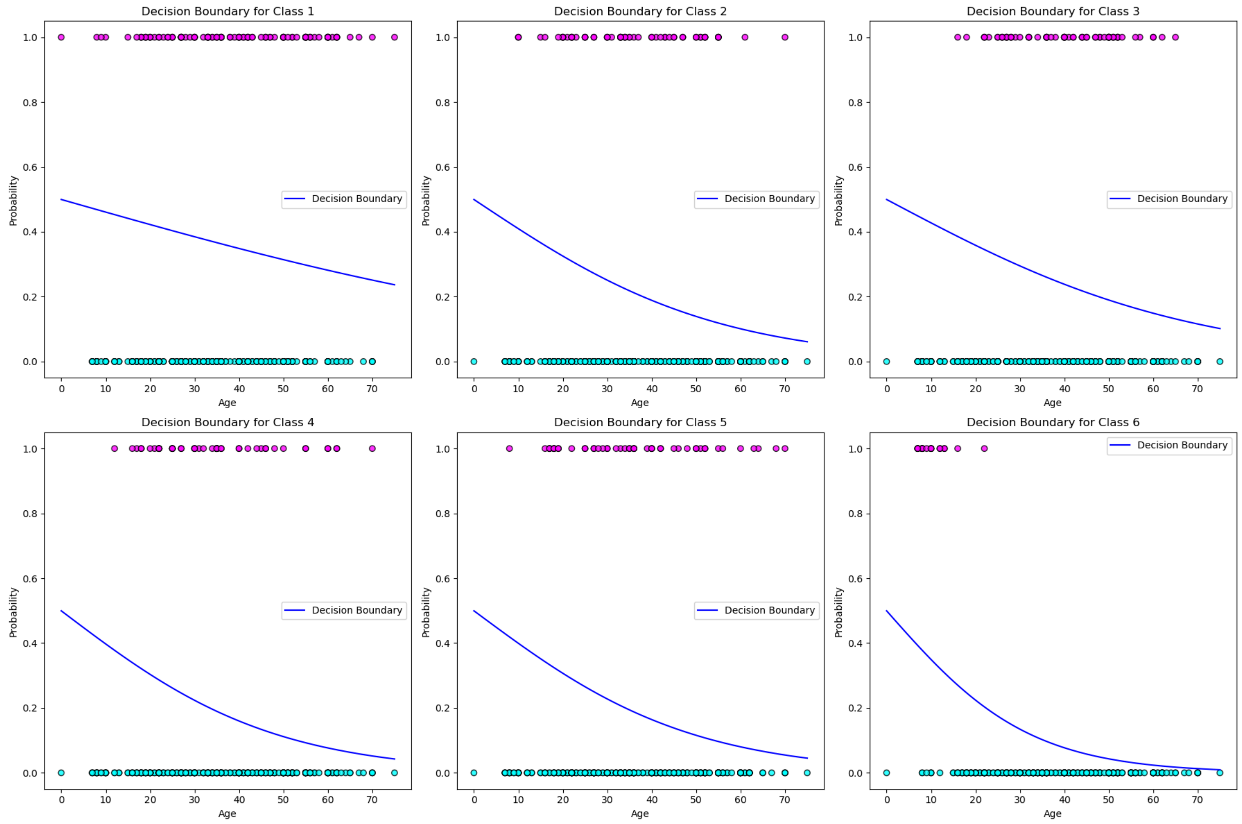
Dermatology Disease Classification

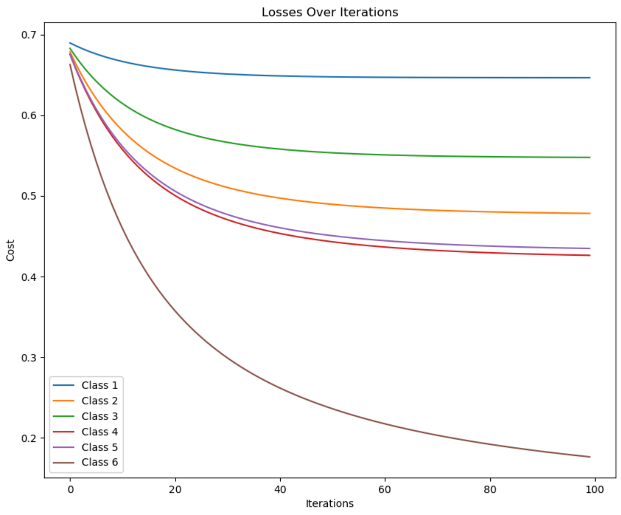
Part 1: Gradient Descent - Classifying Disease by Patient Age

When using the gradient descent algorithm to make a model which classifies disease by patient age, I found that this method was not accurate.

Gradient Descent Accuracy: 31%

This is likely due to the complex nature of the dataset and interactions/importance of each feature. This model was initialized with an alpha = 0.0001 and iterated 100 times. Following is the entropy loss for each iteration of the algorithm for each disease and the decision boundaries for each disease:



