**Diagnosis of Breast cancer using correlated variables**

Ryan Pupia, Syedosama Jaffery, Victor Armani, & MATTHEW FERNANDES.

Minor in Data Science

219:220 Fundamentals of Data Visualization in R: Final project

Prof Bruno Richard, PhD

4/29/2022

Abstract

The purpose of this assignment was to find the best variables to predict whether or not a none biopsied breast tissue mass was malignant or benign using only volumetric data of said cells nuclei.We hypothesize that using the data from our csv with a spearhead method will lead us to three correletory values to diagnosis if a cell is malignant or benign. From this we created a correlation plot. With this correlation plot created we found that our best predictors were more than likely going to be. area\_mean(this is the average surface area of the nuclei in the cell), texture\_worst(standard deviation of gray-scale values, this greyscale is the one viewed when perceiving through a FNA.), smoothness\_worst(local variation in radius lengths, these radiuses are measured from the center of the nucleus to its outer perimeter.) Worst signifies the largest values of said entry EX. smoothness\_worst is the three largest local variations in radius lengths averaged.

Method

The data obtained in the data set was accumulated by the university of wisconsin, the data set contains 570 patients, who were imaged using a Fine Needle Aspirator. Said tool gave all the values used in the dataset.

With a full understanding of our data we created our first correlation plot to find the

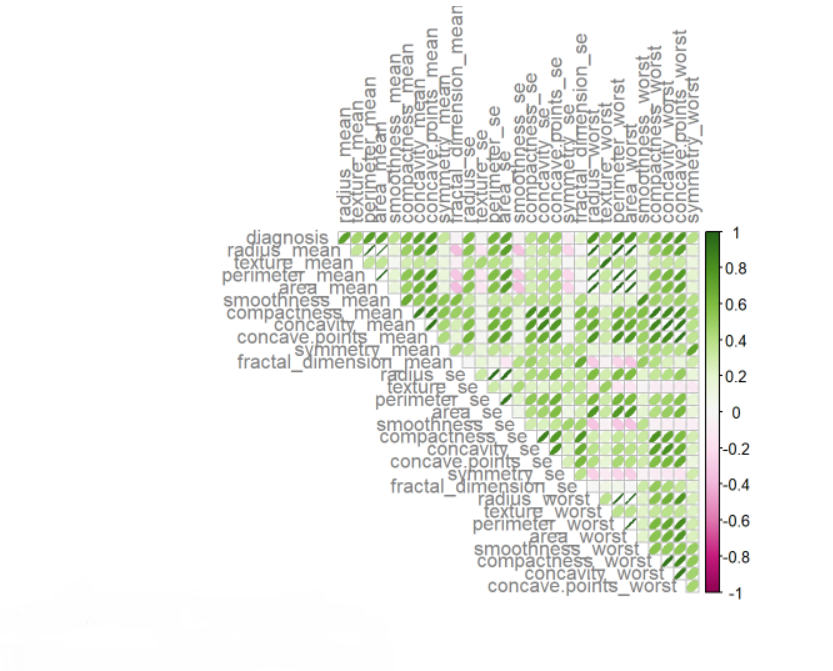
variables that hold the greatest correlation to malignancy in cells as well as least correlation amongst the other variables in the data set. as stated above these variables were Area\_mean, Smoothness\_worst,Texture\_worst.

We then followed this up by creating three histograms that showed the distribution of these variables. This was done to confirm two things. We wanted to make pre-inferences to see if these values would make good predictors, and to determine which analysis methods would be best for these values. We wanted to see if these values were going to be normally distributed. When looking at the histograms it becomes clear that they violate normality. Because our variables also violate normality, we decided to test their statistical difference by use of the Wilcoxon test and achieved p values < .001 for all. This proves that they are all statistically significantly different.

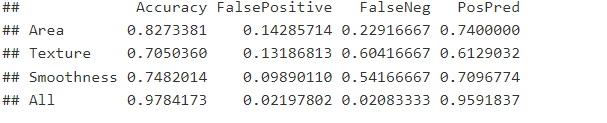
We then created a logistic regression model, and found the slopes, Coefficient of determination, and accuracy which will shown and explained in the interpretation section

Models and interpretations

When looking at the plot it's fairly easy to tell which are most positively correlated with diagnosis, which are radius\_worst,area\_worst,perimeter\_worst. But because these values share strong correlation with other values, we found that it was more than likely best to not test these values. We decided to use area mean because area mean and diagnosis seem to be strongly correlated while also being much less correlatory to the other values in the data set. This same logic goes for our other two predictors which we chose as texture\_worst, smoothness\_worst. While we picked these were picked for our original model this does not suggest that this will provide us with the best model.

