Diagnosis of Breast Cancer

Zack DeNoto

Bellevue University

DSC 680: Applied Data Science

Dr. Fadi Alsaleem

May 2, 2021

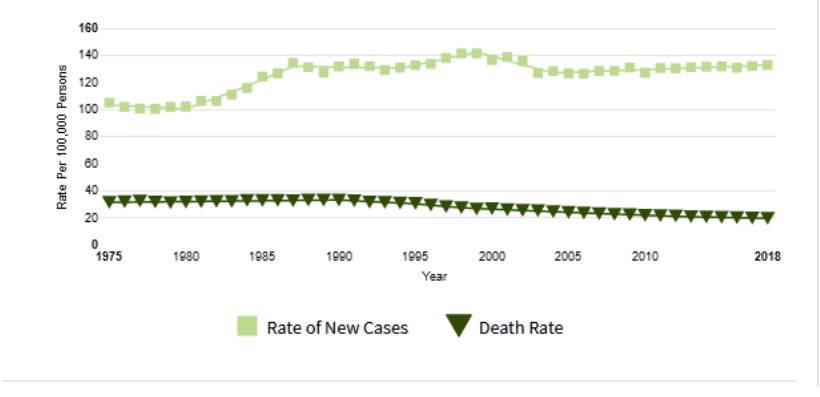
Abstract

The problem that this paper addresses is determining if you can diagnosis if an individual's breast cancer is benign or malignant based on the various factors of the cell's nucleus. This paper will demonstrate if it is possible to make such as diagnosis based on several factors provided in the dataset, as well as determining what factor(s) most affect the diagnosis. This paper will also look at how the different factors are connected and determine which factors have the most correlations to each other. The objective of this project was to find a good dataset, clean the data, determine what factors affect the diagnosis the most, and create the most accurate model(s) possible.

Intro

Every year millions of individuals in the world are diagnosed and killed by cancer. In 2020, there were an estimated 1,800,000 new cases diagnosed with an estimated over 600,000 dead from cancer in the US (Cancer.gov, 2021). Cancer is the second leading cause of death in the US and 1 in 4 deaths in the US is due to cancer (Gis.cdv.gov, 2021). The most common cancer is breast cancer and about 1 in 8 women in the US will develop invasive breast cancer over the course of her lifetime (Breastcancer.org, 2021). Globally, there were 2,300,000 women diagnosed with breast cancer and 685,000 deaths. Though breast cancer is the most common cancer, it is also the most survivable cancer with over 7,800,000 women alive in 2020 who were diagnosed with breast cancer in the past 5 years (WHO.com, 2021). Breast cancer mortality was higher from the 1930's through the 1970's with improvements in survival beginning in the 1980's, and today we have the lowest mortality rate we have ever had, as seen in table 1 (Seer.cancer.gov, 2021).

Table 1 l Chart of rate of new breast cancer cases and deaths from breast cancer.



There are several factors which can increase the chance of being diagnosed with breast cancer, such as age. Breast cancer risk doubles each decade until menopause. After that it slows down, but it is still more common after menopause (Wcrf.org, 2021). After being diagnosed with breast cancer, weight plays a large factor in being diagnosed with a second primary cancer. A study was done linking excess weight and cancer showing that as a woman's BMI increased, so did her risk of developing a second primary cancer. The analysis showed that for every 5 kg (about 11 pounds) increase in weight, a woman's risk of a second primary cancer increased. A higher BMI resulted in a 7% greater risk of primary cancer, a 13 % greater risk of a second primary obesity-related cancer, a 11% greater risk of a second primary breast cancer, and a 15% greater risk of a second primary estrogen-receptor-positive breast cancer (Breastcancer.org, 2021).

Data

The dataset used for this analysis was called Breast Cancer Wisconsin (Diagnostic) Data Set from Archive.ics.uci.edu, retrieved from the link https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29. This dataset contains information on images of a fine needle aspirate (FNA) of a breast mass. The characteristics of the cell nuclei image are the fields in the dataset. The csv file had 33 columns with fields such as radius, texture, symmetry, compactness, fractal dimension, perimeter, concave points, and area. Each of the features has a mean, standard error and "worst" or largest value recorded with four significant digits.

