Genetic Variations – Clustering

Approach:

Attempt to cluster presence of genetic markers into related groups. Identify the best clustering model.

Results

Model	Accuracy	Precision_0	Precision_1	Recall_0	Recall_1	Recall_Diff
Logistic Regression	0.55	0.53	0.56	0.47	0.62	0.16
Logistic Regression - penalty = l1, solver = liblinear	0.55	0.53	0.56	0.47	0.62	0.15
Logistic Regression - penalty = l2, solver = liblinear	0.55	0.53	0.56	0.47	0.62	0.16
Logistic Regression - solver = lbfgs	0.55	0.53	0.56	0.47	0.62	0.16
Random Forest	0.52	0.49	0.54	0.50	0.53	0.03
Random Forest - minsamples = 10%, maxdepth = 10	0.55	0.54	0.55	0.35	0.72	0.37
Random Forest - minsamples = 1%, maxdepth = 10	0.54	0.52	0.55	0.39	0.67	0.28
Random Forest - minsamples = 1%, maxdepth = 20	0.53	0.51	0.54	0.40	0.65	0.25
Random Forest - minsamples = 1%, maxdepth = 50	0.53	0.50	0.54	0.43	0.62	0.19
Random Forest - minsamples = 20%	0.55	0.54	0.56	0.40	0.69	0.28
Random Forest - minsamples = 20%, maxdepth = 5	0.56	0.54	0.56	0.41	0.69	0.28
Random Forest - minsamples = 20%, maxdepth = 10	0.55	0.54	0.56	0.40	0.69	0.29
KNN-5	0.52	0.50	0.54	0.46	0.58	0.12
KNN-6	0.51	0.49	0.54	0.59	0.44	0.16
KNN-7	0.53	0.50	0.54	0.45	0.60	0.15
KNN-8	0.52	0.50	0.56	0.60	0.45	0.15
SVM-Linear, Minority Weighted	0.52	0.00	0.52	0.00	1.00	1.00
SVM-RBF	0.55	0.55	0.56	0.38	0.71	0.33
Neural Network - MI = 800, Activation = ReLu	0.55	0.54	0.56	0.39	0.70	0.31
Neural Network - MI = 800, Activation = Logisitic	0.55	0.53	0.55	0.37	0.71	0.34
DecisionTree (max_depth=3)	0.55	0.52	0.57	0.55	0.54	0.00
DecisionTree (max_depth=5)	0.54	0.52	0.57	0.55	0.54	0.01
DecisionTree (min_samples_split=20)	0.51	0.49	0.53	0.48	0.54	0.07
DecisionTree (max_features="sqrt")	0.50	0.48	0.52	0.50	0.50	0.00
LDA (No Shrinkage)	0.55	0.53	0.56	0.47	0.62	0.15
LDA (Shrinkage=0.5)	0.54	0.52	0.56	0.50	0.58	0.08
LDA (Shrinkage=1.0)	0.54	0.52	0.57	0.53	0.56	0.03

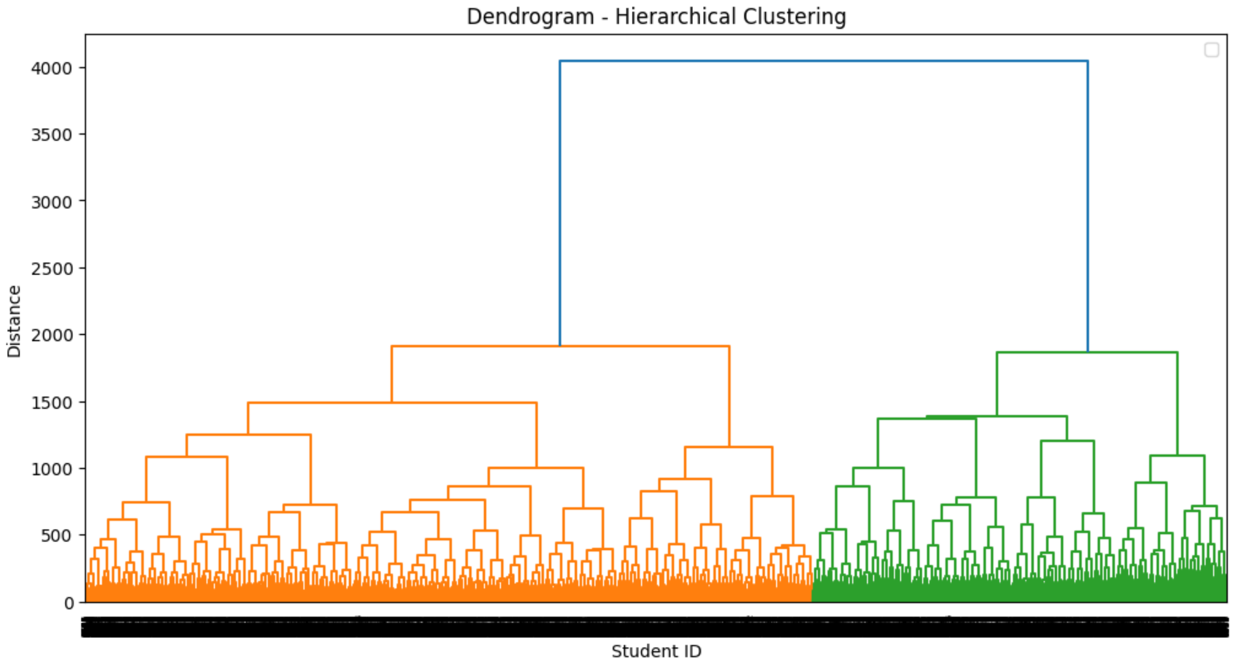
Method:

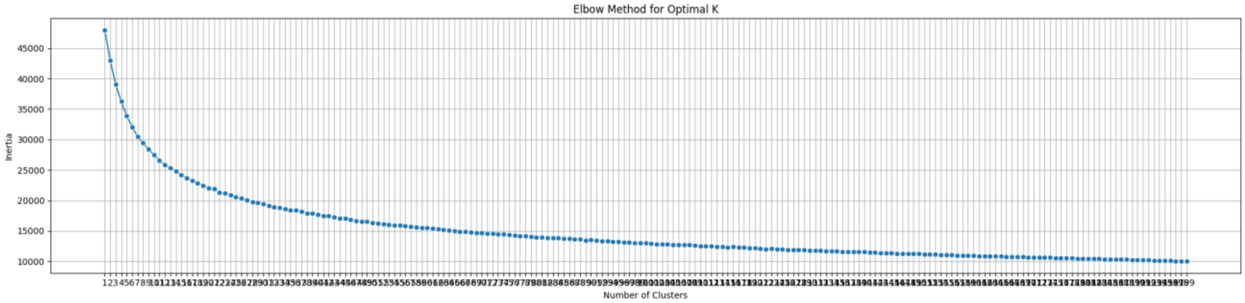
The genetic data was clustred using Kmeans clustering to created 7 different groupings based on visual inspection of the elbow chart. There was no clear bend of the elbow so 7 was selected as the best k value. That was then indexed back to the original data frame and all the non genetic data was kept along with the clustering labels.

Results:

