Using Machine Learning to Predict Breast Cancer Diagnoses

Brandon May

Introduction:

Breast cancer continues to be a significant culprit of morbidity and mortality, even today with our current medical advances. To date, medicine is a relatively untapped area for machine learning and my main goal with this project is to demonstrate that machine learning algorithms could be used to assist in cancer detection as adjunct to physician expertise.

It is no secret that our healthcare is cumbersome, overpriced, and our outcomes are unsatisfactory when compared to other peer countries. In a data-driven world, with the vast amounts of personal health information and data available in electronic health records, the sky is the limit. Could machine learning algorithms be used to improve diagnosis, save lives, and prevent suffering? Up until this point, we have only been limited by data availability. Doctors are fallible just as humans and mistakes and misdiagnoses do happen. There are multiple ways that healthcare can use machine learning from customer service to cancer detection (Sarkar, 2020). Some are using various machine learning algorithms to detect DNA changes that could predict cancer (Ippolito, 2019). Breast cancer detection is yet another potential extension of this methodology. In fact, deep learning is being used to detect breast cancers through mammogram images (Shen *et al.*, 2019).

According to the CDC, breast cancer is the second most common cause of death due to cancer in women (CDC, 2020). Even with modern advances, approximately 39,000 women die from this disease annually (Esserman *et al.*, 2020). Due to increasing adoption of screening mammograms for the appropriate populations, we have reduced mortality from breast cancer by almost 14% (de Gelder *et al.*, 2015). Breast masses are detected using mammography/ultrasound and if suspicious, many women will get a fine needle aspiration or core needle biopsy. This involves using a needle to sample the tissue in question and observing it under a microscope to determine if it has benign or malignant features. Even with a fine needle aspiration, this still cannot detect invasive cancers and further surgical care may be necessary (Joe *et al.*, 2020). Pathology determination can be subjective and one study estimates that sensitivity of approximately 70% in determining tumors and is heavily dependent on training (Ljung *et al.*, 2001). Therefore, the question remains, can machine learning help us predict malignant tumors as an adjunct to physician expertise? Secondary objectives include simplification of measurements and elimination of variables that are redundant and identifying particularly influential variables in the models.

Methods:

The dataset was obtained from the University of California – Irvine Machine Learning Repository and was data was originally collected at the University of Wisconsin in 1995 (Wolberg, M.D. *et al.*¸1995). These datasets have been made available for public use for the purpose of using machine learning algorithms to derive further insights into data science problems. The following links to the dataset directly:

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

This is a CSV file format in which microscopic images of Fine Needle Aspirates (a type of biopsy) of suspicious breast tissue were digitized and analyzed. In these women, they were examining the suspicious masses in those without evidence of metastasis (distant cancer spread) and determined whether these were benign or malignant.

There was a total of 32 variables with 570 different subjects. The target variable is categorical and is coded as M for malignant and B for benign. There are no missing data in the CSV file. There are case identifiers but other demographic information such as age, co-morbidities, and family history are not available.

The data aims to determine if these specific characteristics can be used to predict whether a tumor was malignant or benign. The different variables that are described—in Appendix A at the end of this document—all describe a feature of cell size or cell shape. On a basic level, cancer is detected under a microscope by looking at cell size, shape, as well as various other features of the cell that could be consistent with malignant characteristics. In the example dataset, they used the same ten measurements and computed the mean, standard error, and listed the worst of the values for each of the variables. The values were encoded into Jupyter notebook and the CSV file was loaded as a dataframe. The case identifier variable was dropped from the dataset.

Fundamentally, this is a classification problem in that the purpose is to attempt to classify the data points into either a malignant or benign diagnosis and compare it with the known value. Three different machine learning classification algorithms were used including logistic regression, K nearest neighbors, and Random Forest models.