Methodology

The first step that needed to be done was to examine and clean the dataset. After using R to explore the dataset, it seemed a blank column with the field header "X" was in the dataset unnecessarily and was deleted. After the column was deleted, I looked for any missing values or values that seemed like they could be incorrect. The dataset contained no missing or null values, but did contain values that were 0. At first glance this may be concerning as the rest of the values are greater than 0. However, due

to the nature of the factors being for the breast cancer cell nuclei, if a diagnosis was benign, this potentially could cause the value to truly be 0, and so was left in.

Using histograms to view every factor of the dataset, it was clear there were no outliers and most factors seemed to have a fairly normal distribution. This meant that no outliers needed to be removed from the dataset. There was very minimal data cleaning that needed to happen with this smaller dataset as all the values were quantitative and uniform, leading to the focus of the project being on the modeling.

Once the data was cleaned up, I looked at the correlation between fields compared to the diagnosis. After correlations were looked at, the first goal was to create models for determining the diagnosis to see if I could accurately predict if the diagnosis was benign or malignant. Lazy Predict was used to get a better understanding of what models would be likely to be more accurate. Then I tested several of the models Lazy Predict showed such as K-Nearest Neighbor, Stochastic Gradient Descent (SGD), Logistic Regression, XGBoost and Random Forest. I then repeated this process on several of the other fields such as area, perimeter, radius, and texture. Lastly, I used Keras deep learning to predict the diagnosis as well.

Results

As mentioned in the first methodology section, the first step was to use Lazy Predict to get a better sense of what models to then test out. As seen in the figure below, it seemed like there were many accurate models for determining diagnosis.

Figure 1 l An Image from Lazy Predict showing preliminary model results.

```
clf = LazyClassifier(verbose=0,ignore_warnings=True)
models, predictions = clf.fit(X_train, X_test, y_train, y_test)
print(models)
```

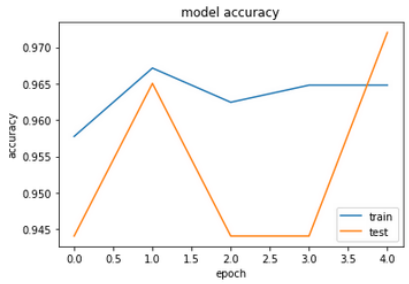
```
100%|██████████| 29/29 [00:01<00:00, 17.15it/s]
```

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score
SVC	0.99	0.99	0.99	0.99
PassiveAggressiveClassifier	0.98	0.98	0.98	0.98
KNeighborsClassifier	0.97	0.97	0.97	0.97
SGDClassifier	0.97	0.97	0.97	0.97
Perceptron	0.97	0.96	0.96	0.96
LogisticRegression	0.97	0.96	0.96	0.96
CalibratedClassifierCV	0.97	0.96	0.96	0.96
LinearSVC	0.95	0.95	0.95	0.95
QuadraticDiscriminantAnalysis	0.95	0.95	0.95	0.95
AdaBoostClassifier	0.95	0.95	0.95	0.95
LGBMClassifier	0.95	0.95	0.95	0.95
XGBClassifier	0.94	0.94	0.94	0.94
ExtraTreesClassifier	0.94	0.93	0.93	0.94
RandomForestClassifier	0.94	0.93	0.93	0.94
RidgeClassifierCV	0.94	0.93	0.93	0.94
RidgeClassifier	0.94	0.93	0.93	0.94
LabelSpreading	0.94	0.93	0.93	0.94
LabelPropagation	0.94	0.93	0.93	0.94
LinearDiscriminantAnalysis	0.93	0.92	0.92	0.93
GaussianNB	0.92	0.92	0.92	0.92
BaggingClassifier	0.92	0.91	0.91	0.92
ExtraTreeClassifier	0.91	0.91	0.91	0.91
DecisionTreeClassifier	0.92	0.91	0.91	0.92
NuSVC	0.92	0.90	0.90	0.91
BernoulliNB	0.90	0.90	0.90	0.90
NearestCentroid	0.91	0.89	0.89	0.91
DummyClassifier	0.45	0.42	0.42	0.44

The accuracies were very high, with most models being above 90% accurate. However, Lazy Predict is just a quick tool to get an idea of how accurate models may be, not how accurate they truly are. I then modeled for diagnosis with Random Forest, Logistic Regression, XGBoost, Stochastic Gradient Descent (SGD), and K-Nearest Neighbor. The accuracies for the models were 95.8%, 96.5%, 97.2%, 67.8%, and 94.4%, respectively. Using Keras for deep learning, I then tested the model for predicting the diagnosis as well. The accuracy of the Keras machine learning was 97.2%, which was in line with many of the tested models as seen in the figure below.