If we look at the log odds for each of them also we see a great correlation with area\_mean having 1 to 1.02 ratio with every increase of 1 unit the chances of malignancy increase by 1.02 times. This is extremely significant as the area mean is a massive number ranging from the high hundreds to low thousands making an incremental change of 1 relatively small making the multiplier more meaningful. The same can be said for texture worst with its ratio being 1 to 1.3. With every unit increase we see about a 1.3 times increase in malignancy probability, and while the range is exponentially smaller than area mean ranging from around 10-50 units the multiplier between 1.02 and 1.3 is 15 times. So for every increase in 1 unit in texture the worst is 15x more meaningful towards malignancy then compared to area mean. The ratio for smoothness worst is 1 to The reason this number is so astronomically large is because the range for smoothness\_worst is only .1 to around .2 meaning that an incremental increase of 1 would boost the probability of malignancy by almost an infinite amount, but even when proportioning to 1/100 of a unit increment the probability increase of malignancy for that would be still extremely significant .

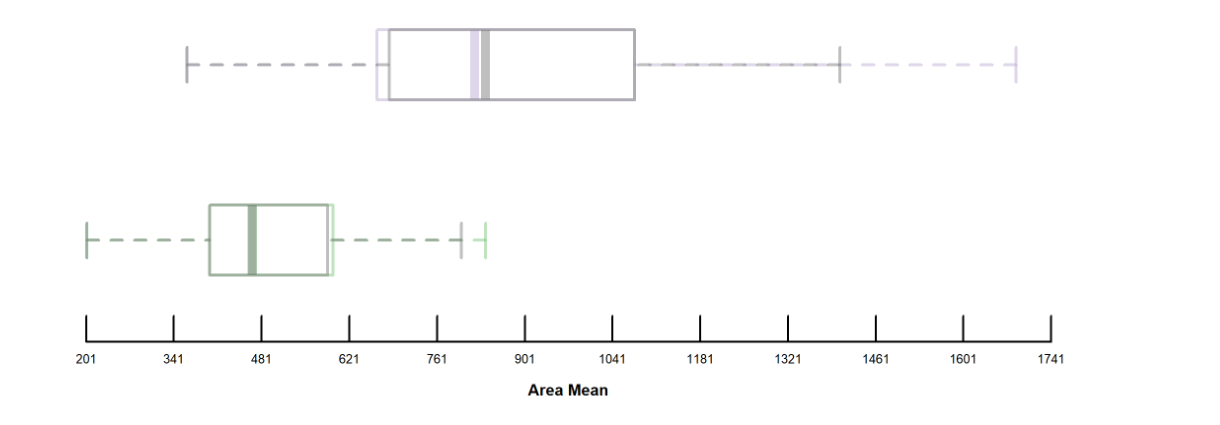
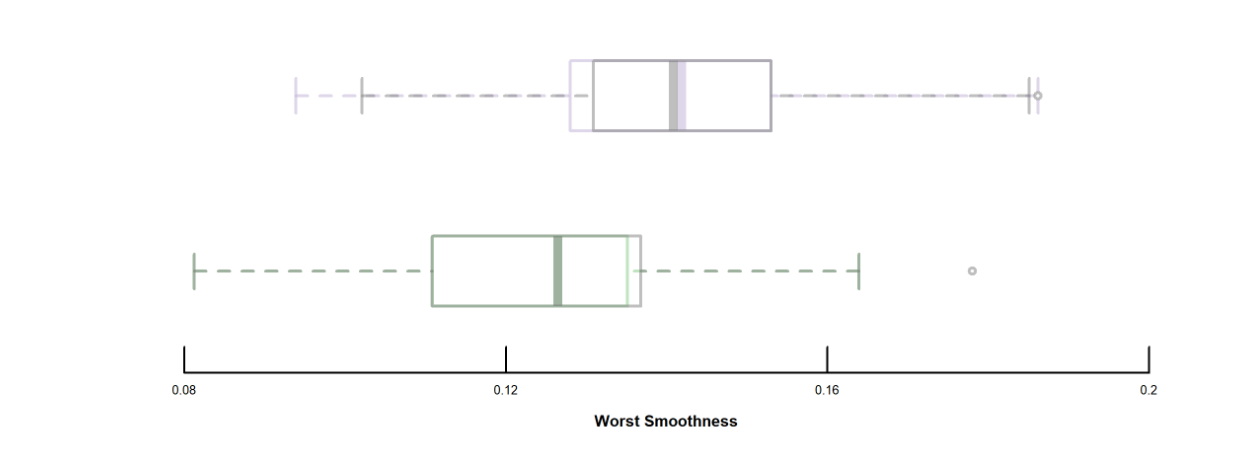
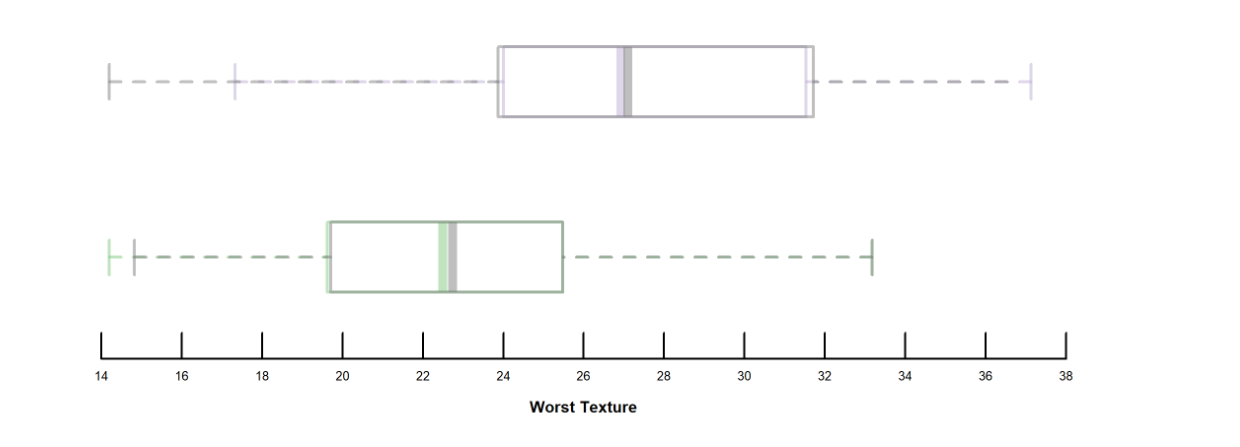
To our surprise our model predicted some relatively very accurate results. With an accuracy of 93.5% we've already accomplished our original hypothesis and then some. With this model we can see that we predict cancer 93.5% of the time, while “only” misdiagnosing a person with cancer 10% of the time. And we only misdiagnose a healthy person 1 in 20 times which correlates to our false pos value .Now after looking at these results we can see that we are letting one in ten people die of cancer because of our models' relatively poor performance. Because of this we created simpler and more accurate models to see if we could make anything better.