Results:

Initial analysis of the dataset noted no missing values in any of the variables. The predictor variables were all decimals and measurements based off the digitized images in the dataset. As stated above, the target variable was categorical with M standing for a malignant diagnosis and B standing for a benign diagnosis. Using a correlation cut-off of 0.95, redundant variables were eliminated including the perimeter_mean, area_mean, perimeter_se, area_se, radius_worst, perimeter_worst, and area_worst measurements. This simplified the dataset down to 24 predictors.

Variable distributions were analyzed to see if any variables were skewed. The target class was unbalanced with approximately 150 more benign diagnoses than malignant diagnoses. Given the unbalanced target class, F1, precision, and recall scores were selected to determine model performance since with an unbalanced target classe, accuracy as a metric may be misleading. The only values that were normally distributed were symmetry_mean and fractal_dimension_mean. The rest were positively skewed. To be as accurate as possible since outliers could be indicative of malignancy, the skewed variables were not corrected.

Using scatterplots, the predictor variables were individually plotted against our target variable. While there were outliers, the malignant diagnoses seemed to have higher worst symmetry scores, worst concave points, worst concavity, worst compactness, and worst texture values. Also, the radius of the imaged cells seemed to be larger with malignant tumors. This is in line with known cancer diagnoses in that cells tend to be asymmetric, unusual, large, and abnormal as compared to a standard cell.

Models were run using Logistic Regression, KNN, and Random Forest algorithms. The confusion matrices and classification reports with the three methods are below:

Logistic Regression:

The logistic regression method performed quite well and the best of the three methods tested. The F1 score for predicting benign lesions was 0.98 and the score for predicting malignant lesions was 0.96. There were only 3 misclassifications based on the confusion matrix. In this case, there were two predicted benign lesions that were actually malignant. A false positive diagnosis of thinking that a lesion was malignant when it was actually benign—despite the psychological distress is less concerning—as a missed cancer diagnosis. The false negatives need to be minimized as much as possible.

KNN:

The KNN model performed the worst of the three even after hyperparameter tuning. The best model used uniform weights, 7 number of neighbors, a p of 1, a leaf size of 1, and the minkowski metric. There were 4 misclassifications in this model with 2 false positives and 2 false negatives. The F1-score for benign and malignant lesion predictions were 0.94 and 0.89 respectively.

Random Forest:

The random forest model performed better than the KNN model but worse than the logistic regression model based on the F1 scores. Hyperparameter tuning was not done specifically for this algorithm due to the fact that random forest algorithms optimize the model with training. The F1 scores for predicted benign and malignant lesions were 0.97 and 0.94 respectively. However, there were 5 false negative classifications. This was the worst number of false negatives out of all of the models and the argument could be made that this actually performed the worst if the goal was to minimize false negatives as much as possible.

Which variables were the most important in these models? Since feature importance is not necessarily applicable to KNN machine learning algorithms, using the Python yellowbrick library, feature importances were calculated and ranked in order of relative magnitude and provided fascinating insights.

