clinical and histopathological while the disease was the outcome variable.
The average accuracy of the random forest was found to be **0.9664285714285714**.
The weighted precision average was **0.98** while the weighted recall average was also **0.98**
This random forest model indicates a high accuracy when all 33 features are taken into consideration.
The robust performance of this model indicates its accuracy in classifying diseases using
Both clinical and histopathological attributes. Hence, this model can be used for classification.

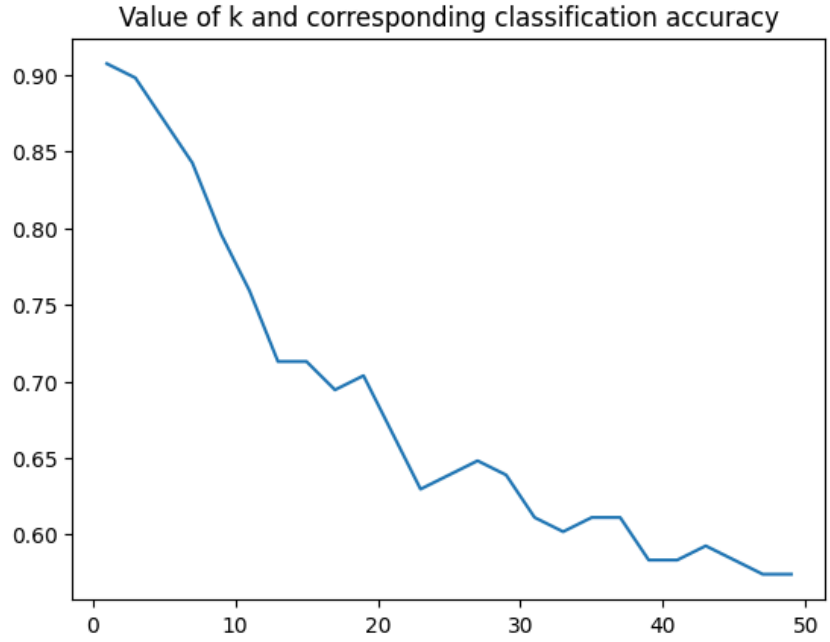


Here in the visualization we can see how the various features impact the model. Since all of them together give a pretty high accuracy, we will continue with this model.

1.3. **Use kNN on the clinical attributes and histopathological attributes to classify the disease type and report your accuracy (*model3*).**

Model 3:

The value for K for optimal accuracy is **1** while the accuracy itself was about
Accuracy= **0.90**
This model was considering all 33 attributes. I tried removing features such as Elongation, Hyperkeratosis, Spongiform but the accuracy dropped to about 80%. Since removing features did not account for better accuracy, we will be considering all of the features.

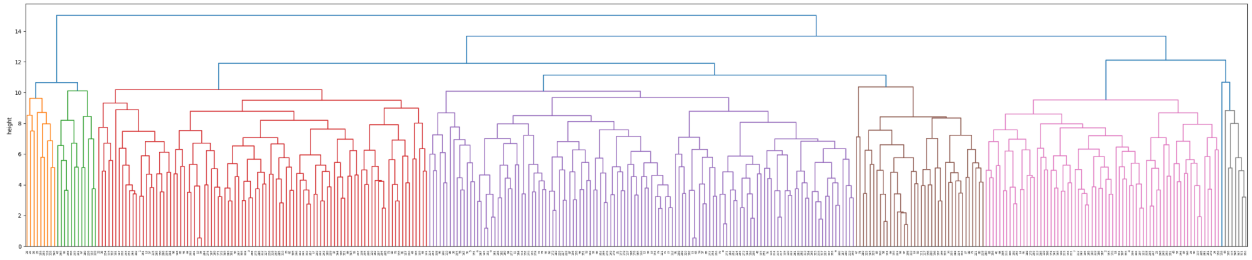


1.4. **Finally, use two different clustering algorithms and see how well these attributes can**

