Marc Riley
DSC 680 Predictive Analytics
Breast Cancer Detection
February 6th, 2022

Contents

Business Problem

Brest cancer affects millions of women in the United States ever year. If found in its early stages and is isolated in one location (the breast) the 5-year survival rate is 99% (ASCO 2021). Once it has spread the survival rate ranges from 85%-28% that is why it is so important to accurately and quickly diagnosis a tumor or tissue as malignant or benign

Background

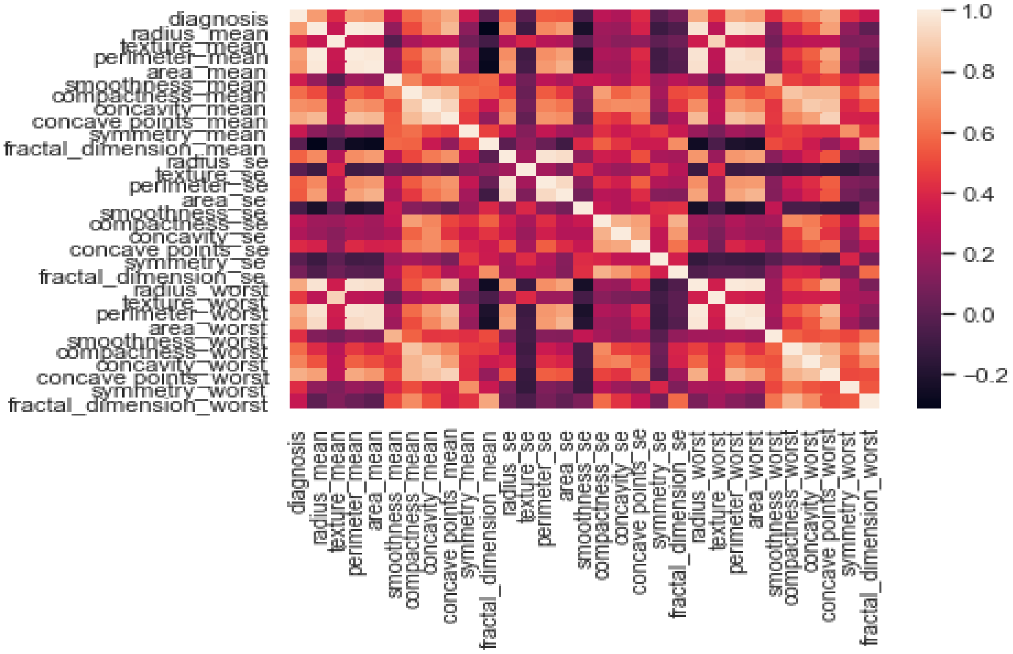
Breast cancer is the growth of cells that cannot be controlled. After time this growth will continue to multiple until a tumor is formed. This tumor can either be benign (not dangerous) or malignant (cancerous). A malignant tumor left untreated or undetected can be potentially fatal. (Smith, 2026) Each year nearly 2 million women are diagnosed with breast cancer every year (Lancer, 2018). Since an inaccurate diagnosis can be fatal it is important that once a test is conducted an accurate result is given. Machine learning can look at a group of variables and determine if a tumor is malignant or benign.

Data

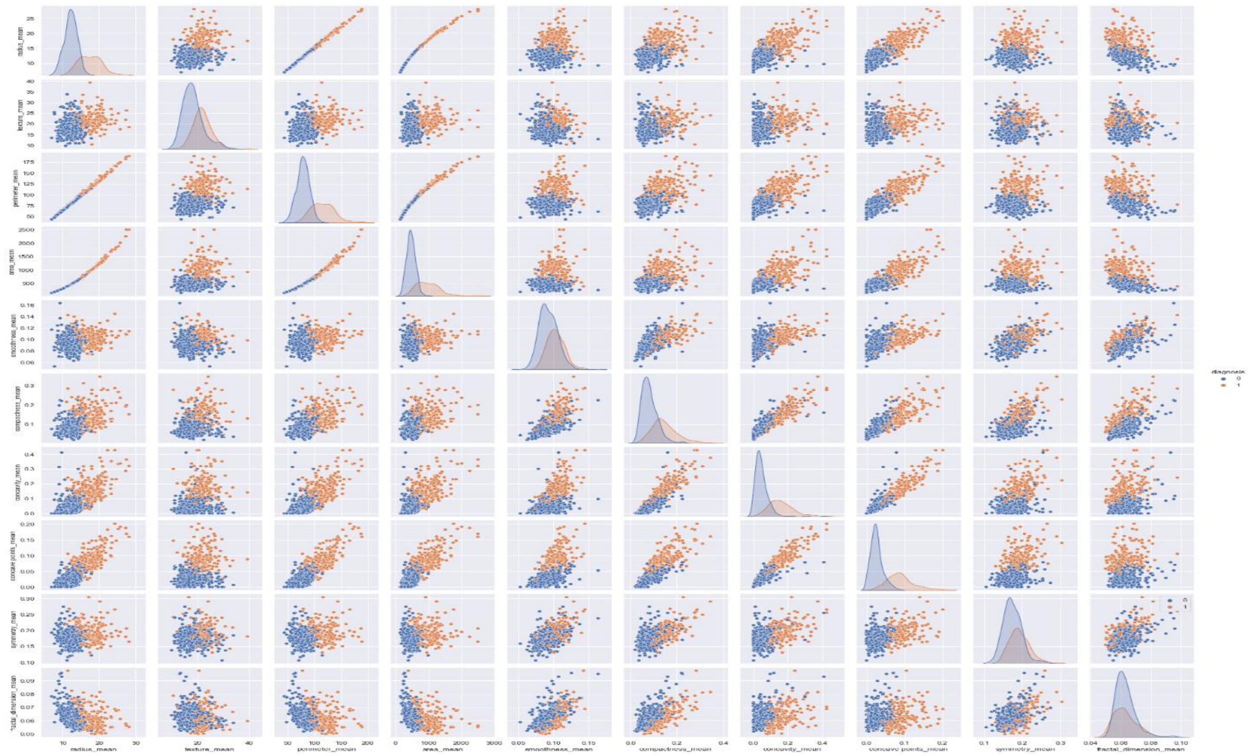
The data provided by Kaggle was originally obtained by patients in Wisconsin. A digitalized picture is provided of the fine needle aspirate of the concerned breast mass. There are roughly 600 images captured in this data set. Each image contains 32 variables including symmetry, concave points, area, texture, and smoothness. The target variable is the diagnosis