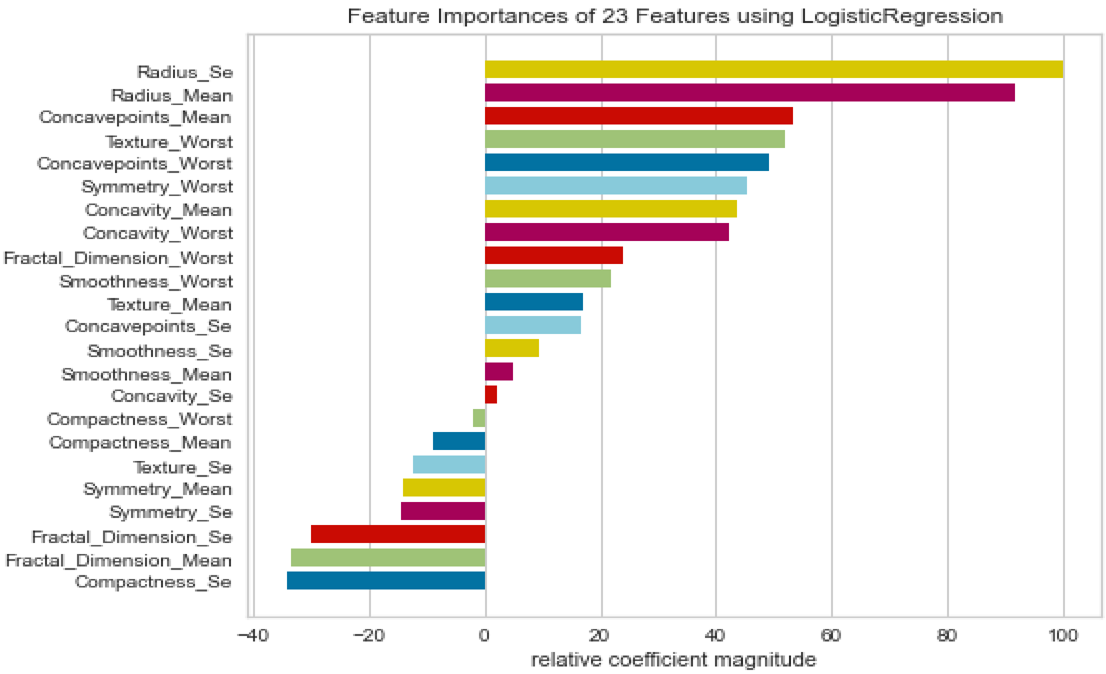

Figure 1. Ranked Feature Importance Based on Coefficients of Logistic Regression Model

As seen above, the variables that had the most relative importance using a logistic regression model were radius_se, radius_mean, concavepoints_mean, texture_worst, concavepoints_worst, symmetry_worst, concavity_worst, and concavity_mean. These are based on the relative coefficients of the model themselves. Logically this makes sense in that the size of the cell signified by radius would help in determining a malignant lesion. Further, many of the variables were of the worst value category. This could suggest that the values that are most abnormal could be associated with predicting malignant or benign lesions.

