1: Selecting the Dataset

For this assignment, we used the Wisconsin Breast Cancer Dataset. It was preprocessed by adding a feature header to the top of the file. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

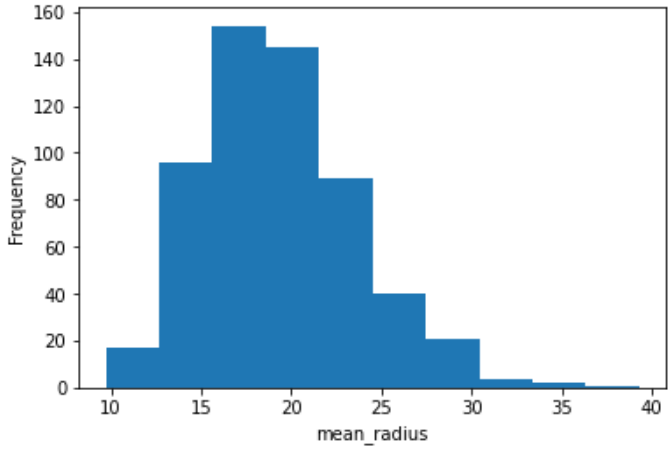
2: Loading the Data

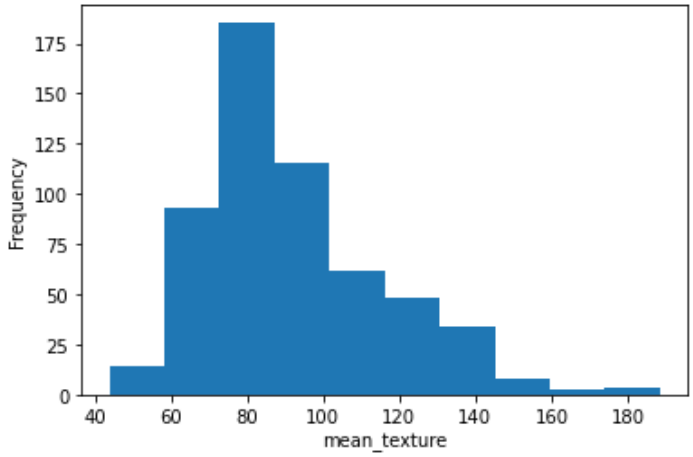
For loading the data we trimmed out unique info (the first column for UID) and made it into four parts. The first part was the data, including the target in the first column, and all of the samples and predictors. The second part was just the target column. For the third part, it was just the target name from the header of the target column. The fourth and last part was an array of all of the feature names in the same order as the corresponding data, also from the header.

3: Meet the Data

The data has 30 features, which are as follows: mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness, mean concavity, mean concave points, mean symmetry, mean fractal dimension, standard error of radius, standard error of texture, standard error of perimeter, standard error of area, standard error of smoothness, standard error of compactness, standard error of concavity, standard error of concave points, standard error of symmetry, standard error of fractal dimension, largest radius, largest texture, largest perimeter, largest area, largest smoothness, largest compactness, largest concavity, largest concave points, largest symmetry, largest fractal dimension. The target is ‘diagnosis.’ There are 569 samples. The first five rows are printed at the end of the document due to their size.

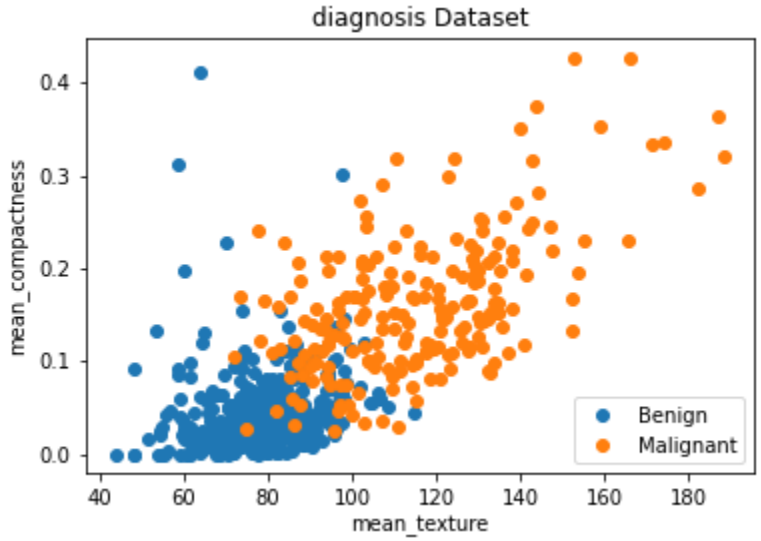
Here are the histograms of the first two features of the dataset, mean radius and mean texture:



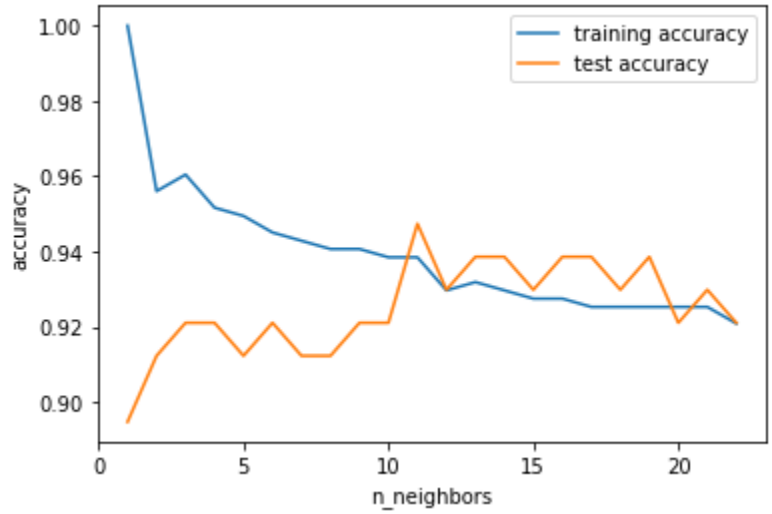


This shows that at between 15 and 20 mean radius and at about 80 mean texture is the concentration of our results.

The pair plot below shows that there is a correlation between malignant diagnosis and the texture and compactness.

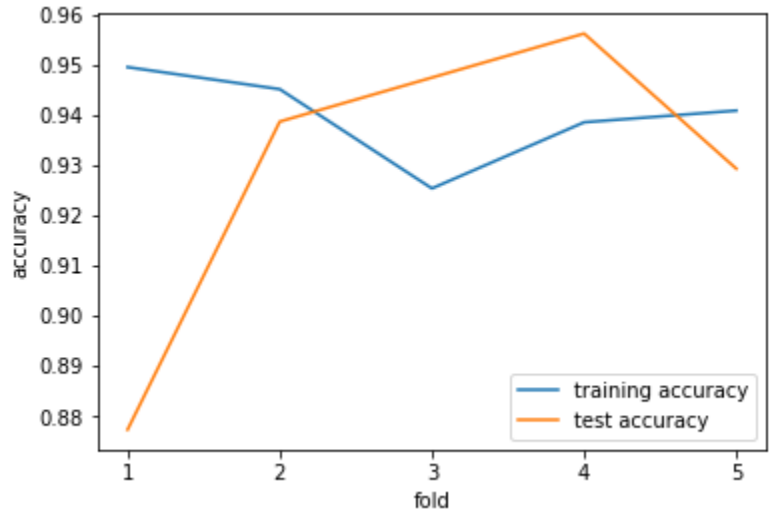


4: Model Development and Training



Looking at the relationship between the accuracy and number of neighbors, 11 seems to be about where you get the best results. After that, the model starts having diminishing returns and loss testing accuracy.

5: k-NN with Cross Validation



	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Mean
Training Accuracy	.949	.945	.925	.938	.940	.940
Test Accuracy	.877	.939	.947	.956	.929	.930

The model from step 4 does not seem to be cross validated. This is evident due to the fact that the mean training accuracy is higher than the mean test accuracy, however, at 11 neighbors for step 4 (the number of neighbors used for step 5), the test accuracy is higher than the training accuracy, which only happened at 2 of the 5 folds.

diagnosis	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension
M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744
M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883

stderr radius	stderr texture	stderr perimeter	stderr area	stderr smoothness	stderr compactness	stderr concavity	stderr concave points	stderr symmetry	stderr fractal dimension
1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193
0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186	0.0134	0.01389	0.003532
0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832	0.02058	0.0225	0.004571
0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661	0.01867	0.05963	0.009208
0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688	0.01885	0.01756	0.005115

largest radius	largest texture	largest perimeter	largest area	largest smoothness	largest compactness	largest concavity	largest concave points	largest symmetry	largest fractal dimension
25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601	0.1189
24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275	0.08902
23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613	0.08758
14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.173
22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364	0.07678