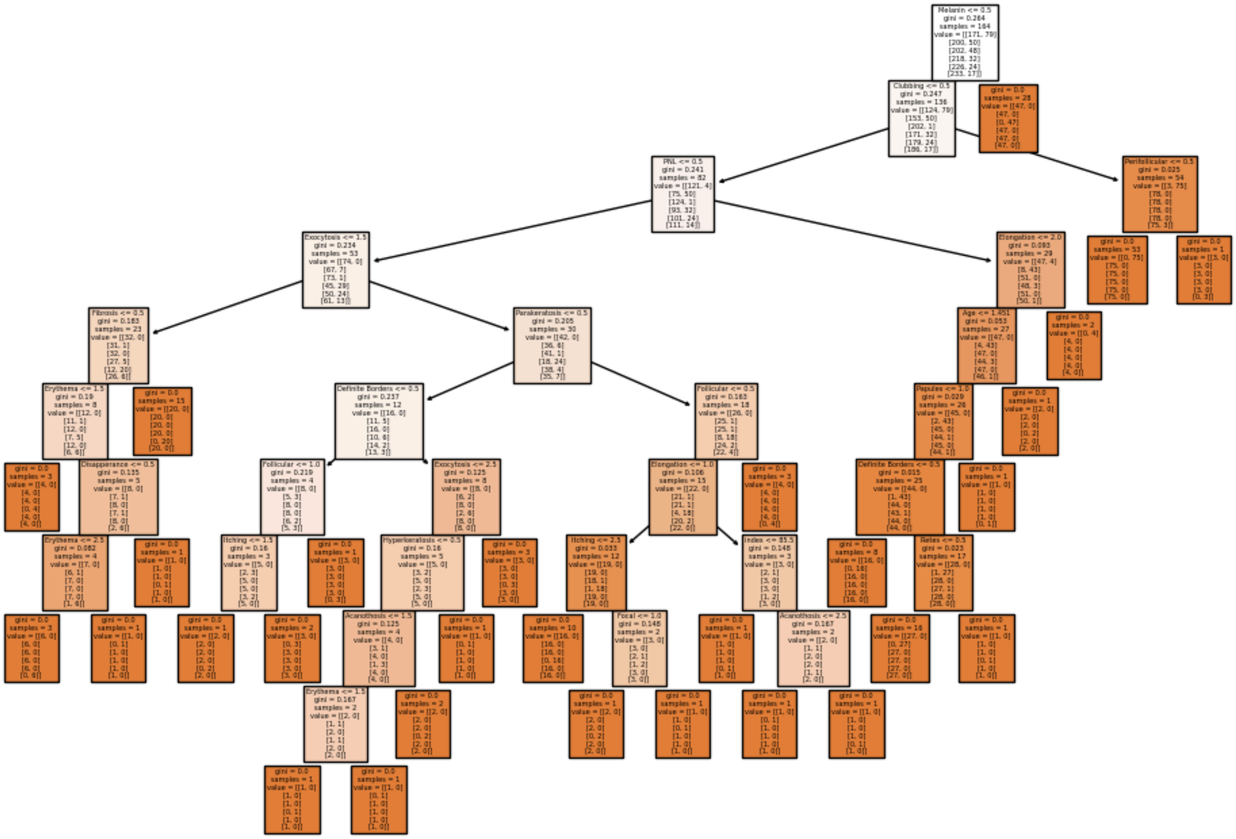


Part 2: Random Forest on Clinical and Histopathological Features

Next, a random forest classifier was fit to all features in the data set. The data was split into random training (70%) and testing (30%) subsets, then fit to a unique random forest model. This process was completed 50 times and proved to be very accurate at classifying disease types. Disease types were one-hot encoded and patient age was normalized.

Mean Random Forest Accuracy (50 Iterations): 94.57%

To help visualized this process, here is the first random tree of the final iteration:



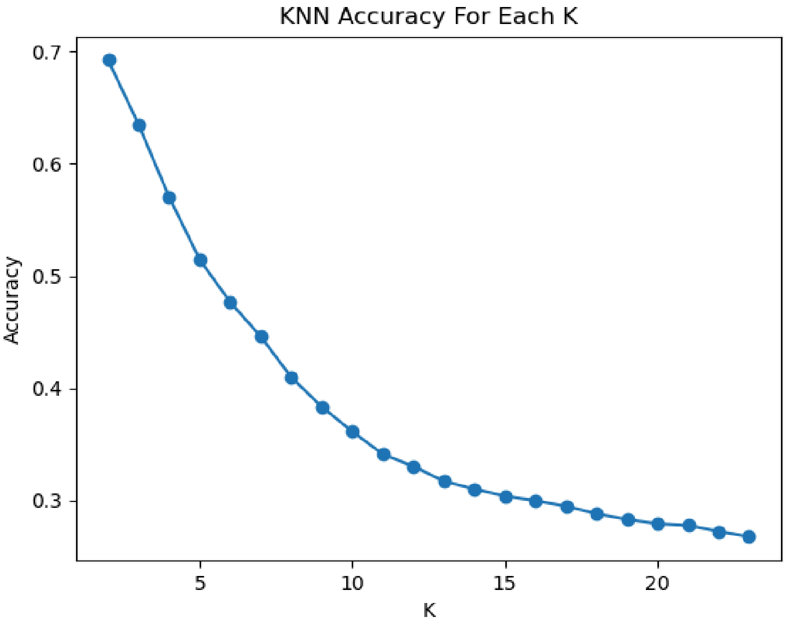
