Automation of Analysis for Cancer Imaging Screening

Eric Dickey, Emily Phillips, Jesse Zamora

Department of Data Science, Bellevue University

DSC 630: Predictive Analytics

Professor Fadi Alsaleem

March 2, 2022

Executive Summary

Breast cancer is one of the most commonly diagnosed cancers in the world which can carry devastating effects for those who develop it. Fortunately, great enhancements in medicine have approved it's treatability and survival rate. One of those enhancements is detection of symptoms early, such as examining the cells taken from a tissue sample. Historically, these cell samples would need to be evaluated by a medical professional through a manual process which is both time consuming and open to human error through mis-interpretation of analysis. Machine learning can help to improve the process of evaluating the tissue samples taken for breast cancer diagnostic imaging screening by leveraging the power of computers and algorithms to detect anomalies and cell structures which may be indicative of breast cancer.

Machine learning algorithms can help to evaluate the size, structure, and shape of tissue sample cells to apply mathematical algorithms to classify a sample as either malignant or benign. This form of artificial intelligence can process historical tissue sample data and learn what values and combination of values indicate a sample is statistically likely to either be malignant or benign. This mathematical determination based on probability can have it's performance measured to evaluate how accurate the predictions are and human data scientists can build models to optimize their performance.

By leveraging technology to evaluate diagnostic imaging in breast cancer screening, the medical community can enhance the value of early detection, which ultimately results in a more effective and efficient diagnostic imaging process that exceeds the performance of manual reviews done by humans.

Abstract

Breast cancer is the most diagnosed cancer in the world according to the World Health Organization. The mortality rate did not change from the early 1900s and only recently began improving in the 80s. This is due to many countries beginning to adopt early detection programs along with an improvement in the capabilities of science and health care. Diagnostic imaging is used to screen for breast cancer and is recommended as a preventative measure for women due to the positive impact early detection can have on treatment. Part of the diagnostic screening includes taking a tissue sample and examining the sample to determine if it is a malignant or benign diagnosis. This classification is critical as a mis-diagnosis can result in expensive medical bills, unnecessary invasive procedures, delays in treatment, overtreatment, and emotional damage (CDC, 2021).

Data science can help to address these issues by applying an algorithm to determine the statistical probability that a diagnosis is malignant or benign and classify the image accordingly. A Support Vector Machine Learning model can accurately classify an image with 94.55% accuracy and other algorithms show potential to increase this accuracy even greater. By leveraging machine learning techniques, breast cancer screening can become faster and more accurate than manual reviews, resulting in a net positive impact for the patients and medical community. The use of machine learning to examine diagnostic images creates potential for use beyond breast cancer screening and should be explored as a next generation solution to improve diagnostic image reviews in a wide range of use cases.

Introduction

There are many significant events in life which people may think they would never be able to predict or gain awareness of; one of these being the onset of various diseases, cancers, etc. However, with the power of predictive analytics, especially in the medical field, this mystery does not seem so dooming! Strides are being made in being able to predict whether a tumor is malignant or benign, which symptoms or side effects can predict illnesses, and so on. Through this project, the focus will be around insights on breast cancer.

According to the American Cancer Society, "the average risk of a woman in the United States developing breast cancer sometime in her life is about 13%. This means there is a 1 in 8 chance she will develop breast cancer". Additionally, breast cancer is the second leading cause of cancer death in women ("Breast Cancer Facts and Statistics", 2022). However, due to women finding breast cancer earlier in recent years, the breast cancer death rates in women have begun to decline. Early identification is directly related to proper screening, which comes along with being able to accurately predict breast cancer tissue images' diagnoses.

The predictive analytics in this project will utilize past breast screening tissue image data to classify future images' diagnoses as either benign or malignant. A benign diagnosis refers to a diagnosis that the breast tissue is non-cancerous, whereas a malignant diagnosis would specify that the tissue is harmful and cancerous. Through selection of appropriate models and model specifications, the goal and scope are to identify a model which has a high accuracy in identifying tissue images of concern, which can be extremely beneficial for early identification and quick intervention for the longevity of patients' lives.