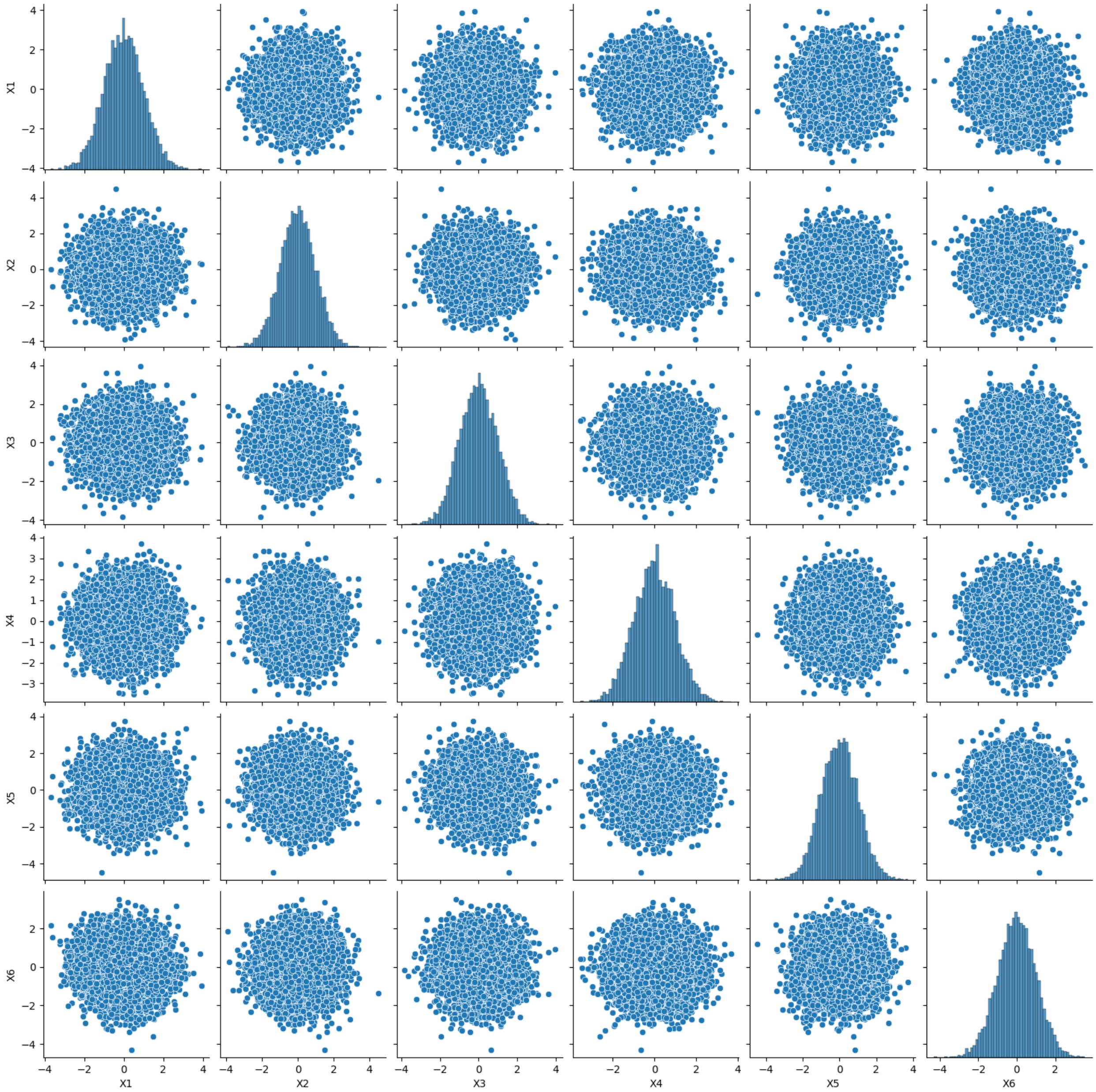
The best model is the decision tree with a max depth of 3 with the highest accuracy of 55% and a very even split between the recall values. All the models performed poorly with low accuracy well into the 50's indicating that none are acceptable. Cross validation was run on the top 5 selected models and of them the decision tree with a max depth of 3 performed well with tight scores across accuracy and recall.