Part 3: k-Nearest Neighbors on Clinical and Histopathological Features

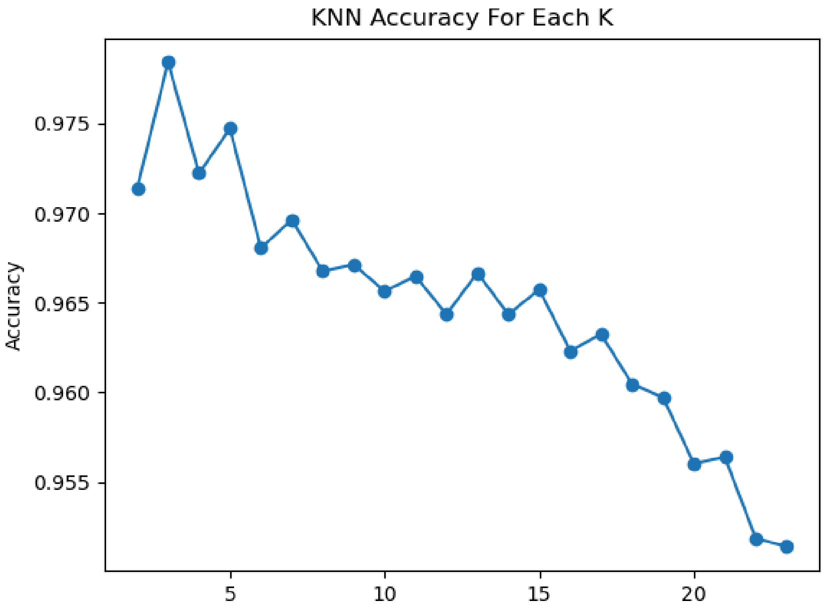
Initially, using KNN to classify each disease did not prove to be as accurate as I had hoped. For the initial model, I had used all features (age normalized) and evaluated the mean accuracy over 100 iterations for k values 2 - 24. The value of k = 24 was found by taking the square root of N = 358 (observations) and adding 5. This range provides comprehensive possibilities of k for testing. Additionally, the data was split into random training (70%) and testing (30%) subsets, then fit to each KNN model. Following is the highest performing model:

K-Neighbors: 2

KNN Model Accuracy (100 iterations): 69.22%

To try to improve results, I then Sequential Feature Selection to determine the most important features to use for this model. This was be done by comparing the accuracy of models with the 33 (n-features - 1) through 17 (n-features / 2) best features.





The following model performs substantially better and proves to be the most accurate when predicting patient disease.

Number of Features: 22

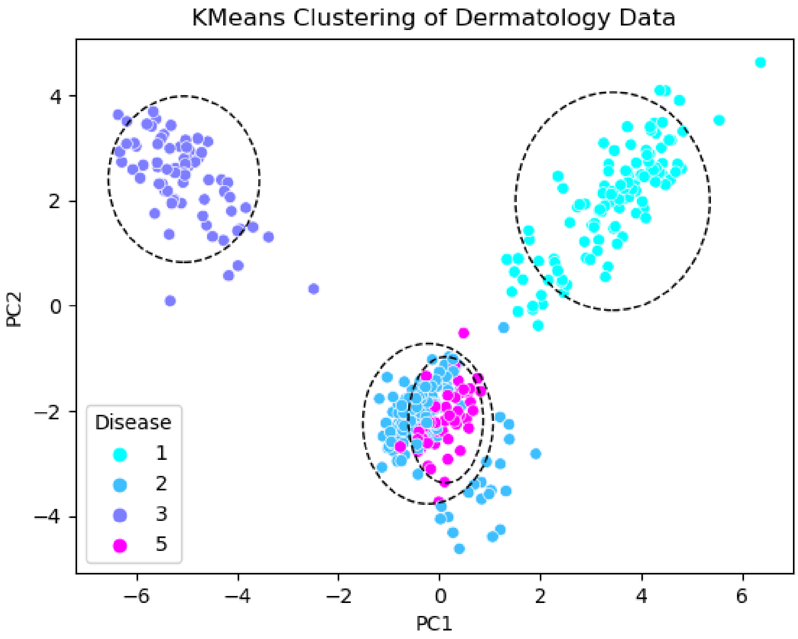
Features Used: 'Koebner', 'Polygonal', 'Papules', 'Oral', 'Knee', 'Scalp', 'Family History', 'Melanin', 'Fibrosis', 'Clubbing', 'Elongation', 'Thinning', 'Spongiform', 'Munro', 'Focal', 'Disappearance', 'Vacuolization', 'Spongiosis', 'Retes', 'Follicular', 'Perifollicular', 'Band-like'