Figure 2 l An Image from Keras machine learning and the corresponding accuracy graph.

```
569/569 [==============================] - 0s 121us/step
Accuracy: 96.31
Train on 426 samples, validate on 143 samples
Epoch 1/5
426/426 [==============================] - 0s 106us/step - loss: 0.1029 - accuracy: 0.9
577 - val_loss: 0.1040 - val_accuracy: 0.9441
Epoch 2/5
426/426 [==============================] - 0s 129us/step - loss: 0.0856 - accuracy: 0.9
671 - val_loss: 0.0812 - val_accuracy: 0.9650
Epoch 3/5
426/426 [==============================] - 0s 106us/step - loss: 0.0861 - accuracy: 0.9
624 - val_loss: 0.1156 - val_accuracy: 0.9441
Epoch 4/5
426/426 [==============================] - 0s 103us/step - loss: 0.0917 - accuracy: 0.9
648 - val_loss: 0.0958 - val_accuracy: 0.9441
Epoch 5/5
426/426 [==============================] - 0s 136us/step - loss: 0.1075 - accuracy: 0.9
648 - val_loss: 0.0773 - val_accuracy: 0.9720
```

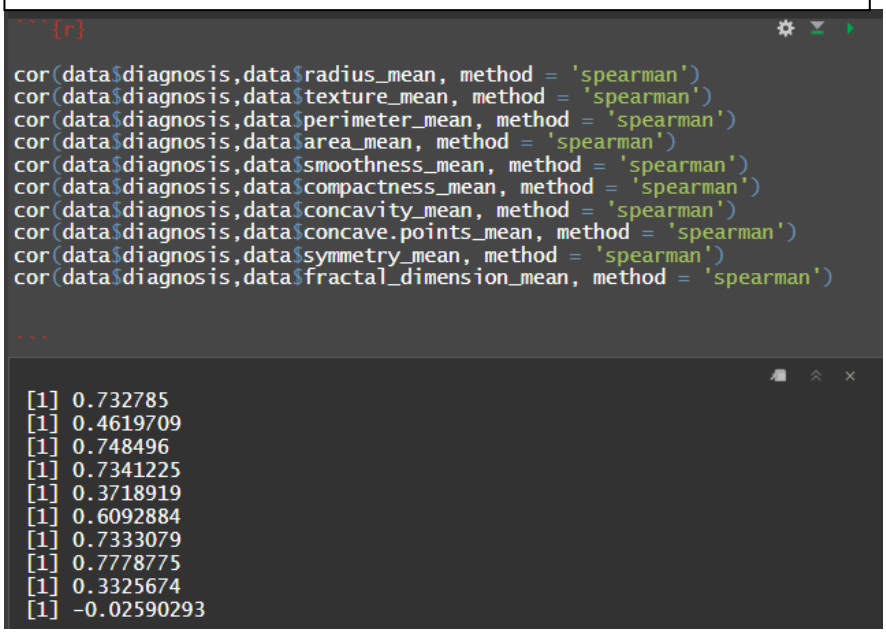


These results showed that you cannot always take the results of Lazy Predict to be accurate, as SGD showed an accuracy of 97% but only had 67.8% accuracy tested separately. Also, these results show that based on the ten factors of the breast cancer mass cell, one can predict if the diagnosis will be benign or malignant with high accuracy.

Next, I wanted to see if it was possible to accurately predict some of the factors such as area, perimeter, radius, and texture. For every factor tested, I had to create new variables for the train and test by isolating the factor tested and running the same five models for each of the factors. Radius was

the factor with the highest accuracies of 83.2%, 33.6%, 91.6%, 16.8%, and 79.7% for Random Forest, Logistic Regression, XGBoost, Stochastic Gradient Descent (SGD), and K-Nearest Neighbor respectively. Texture was the factor with the second highest accuracies of 20.3%, 11.9%, 26.6%, 5.6%, and 7.7%. Perimeter was the factor with the third highest accuracies of 22.4%, 4.2%, 23.8%, 1.4%, and 16.1%. Lastly was area with accuracies of 0.7%, 0%, 1.4%, 0%, and 0%. The results of the accuracies for the factors indicate that some factors such as radius or texture are much easier to predict knowing the other factors, whereas perimeter and area are much harder to predict if you know the other factors. When looking at the correlations between all ten factors and the diagnosis, perimeter had a higher correlation with diagnosis compared to radius as seen in the figure below.

