Pratishrut Kamal
Data Scientist Candidate
Take-Home Submission Report
Due: 6/28/2022

Classification of Malignant/Benign Tumors Based on Numerical Data from CT Scans

Problem: Creating a machine learning model that can classify whether, based on numerical data from CT Scans of tumors, whether the Tumor is a Malignant or Benign Tumor

The two models that were created for this had a high variance dataset and low bias as we are only dealing with numbers

1.Data Accuracy

1.1 Problem: Unneeded/Repetitive/Biased Data

- fractal_dimension_mean
- radius_se
- texture_se
- perimeter_se
- area_se
- smoothness_se
- compactness_se
- concaviy_se
- concave points_se
- symmetry_se
- fractal_dimension_se
- radius_worst
- texture_worst
- perimeter_worst
- area_worst
- smoothness_worst
- compactness_worst
- concavity_worst
- concave points_worst
- symmetry_worst
- fractal_dimension_worst

texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
17.33	184.60	2019.0	0.16220	0.66560	0.7119	0.2654	0.4601	0.11890
23.41	158.80	1956.0	0.12380	0.18660	0.2416	0.1860	0.2750	0.08902
25.53	152.50	1709.0	0.14440	0.42450	0.4504	0.2430	0.3613	0.08758
26.50	98.87	567.7	0.20980	0.86630	0.6869	0.2575	0.6638	0.17300
16.67	152.20	1575.0	0.13740	0.20500	0.4000	0.1625	0.2364	0.07678
...	...	...	...	...	...	...	...	...
26.40	166.10	2027.0	0.14100	0.21130	0.4107	0.2216	0.2060	0.07115
38.25	155.00	1731.0	0.11660	0.19220	0.3215	0.1628	0.2572	0.06637
34.12	126.70	1124.0	0.11390	0.30940	0.3403	0.1418	0.2218	0.07820
39.42	184.60	1821.0	0.16500	0.86810	0.9387	0.2650	0.4087	0.12400

Another reason for this decision was to increase the accuracy by including less variables in the calculation and to increase the speed that people could enter these metrics because there would be less metrics to enter and measure in a real medical setting

1.2 Resolving the Issue

- Removing these data points from the training and testing data

Pratishrut Kamal
Data Scientist Candidate
Take-Home Submission Report
Due: 6/28/2022

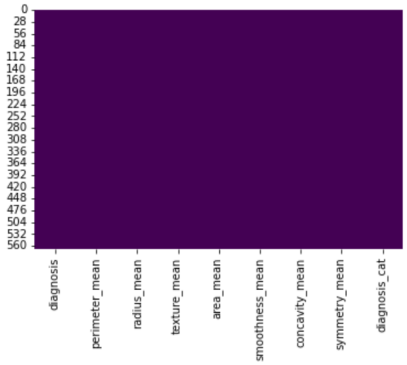
2.Data Completeness

2.1 Problem:

Less than 50% of the data provided to us was left unused because of biased, unneeded, or repetitive, one positive is that none of the data we needed had null values

```
sns.heatmap(df.isnull(),cbar=False,cmap='viridis')
```

```
<AxesSubplot:>
```



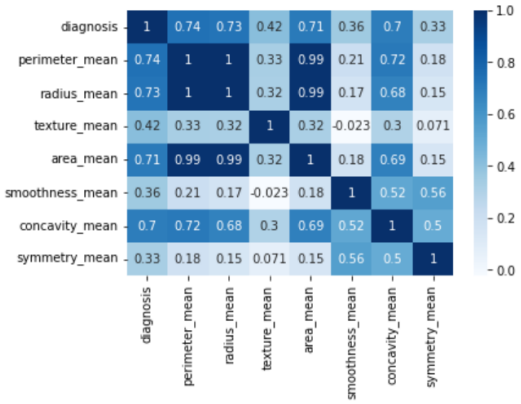
2.2 Resolving the Issue

- Making sure that when we record data is that we only have

3.Multicollinearity of Predictors

3.1 Problem: Observed multicollinearity among the data that was used to train the dataset:

- perimeter_mean
- radius_mean
- texture_mean
- area_mean
- smoothness_mean
- concavity_mean
- symmetry_mean



Variance Inflation Factor for each Category

perimeter_mean -> 26.397601885840036
radius_mean -> 24.923041341095328
texture_mean -> 22.27751654323873
area_mean -> 53.68920510079189
smoothness_mean -> 14.582527045706449
concavity_mean -> 89.69632773970461
symmetry_mean -> 15.119175980629615