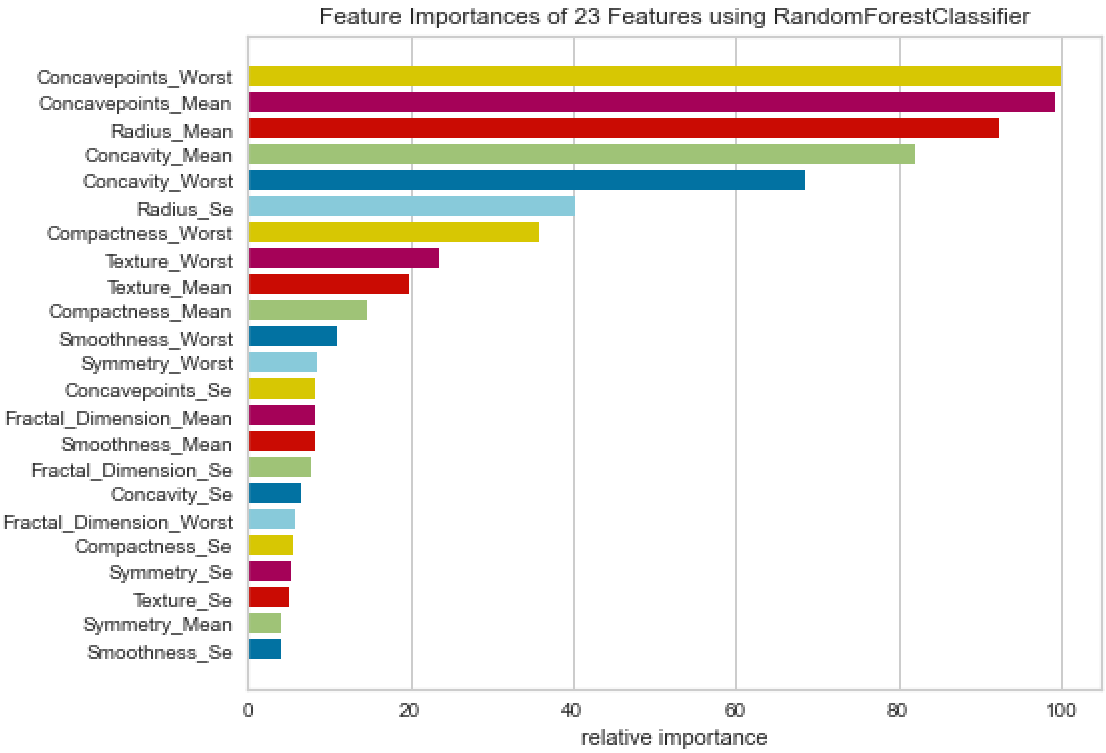

Figure 2: Ranked Feature Importance Based on Coefficients of Random Forest Model

The Random Forest Algorithm above had somewhat different results. In this model, concavepoints_worst, concavepoints_mean, radius_mean, concavity_mean, and concavity_worst were all ranked highly with relative importance. These are similar to the logistic regression model as well. This would seem to suggest that radius_mean, concavepoints_worst, and concavepoints_mean all seem to be particularly important predictors for whichever of the two models you are considering.

Conclusions:

Data science has traditionally been used in corporations and technology-predominant organizations. Healthcare generates a staggering amount of data that can be used in machine learning algorithms. The fact of the matter is that even with excellent care, mistakes and bad outcomes still do occur. Breast cancer is just one of major cause of morbidity and mortality for women worldwide. Even with advanced screening techniques, 39,000 women still die each year from this disease. The diagnosis can still be subjective, and mistakes do happen (both false positives diagnoses and false negative diagnoses). The main question to answer is if healthcare can begin leveraging machine learning techniques to improve outcomes.

There are some limitations to this dataset. First, there are only approximately 500 different samples to train the algorithm. A robust model would need thousands, perhaps even millions of subjects to fine tune its performance. Second, the dataset was constructed from sample images from 1995. This is quite dated. For further implementation of a model to predict malignancy based on images, there would have to be much more updated and realistic data. Besides needing more data points, the fact remains that population health is dynamic and cancer diagnosis and treatment can change drastically and quite rapidly with current research and advancements.

In general, the gold standard for diagnosis of cancer is based on a microscopic examination of the biopsy in question. This is performed by specially trained physicians. Further, cancer is usually detected based on some general guidelines that tumor cells tend to be abnormal in size, abnormally shaped, and can have features inside the cell that can be associated with more aggressive malignancies. The variables that were tested in this dataset were measures of either cell size or cell shape.

While the logistic regression model performed the best of the three models tested based on F1 scores and the low number of false negatives, there were some variables that were important in both this model and the Random Forest model. Radius_mean, concavepoints_mean, and concavepoints_worst were all ranked highly based on relative importance. The logistic regression model seemed to favor the variables that were the worst measurements while the mean and standard error measurements were favored the most. While another model could potentially rank other variables differently, the fact that two of the algorithms were picking variables that are widely known to be helpful in diagnosis malignant cells demonstrates that machine learning could have an impact on diagnosis.

The goal is not to replace doctors; on the contrary, doctors are required to be on the forefront treating patients and are still experts in the field. However, the goal should be to use machine learning models as an adjunct to flag potentially high-risk findings that should be further investigated before disregarding. This could be something as simple as the model indicating that there are several highly suspicious features of malignancy and have a physician review for final diagnosis to either concur or dispute that result. Outcome improvement is paramount; a healthcare system must strive to deliver the best quality care as possible. This is a basic tenet of treating patients and this project suggests that machine learning could potentially help healthcare workers improve outcomes and make patient lives better.