Figure 3 l An Image from R comparing correlations of dataset factors to the diagnosis variable.



Texture had a very low correlation of 0.462, which was unexpected because of its' higher accuracies using the five models, compared to perimeter with a correlation of 0.748 and area that had correlation of 0.734, both of which had lower accuracies.

Conclusion

Before this analysis on breast cancer, I did not know what to expect on whether or not the diagnosis would be able to be predicted based on factors of the cell nuclei. Cancer is a complicated

subject and I thought low accuracies would be achieved for modeling based on the ten factors of the cell mass. I was pleasantly surprised to get high accuracies in predicting the diagnosis and see that some of the other factors could be predicted accurately as well. This analysis shows that there are strong relationships between the factors in the nucleus of the breast cancer cell and the diagnosis of benign or malignant. It gives me hope that these relationships can be used in the future to help make advancements in breast cancer treatment. Advancements in breast cancer treatment and understanding come from conducting studies researching various impacts on diagnosing breast cancer. One study in particular looked at the hormonal factors in breast cancer and looked at the relation between risk of breast cancer and use of hormone replacement therapy (HRT). In that study it was found that the risk of having breast cancer diagnosed is increased in women using HRT and increased with the increasing duration of use (Calle, 1997). Another study focused on Vitamin D and breast cancer biomarkers in female patients where there was no statistical difference found between the Vitamin D and placebo on the patients (Wood, 2010).

This analysis, however, looked at the various factors in the nuclei of breast cancer cells. With the Covid-19 vaccine research having so much funding this past year, the technology behind the vaccine, specifically RNA (mRNA) research, could help lead towards a cure for breast cancer many other diseases such cancer, HIV, and more (Terry, 2021). RNA research is looking to create treatments by looking at the diseases at a cellular level to help train the immune system, just like the factors in this analysis. With mRNA research and other analyses like this one, I am hopeful that advances in the treatment of breast cancer and other cancers will continue to progress to help save many lives.

References

Cancer.gov. (2021). Cancer Statistics. Retrieved from https://www.cancer.gov/about-cancer/understanding/statistics

Breastcancer.org. (2021). U.S. Breast Cancer Statistics. Retrieved from https://www.breastcancer.org/symptoms/understand_bc/statistics

WHO.com. (2021). Breast Cancer. Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Wcrf.org. (2021). Breast Cancer How Diet, Nutrition, and Physical Activity Affect Breast Cancer Risk. Retrieved from https://www.wcrf.org/dietandcancer/breast-cancer

Breastcancer.org. (2021). For Women First Diagnosed With Breast Cancer, Risk of Second Cancer Goes Up as Weight Increases. Retrieved from https://www.breastcancer.org/symptoms/understand_bc/statistics

Calle, E. E., & Heath, C. W. (Eds.). (1997, November). Breast cancer and hormone replacement therapy: *Collaborative reanalysis of data from 51 epidemiological studies of 52,705 women with breast cancer and 108,411 women without breast cancer.* https://profiles.wustl.edu/en/publications/breast-cancer-and-hormone-replacement-therapy-collaborative-reana

Wood, E. Marie. (2010, October). Vitamin D and Breast Cancer Biomarkers in Female Patients . https://clinicaltrials.gov/ct2/show/results/NCT01224678?recrs=e&cond=Breast+Cancer&draw=2&rank=10&view=results

Seer.cancer.gov. (2021). Cancer Stat Facts: Female Breast Cancer. Retrieved from https://seer.cancer.gov/statfacts/html/breast.html

Gis.cdc.gov. (2021). Leading Cancer Cases and Deaths, All Races/Ethnicities, Male and Female 2017. Retrieved from https://gis.cdc.gov/Cancer/USCS/DataViz.html

Terry, M. (2021). mRNA Tech Used in COVID-19 Vaccines Could be Used to Cure HIV, Cancer and More. Retrieved from https://www.biospace.com/article/mrna-tech-used-in-covid-19-vaccines-could-be-used-to-cure-hiv-cancer-and-other-diseases/