EDA

The following steps were taken to explore and analyze the data set. The data was checked for missing values one column (“Unnamed: 32”) was found to have 569 missing values. This column along with “id” were removed from the data frame. Next thing conducted was a correlation test



As you can see some of these variables have high correlation with each other this has the possibility of multicollinearity. Sever multicollinearity has the potential of making the estimates sensitive to minor changes. As of now I will leave these variables alone. The column “diagnosis” was split to see how many malignant vs benign cases are in the data set (212/357). Outliers were examined and left in since cancer cells can take many different shapes and sizes. Finally the variables were all plotted so we can see the difference between benign and malignant

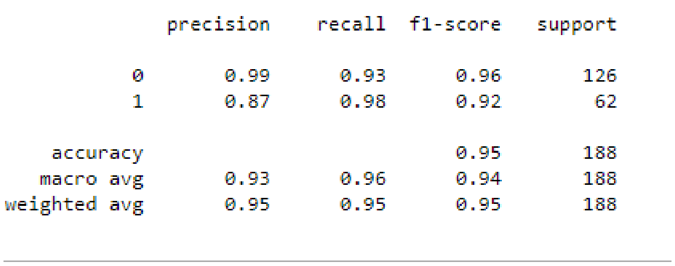


Methods

The main technique used to detect cancerous cells will be K-Nearest Neighbor. This method operates by grouping different data points together that are similar. Once the algorithm is trained a new input will be placed into a group. The group will hopefully tell us if the mass is malignant or benign. A random forest model will also be implemented to compare results.

Analysis

The Results for the K-Nearest Neighbor model are below



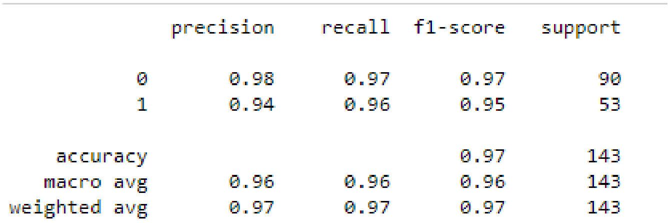
```
              precision    recall  f1-score   support

           0       0.99      0.93      0.96       126
           1       0.87      0.98      0.92        62

    accuracy                           0.95       188
   macro avg       0.93      0.96      0.94       188
weighted avg       0.95      0.95      0.95       188
```

These results are very promising. The precision and recall are both around 90-99 percent. Step two is to build a random forest model and see if the results are comparable or even better. The accuracy is at roughly 94% which is good but when it comes to determining if someone has cancer the more accurate the better.

The second model tested was a Random Forest model these work by deploying many decision trees and combine the results. They are very versatile but can sometimes be hard to interpret and take more computation than a regular decision tree. When ran on our breast cancer data set it performed slightly better than the KNN model. Although 2-6% across all areas doesn't seem like much it can make a big difference especially when it comes to accurately predicting breast cancer.



```
              precision    recall  f1-score   support

           0       0.98      0.97      0.97        90
           1       0.94      0.96      0.95        53

    accuracy                           0.97       143
   macro avg       0.96      0.96      0.96       143
weighted avg       0.97      0.97      0.97       143
```

Conclusion

Both models performed very well across the board with the random forest doing slightly better.

Assumptions

Some assumptions that have been made for this model include all test results have these 32 variables needed to determine if a tumor is cancerous. This data set is representative of all women.

Challenges

The first challenge was seeing what variables were needed. The first test run of the model produced an accuracy of 68%. That was due to the id variable being left in. The second challenge is to try and get the precision increased for malignant tumors. This challenge was fixed by deploying a different model.

Future Applications

This model could be applied to different types of cancer not just breast.

Ethical Concerns

Some ethical considerations could be if the algorithm creates clusters based heavily off demographic information it might negatively impact a certain group of individuals with false positives/negatives. The machine is trained off women in Wisconsin the sample might be bias due to the location and ethnicity restrictions. A positive sample might look different to a woman in Sudan than it does to woman in Wisconsin. Finally in order to get this data a specific type of biopsy needs to be conducted. Some groups of people might not have access to this technology.

References

Cancer statistics - facts on cancer. Paul & Perkins. (2013). Retrieved February 6, 2022, from https://paulandperkins.com/cancer-statistics/

https://www.kaggle.com/gargvg/breast-cancer-eda-classification/data

Lancet. Author manuscript; available in PMC 2018 Oct 16.

*Published in final edited form as:*Lancet. 2017 Feb 25; 389(10071): 847–860.

Published online 2016 Nov 1. doi: 10.1016/S0140-6736(16)31392-7

Smith. (2016, August 21). *How does breast cancer get misdiagnosed or go undetected?* Shevlin Smith. Retrieved January 27, 2022, from https://www.shevlinsmith.com/blog/2016/august/how-does-breast-cancer-get-misdiagnosed-or-go-un/