Methods

The dataset chosen for this project was sourced from the University of California Irvine's Machine Learning Repository. Considering the source of this data originates from a respected educational research institute, there are limited concerns regarding the integrity or quality of the data. The dataset contains features that were extracted from a digitized medical image used to form a breast tissue sample and has a defined target variable that classifies the data points as malignant or benign. The variables within the dataset contain various measurements related to the cell structure, including the following for each cell nucleus:

- Radius (mean of distances from center to points on the perimeter)
- Texture (standard deviation of gray-scale values)
- Perimeter
- Area
- Smoothness (local variation in radius lengths)
- Compactness (perimeter^2 / area - 1.0)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry
- Fractal dimension ("coastline approximation"

The mean, standard error, and worst (mean of three largest values) were computed for each image resulting in a total of 30 features. This data set is suitable for a supervised machine learning project as it has a clear target variable defined, which is the diagnostic value of malignant or benign.

Along with having a clear target variable, the dataset has no missing data and no duplicates which need to be handled, so it helps with being able to produce clear transformations before modeling. The columns "Unnamed: 32' and 'id' were also removed during the exploratory data

analysis phase, since the former only contained null values and the latter is not relevant to the data problem of diagnosing breast tissue images for breast cancer.

With the dataset cleansed, distributions and relationships between variables were explored. The target variable has a relatively even class distribution between the two classes: ‘B’ and ‘M’, 357 cases and 212 cases respectively. Due to these counts, the classes do not seem to be imbalanced, which will be taken into account for modeling for handling the target variable.

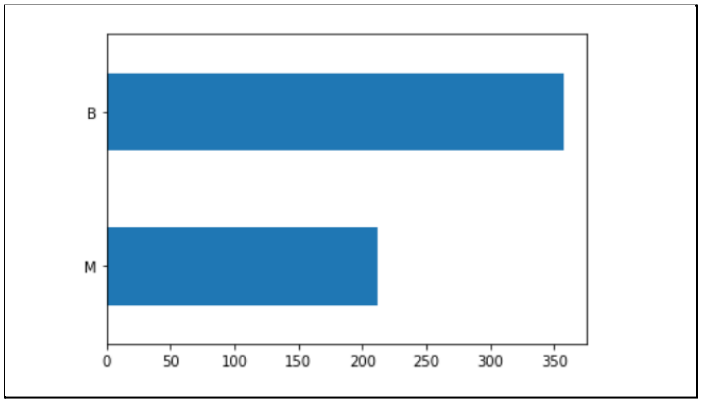

Figure 1: Distribution of the target variable ‘diagnosis’

In terms of distributions for the predictor variables, boxplots and histograms were visualized to decipher normality and possible presence of outliers. Most of the distributions for the numerical predictor variables in the dataset leaned towards a positive skew, where the majority of the points are at the lower value range. There are a few variables that have relatively normal distributions such as texture_mean, smoothness_mean, symmetry_mean, texture_worst and smoothness_worst. Their distributions are mostly centered around the mean and follow a relatively close Gaussian distribution. The histograms showcased a few outliers, which were further explored in boxplots.