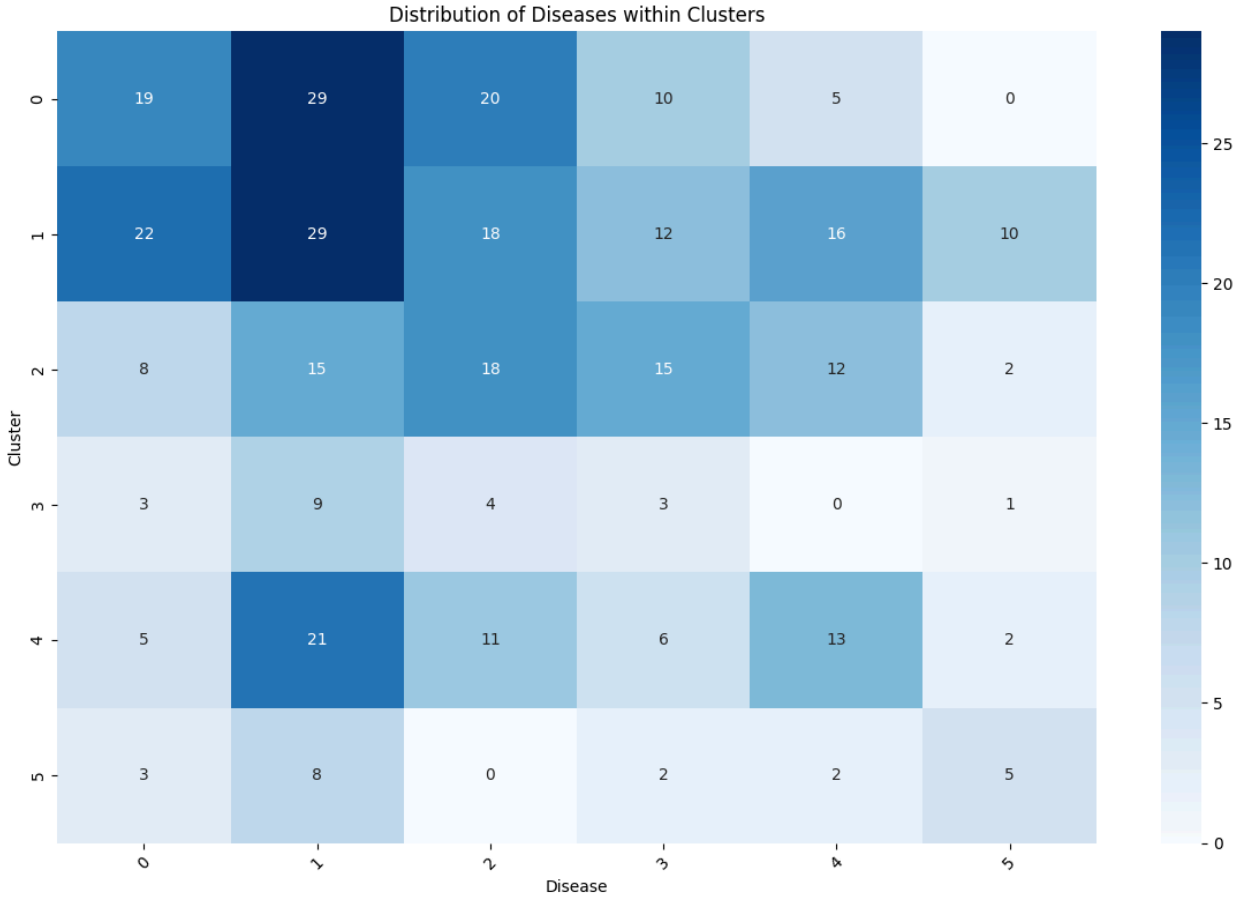
determine the disease type (*model4* and *model5*)

Model 4:

Here we modeled using an agglomerative algorithm.
We took **k = 6**, and used this to build this model.
The code was able to predict the clusters and classify the diseases accordingly.



This gives us a better insight as to which disease belonged to which clusters and how they were grouped. For example, Disease 4 and Cluster 1 had 26 datapoints in them.