References Cited:

Centers for Disease Control. (2020, June 08). Breast Cancer Statistics. Retrieved September 03, 2020, from https://www.cdc.gov/cancer/breast/statistics/index.htm

De Gelder, R., Heijnsdijk, E. A., Fracheboud, J., Draisma, G., & Koning, H. J. (2014). The effects of population-based mammography screening starting between age 40 and 50 in the presence of adjuvant systemic therapy. *International Journal of Cancer, 137*(1), 165-172. doi:10.1002/ijc.29364

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Esserman, L. J., MD, MBA, & Joe, B. N., MD, PhD. (2019, May 30). Diagnostic evaluation of women with suspected breast cancer. Retrieved September 04, 2020, from https://www.uptodate.com/contents/diagnostic-evaluation-of-women-with-suspected-breast-cancer?search=breast cancer diagnosis

Ippolito, P. (2019, September 12). Understanding Cancer using Machine Learning. Retrieved September 04, 2020, from https://towardsdatascience.com/understanding-cancer-using-machine-learning-84087258ee18

Joe, B. N., MD, PhD, & Esserman, L. J., MD, MBA. (2019, May 16). Breast biopsy. Retrieved September 04, 2020, from https://www.uptodate.com/contents/breast-biopsy?search=fna breast

Ljung, B., Drejet, A., Chiampi, N., Jeffrey, J., Goodson, W. H., Chew, K., . . . Miller, T. R. (2001). Diagnostic accuracy of fine-needle aspiration biopsy is determined by physician training in sampling technique. *Cancer, 93*(4), 263-268. doi:10.1002/cncr.9040

Sarkar, T. (2020, April 30). AI and machine learning for healthcare. Retrieved August 30, 2020, from https://towardsdatascience.com/ai-and-machine-learning-for-healthcare-7a70fb3acb67

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., Mcbride, R., & Sieh, W. (2019). Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Scientific Reports, 9*(1). doi:10.1038/s41598-019-48995-4

Wolberg, W. H., M.D., Street, W. N., Ph.D., Heisey, D. M., Ph.D., & Mangasarian, O. L., Ph.D. (1995). Computerized Breast Cancer Diagnosis and Prognosis From Fine-Needle Aspirates. *Archives of Surgery, 130*(5), 511. doi:10.1001/archsurg.1995.01430050061010

Appendix A - Variable Descriptions:

<u>Variables:</u>

Id: Number signifying unique samples (Integer)

Diagnosis: M for Malignant, B for Benign (Categorical) – the Target Variable

Radius_mean: Mean of distance from center to points on the perimeter of tumor cell, cell size measure (Float)

Texture_mean: Mean of grey-scale values of image (Float)

Perimeter_mean: Expression of both cell size and shape (Float)

Area_mean: Mean area of cell size (Float)

Smoothness: Mean of cell smoothness and shape (Float)

Compactness: Mean of cell compactness and shape (Float)

Concavity_mean: Mean of cell concavity of image (Float)

Concave points_mean: Mean of concave points (Float)

Symmetry_mean: Mean of cell symmetry (Float)

Fractal_dimension_mean: Mean of fractal dimension, measure of cell shape (Float)

Radius_se: Standard error of distance from center to points on the perimeter of tumor cell, cell size measure (Float)

Texture_se: Standard error of grey-scale values of image (Float)

Perimeter_se: Standard error of both cell size and shape (Float)

Area_se: Standard error of cell size area (Float)

Smoothness_se: Standard error of cell smoothness (Float)

Compactness_se: Standard error of cell compactness (Float)

Concavity_se: Standard error of cell concavity of image (Float)

Concave points_se: Standard error of concave points (Float)

Symmetry_se: Standard error of cell symmetry (Float)

Fractal_dimension_se: Standard error of fractal dimension (Float)

Radius_worst: Worst measurement of distance from center to points on the perimeter of tumor cell, cell size measure (Float)

Texture_worst: Worst measurement of grey-scale values of image (Float)

Perimeter_worst: Worst measurement of both cell size and shape (Float)

Area_worst: Worst measurement of cell size area (Float)

Smoothness_worst: Worst measurement of cell smoothness (Float)

Compactness_worst: Worst measurement of cell compactness (Float)

Concavity_worst: Worst measurement of cell concavity of image (Float)

Concave points_worst: Worst measurement of concave points (Float)

Symmetry_worst: Worst measurement of cell symmetry (Float)

Fractal_dimension_worst: Worst measurement of fractal dimension (Float)