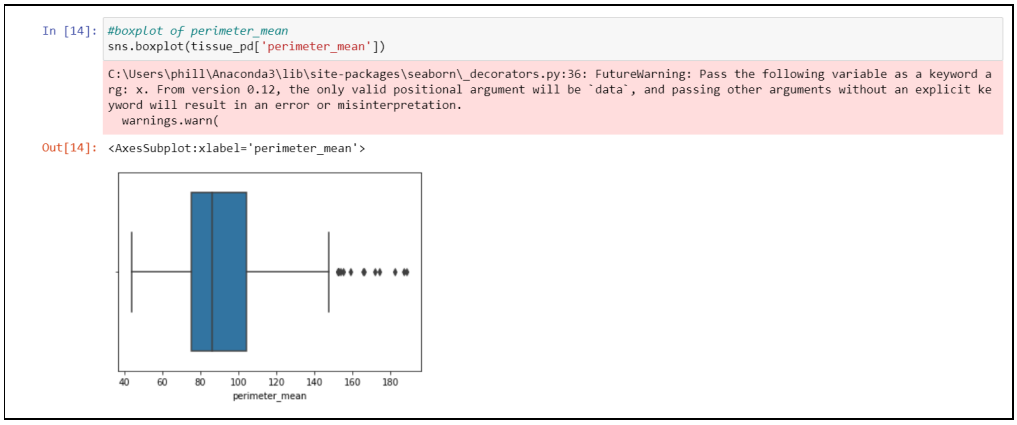

Figure 2: Box Plot of 'perimeter_mean' predictor variable

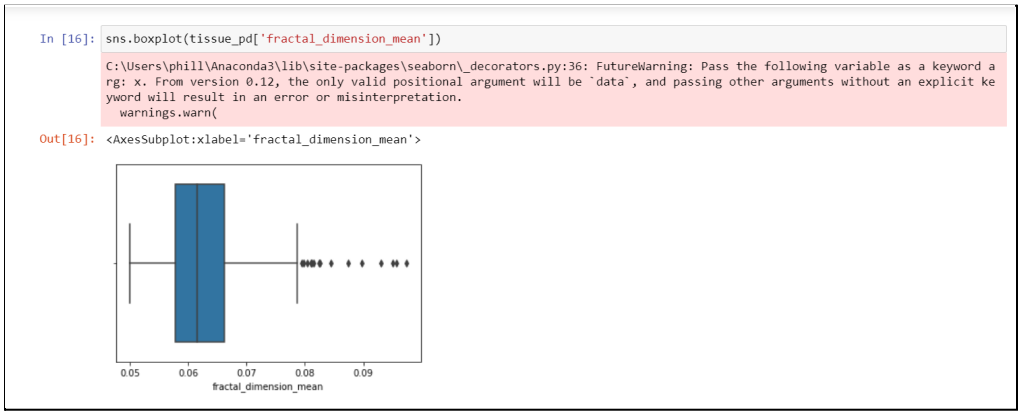

Figure 3: Box Plot of 'fractal_dimension_mean' predictor variable

As seen from the boxplots above, there are clear outlier points beyond the maximum line which represents Q3 + 1.5*IQR. However, as mentioned in the original paper for the research study which originated this dataset, *Nuclear Feature Extraction for Breast Tumor Diagnosis,* "the extreme values are the most intuitively useful for the problem at hand, since only a few malignant cells may occur in a given sample" (Street;Wolberg;Mangasarian, 1992, p. 5). Therefore, due to this reasoning, no outliers were removed from the dataset since the extreme values are integral to the decisioning around diagnoses.

Besides the distributions for the variables, correlation plots were depicted as well to explore variable relationships. There were quite a few relationships that featured high positive correlations; since this could be a sign of collinearity, variance-inflation factors (VIF) were

calculated to decide existence. Variance inflation factor “measures how much the behavior (variance) of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables” (Investopedia Team, 2021). In the Python function for calculating VIF and reducing features, the boundary for feature removal based on VIF is 5.0. By utilizing the VIF of the variables, the variables that had collinearity with each other were dropped after function execution. All features were dropped except for the following: texture_se, area_se, concavity_se and concavity_worst. These features will be used further for modeling.

For model selection, several models which are appropriate for a classification model will be used to evaluate performance accuracy, best parameters to use, and best estimators within the model. Using multiple algorithms is appropriate as it allows for hyperparameter definitions to be set for each model. The models being evaluated include the following:

- K-Nearest Neighbor - Straight forward pattern recognition model which allows the testing of several k values and leaf sizes to determine the best performance - Naive Bayes - Calculates the possibility of whether a data point belongs within a certain category - Decision Tree Classifier - A decision tree is a supervised learning algorithm that performs strong in classification problems	- Support Vector Machine - A support vector machine (SVM) uses algorithms to train and classify data within degrees of polarity, which works well for complex data if a decision is needed beyond x/y - Random Forest Classifier - Expands beyond a decision tree by constructing multiple decision trees to remediate forcing a binary decision

The models were evaluated using GridSearchCV, which allows for several models to be evaluated and their performance measured using a user defined metric. The advantage of using GridSearchCV is the ability to evaluate a model by going through the various combinations of user defined parameters to determine the best combination based on the defined metric of

accuracy. This method also has limitations, however, with the best parameters falling within the user defined parameters range.