Appendix

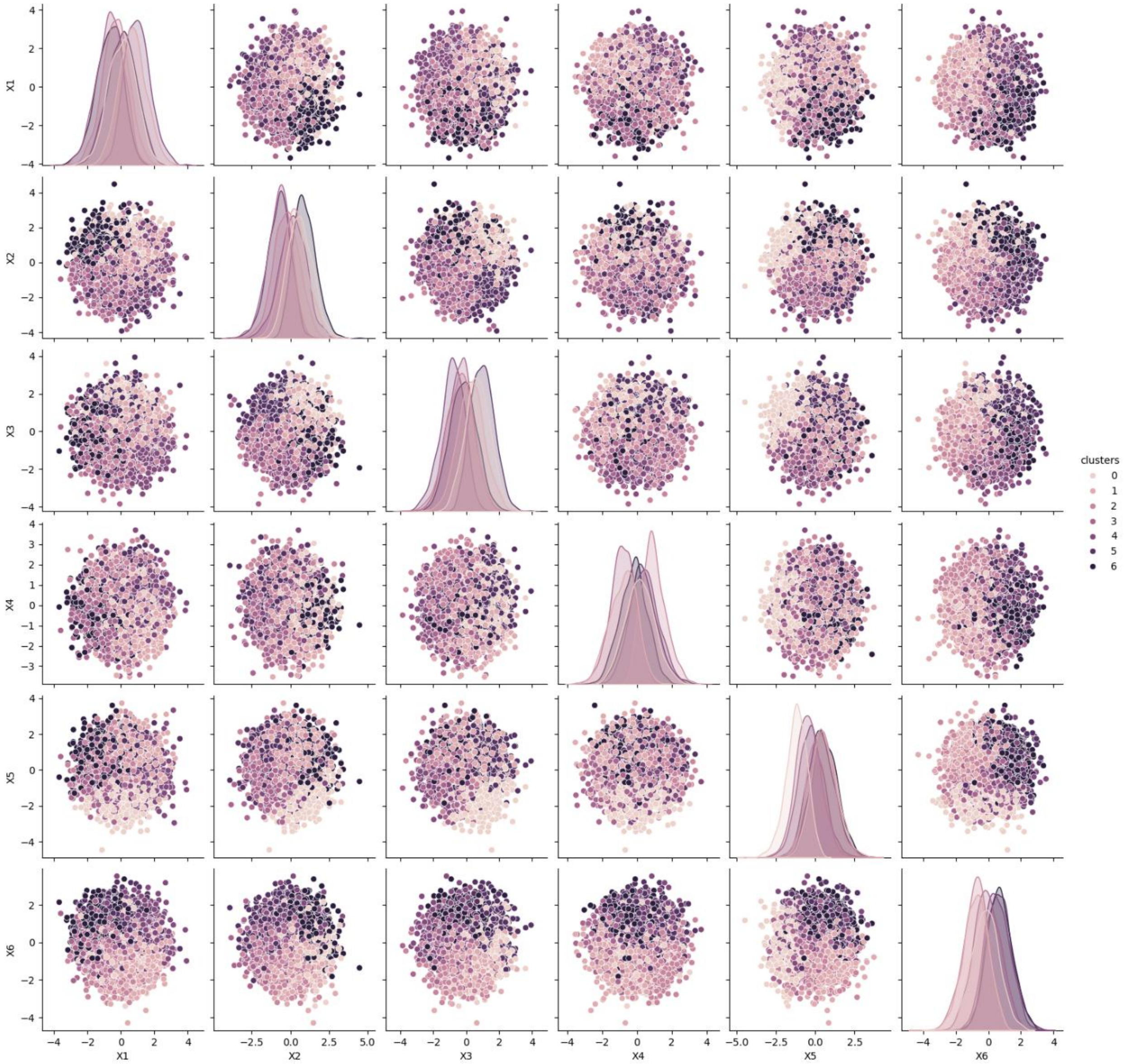


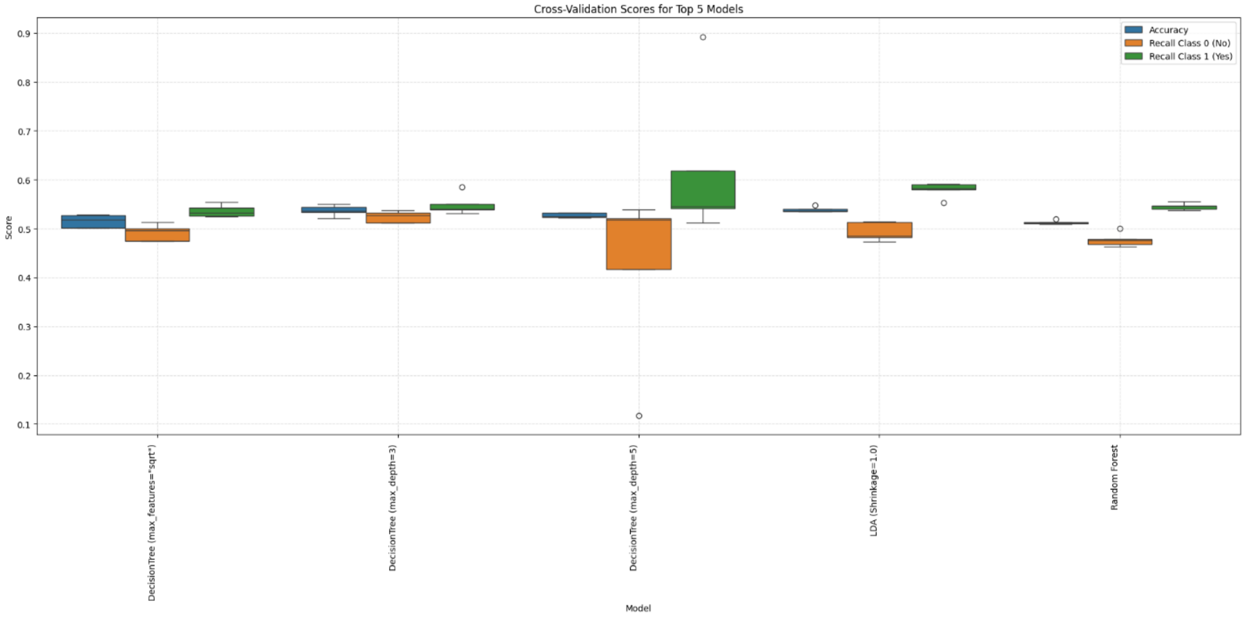


Pairplot of Unclustered Data



Pairplot of clustered Data




Cross-Validation Scores for Top 5 Models
Accuracy
Recall Class 0 (No)
Recall Class 1 (Yes)
Score
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
DecisionTree (max_features="sqrt")
DecisionTree (max_depth=3)
DecisionTree (max_depth=5)
LDA (Shrinkage=1.0)
Random Forest
Model