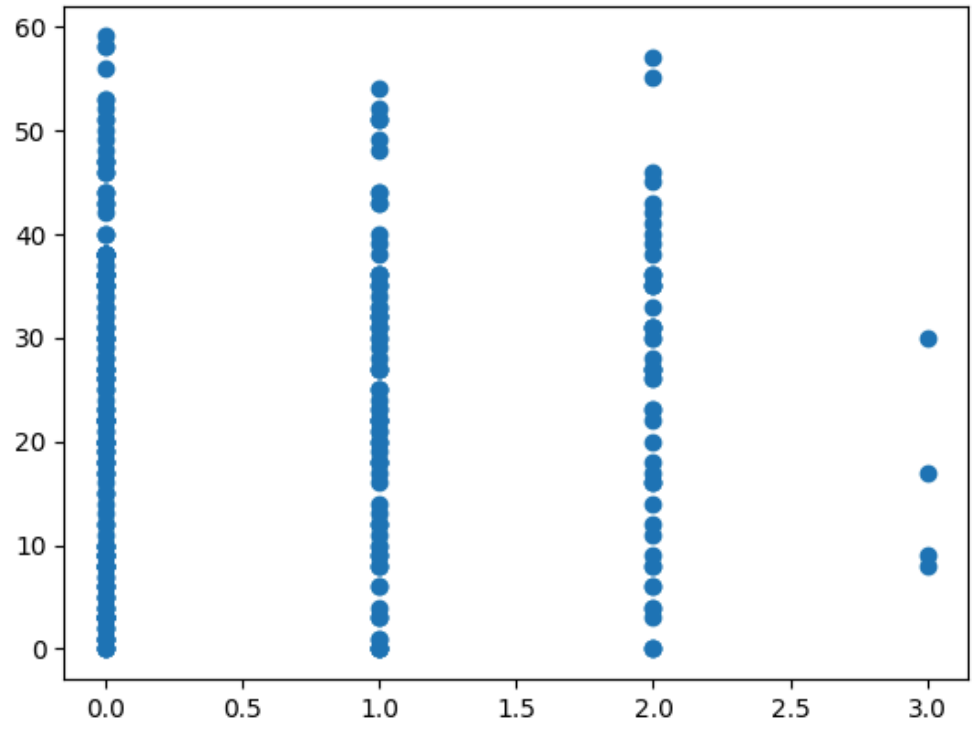


However, the accuracy of this model is quite low averaging at about **0.3**

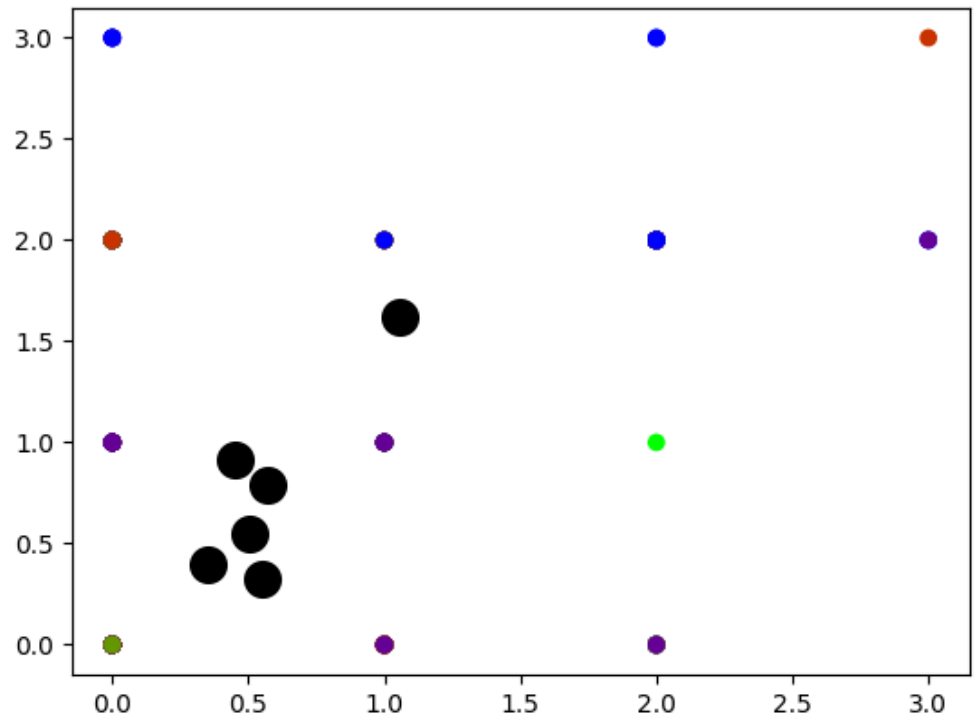
This may not be the best model to classify the diseases

Model 5:

For the divisive algorithm, we again consider 33 attributes and we will see how well this model classifies the diseases. The silhouette score was calculated to be about 0.26. Below we see the scatter plot of the dataset:



And here is the plot for the divisive algorithm:
The centroids of the clusters are tightly packed while the accuracy of this model is about 2%.



Now, compare and contrast the five models you built.
The techniques that have been used above display varying levels of success in classifying and clustering diseases based on all the attributes. Gradient descent model converged in about 35 iterations with a negative gradient that would indicate a decreasing cost. Random forest gave a very good accuracy when all features were considered. KNN also indicated high accuracy for classifying the diseases. Grouping diseases in clusters indicated relatively lower accuracy averaging at around 30% indicating poor performance.

Out of all the models, Random forest and KNN demonstrated higher accuracy in disease classification as compared to the clustering algorithms. This suggests that supervised learning methods might be more suitable for this classification. The availability of labeled data is important for the classification tasks.