So we originally thought to create a model with only one predictor to see how good our predictors themselves were alone. Alone as guessed they're all poor and essentially useless none of them besides area even had an accuracy above 80% besides Area\_mean. But what was very interesting is when we put every predictor into the model. We got a 97.8% accuracy, a 2 percent chance of misdiagnosing cancer whether it be false positive or a false negative. Which means that our model is accurately predicting someone with cancer 96% of the time.

Understandably this model is much too broad as it uses 30 predictors and wouldn't be viable in most applications especially because we added some 20+ variables for a barely 4% increase in accuracy and positive prediction. That's not to say it's meaningless. Our false negatives are 5x times better which means we're only sending people with cancer home 1 in 50 times instead of 1 in 10.

But what if we had an AI who could compile the data alone from the digitized images from the FNA(Fine needle Aspirator) and could within seconds generate meaningful values and diagnosis’. This would make this model very useful, in doctors offices for quick testing as the FNA can be done in patient, and tested in office by the AI before being sent to the Pathologist for confirmation on its diagnosis.this can be an extreme stress reducer for patients who have to wait a week or more to hear back from a pathologist, they can receive instant preliminary results within an hour.

To visualize our model's distribution against the true distribution we made box and whisker plots. In which we see the true distribution of malignancy as the purple graph on top and green as benign. If our model's distribution is accurate we expect to see almost perfect overlap, For our three predictors graphs. As we can see the overlap on our three predictors is near perfect when predicting benign but can sometimes mildly struggle when predicting malignancy. This is to be expected as our model wasn't 100% perfect. There are some outliers made by our model as seen by the malignant worst texture, and area mean, but the overlap is near perfect giving even more credence to the fact that our model is accurate.

Conclusion

In conclusion from our interpretation, we can say that our predictors are good but this model is more or less useless in the real world. But, perhaps with the advent of better technology an AI can self analyze the sample from the FNA and give accurate measurements so that it can also in turn give a pre-diagnosis for breast cancer patients. This can alleviate patient anxiety and give pathologists and general doctors a starting point for treatment. With that being said with more data we can also train an AI to accurately predict the stage of the cancer as well as its type ie.(ductal carcinoma in situ, invasive ductal carcinoma, inflammatory breast cancer, and metastatic breast cancer).