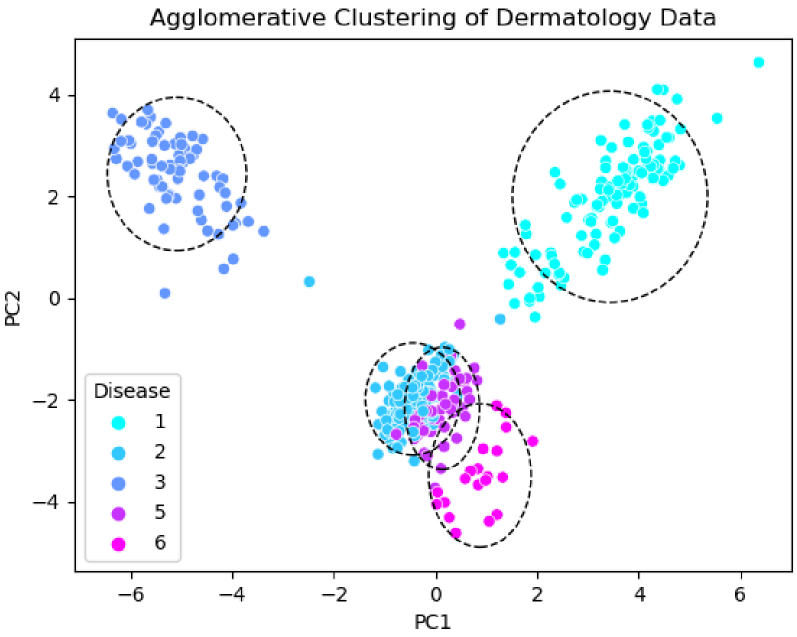
K-Neighbors: 3

KNN Model Accuracy (100 Iterations): 97.84%

Part 4: KMeans and Agglomerative Clustering on Clinical and Histopathological Patient Data

When using clustering methods to classify patient data, I classified each cluster using 6 clusters given the 6 disease types. All features were normalized for both KMeans and Agglomerative clustering methods. (I had initialized these models with only patient age standardized at first, but accuracy scores were severely affected, both > 35%) Data splitting was not performed to evaluate overall model classification. Given that these clustering models do not inherently label each cluster with the associated disease type, the resulting disease clusters were created by mapping the initial model clusters to the disease most present in the cluster. Due to the nature of the clusters, not every disease was represented in each model. To help visualize these clusters, primary component decomposition was performed to find the 2 most impactful features in the dataset.


KMeans Clustering of Dermatology Data


Agglomerative Clustering of Dermatology Data

The clusters in KMeans only represented 4 of the 6 diseases, with one of the clusters being embedded in another.

KMeans Model Accuracy: 81.01%

Agglomerative clustering performed slightly better as it had more distinguished clusters which represented 5 of the 6 diseases.

Agglomerative Model Accuracy: 86.31%

Overall Results:

Using Gradient Descent, Random Forest, k-Nearest Neighbors , KMeans Clustering, and Agglomerative Clustering, shows important distinctions regarding the performance of different machine learning methods on a single dataset. Gradient Descent yielded a low accuracy of 31% when classifying diseases based solely on patient age, which is likely due to the complex nature of the data. This highlights the limitations of using linear models with insufficient features to classify complex data.

Accuracy metrics greatly improved when implementing more robust models. Random Forest classification demonstrated a high mean accuracy score of 94.57% over 50 iterations. By utilizing all features in the dataset, with little manipulation, this method was able to effectively capture the relationships between features and provide reasonably accurate predictions. Through this, Random Forest Classification shows the importance of using a diverse set of features when making predictions. KNN, on the other hand, showed the importance of being more selective with included features. When using all features in the dataset, KNN achieved a max mean accuracy of 69.22%. However, after applying Sequential Feature Selection, the model's mean accuracy improved significantly to 97.84%. The final model reduced the included features to 22 from 34, indicating that feature selection can greatly improve the performance of algorithms sensitive to irrelevant or redundant data.

KMeans and Agglomerative Clustering showed the abilities of unsupervised learning when grouping patient data based on similar features. KMeans performed moderately well and achieved an accuracy of 81.01%. However, it only represented 4 of the 6 diseases. Agglomerative Clustering performed slightly better with an accuracy of 86.31% and was able to identify 5 of the 6 diseases. This is important because despite the models not inherently labeling clusters, they were still able to distinguish the underlying structure of the data. While in this instance the target labels were pre-defined, these results highlight the ability of clustering methods to show identifying features of specific groups of data. One thing to note, though, is the reasonable amount of additional data manipulation required to achieve these results.