Results

The results of the model evaluation with the aforementioned features identified in the VIF evaluation concluded that Support Vector Machine Learning had the strongest performance with 94.55% accuracy. Next in performance was Random Forest Classification, which had 93.50% accuracy with a maximum of 3 features using Gini Importance. The other models did not have sub-par performance, with K-Nearest Neighbor having the worst performance at 89.29%. This strong performance amongst all the models highlights their ability to solve classification problems in a supervised learning model. Furthermore, the performance was evaluated using a confusion matrix, which determines the number of true positives, false positives, true negatives, and false negatives. The precision and recall was calculated to help further evaluate the performance of each algorithm. Recall helps to determine the ability to find all relevant instances of a class in a data set, whereas precision is used to determine the proportion of data points that the model says exists within the relevant class were indeed relevant. While all models had precision and recall scores which were considered good (> 80%), Support Vector Machine algorithm had both the highest recall and precision performance, supporting it's position as the best algorithm to power the machine learning model.

```
--------------------------------------------------------------------------------------------------------
KNN Results
Best Accuracy : 0.8928571428571429
Best Parameters : {'weights': 'distance', 'n_neighbors': 9, 'leaf_size': 29, 'algorithm': 'ball_tree'}
Best Estimator : KNeighborsClassifier(algorithm='ball_tree', leaf_size=29, n_neighbors=9,
                     weights='distance')
--------------------------------------------------------------------------------------------------------
Naive Bayes Results
Mean Accuracy : 0.9121867167919799
--------------------------------------------------------------------------------------------------------
Decision Tree Classifier Results
Best Accuracy : 0.9156641604010025
Best Parameters : {'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 3}
Best Estimator : DecisionTreeClassifier(max_features='auto', min_samples_split=3,
                       random_state=42)
--------------------------------------------------------------------------------------------------------
Support Vector Machine Results
Best Accuracy : 0.9455513784461151
Best Parameters : {'C': 1000, 'kernel': 'linear'}
Best Estimator : SVC(C=1000, kernel='linear', random_state=42)
--------------------------------------------------------------------------------------------------------
Random Forest Classification Results
Best Accuracy : 0.9350256171401956
Best Parameters : {'bootstrap': True, 'criterion': 'gini', 'max_depth': None, 'max_features': 3, 'min_sampl
es_leaf': 1, 'min_samples_split': 2}
Best Estimator : RandomForestClassifier(max_features=3, random_state=42)
--------------------------------------------------------------------------------------------------------
```

Figure 4: Five Models - Evaluation Results (Accuracy)

Figure 5: Five Models - Evaluation Results (Precision and Recall)

Discussion/Conclusion

The use of machine learning to evaluate breast cancer tissue diagnostic images can help to supplement the manual review for preventative and diagnostic screening. This case study shows that a machine learning model can operate within a high degree of accuracy, which allows for expedited processing of images and could be scaled globally. Support Vector Machine Learning was shown to have the highest accuracy at 94.55%, provided only the features used from the VIF test were used. The Random Forest Classification also performed well, with 93.50% accuracy, however the max number of features were 3 as calculated by Gini Importance. Additional Feature Importance Engineering and exploration of different methodology to extract features may improve that classification performance to achieve an even greater accuracy. Additionally, to scale this case study and to help normalize the data across the global community, standardization of image resolution is recommended. Machine learning can help to improve manual tasks with greater speed and accuracy, resulting in numerous benefits and this case study is no exception.

Acknowledgements

Acknowledgement and accolades are given to the University of California Irvine's Machine Learning Repository for compiling this valuable dataset in support of a great public health cause that can potentially improve the survivability of breast cancer. Additionally, the American Cancer Society has continuously driven the effort to educate communities about the importance of breast cancer awareness. Lastly, the Python developer community has made tremendous advancements in developing libraries that help to expand the use of machine

learning applications and drive the mission of expanding machine learning with developers by continually updating and improving the performance of various libraries.

References

Centers for Disease Control and Prevention. (2021, September 22). What is breast cancer screening? Centers for Disease Control and Prevention. Retrieved from https://www.cdc.gov/cancer/breast/basic_info/screening.htm.

Street, N. W., Wolberg, W. H., & Mangasarian, O. L. (1992). (rep.). *Nuclear Feature Extraction for Breast Tumor Diagnosis*.

Team, T. I. (2021, October 31). *Variance inflation factor (VIF)*. Investopedia. Retrieved January 15, 2022, from https://www.investopedia.com/terms/v/variance-inflation-factor.asp

breastcancer.org. (2022, February 9). *Breast Cancer Facts and Statistics*. Breast cancer facts and statistics. Retrieved March 2, 2022, from https://www.breastcancer.org/facts-statistics