Methods

Because the data I collected does not provide enough information, and I would need NLP to proceed my goals which I am half the way. I used sklearn built in breast cancer data. First, the Heatmap shows how columns are related and distributed. Then I plot a pair plot to help understanding the detail comparing to the heatmap.

The attribute target is used as y dataset for training, because it is the label. The x data would be the breast cancer data without the attribute target. Then the datasets are separated to test and train for prediction.

After predicting, the confusion matrix shows how both test and train are distributed compared to the prediction results. Then the decision tree at the end shows classification.

Class distribution

By looking at the heatmap and the pair plot, the worst breast cancer cases are most correlated to the mean of texture, area and perimeter.

Baseline model for comparison

ACCURACY: 0.9035087719298246

NEGATIVE RECALL (Y=0): 0.8805970149253731

NEGATIVE PRECISION (Y=0): 0.9516129032258065

POSITIVE RECALL (Y=1): 0.8805970149253731

POSITIVE PRECISION (Y=1): 0.9516129032258065

Final results

The accuracy is high since the data is not imbalanced. The recall score shows the true positive out of actual positive variables. The precision score shows the correctness of precisions. The prediction is accurate!

Conclusions:

The figure shows that breast cancer has a high risk to be the worst, because it is correlated to the mean of parameters. And the errors on the diagnose is mostly error.