Aditya Sumbaraju

Bellevue University

DSC 500 Introduction to Data Science

Professor Fadi Alsaleem

Sept 12, 2021

Introduction

Essentially, data mining is the act of discovering innovative data models that are possibly helpful, valid, and finally comprehensible. An enormous quantity of medical data is being generated due to medical database management technologies and methods that are cutting-edge (Bakr et al.,2020). Breast cancer research data will be analyzed utilizing supervised learning techniques. A breast cancer dataset is analyzed using bioinformatics classification techniques.

The first data sets have been classified, and an optimal algorithm for the detection and prediction of breast cancer has been developed with the assistance of many volunteers. The proposal enables us to choose the most effective classification for breast cancer prediction, assess different approaches for mining data from the breast cancer dataset, and determine the most successful performance-based strategy for disease prediction.

Background

Machine learning may now be included in the armory of a predictive analyst, which may help them perform better at their jobs. Machine learning includes predictive analytics as one of its components. Machine learning algorithms may be capable of reliably detecting the development of breast cancer soon (Bakr et al.,2020). When medical practitioners have a more accurate prediction of their patients' breast cancer survival, they can make more informed treatment planning decisions, avoid unnecessary therapy and thus lower economic costs, more effectively enroll and exclude patients from randomized trials, and develop palliative care and hospice.

Problem Statement

Breast cancer is diagnosed in a large number of persons each year. A challenging diagnosis is required since the illness has a human origin. As a consequence of these and other incidents, the use of machine learning and computers to detect this disease has risen significantly in recent years. We evaluate several classification learning algorithms to prevent benign and malignant cancers based on a breast cancer dataset. We will be focusing on machine learning algorithms in this course. Examples of algorithms include Nave Bayes, the nearest neighbor technique, logistic regression, enhanced learning, and vector support technology, to name a few examples.

Scope

The (principal *component analysis)* PCA's primary objective is to reduce the size of a data set that comprises multiple variables that are strongly or weakly related to one another while keeping the maximum amount of variation possible (Khourdifi & Bahaj,2018). Each variable is translated into a new set of orthogonal vital elements (or PCs), yielding an orthogonal set of major orthogonal elements. The PCs are arranged so that the change in each of the original variables diminishes as we progress down the list. Because the original components are generally kept, a significant portion of their variety in the first element is preserved using this technique. As with the vectors of Manfred Eigen, the essential elements of a covariance matrix are orthogonal. The PCA dataset must be scaled. The results are affected by relative scaling. In layman's terms, this is a strategy for summarizing data. Consider placing several bottles of wine on a wooden board. Numerous characteristics separate one wine from another. However, redundancy may occur if more than one person resides in the same home. PCA can only summarize each wine in the stock

with fewer characteristics (Khourdifi & Bahaj,2018). The user is presented with an intuitive depiction of this object, a projection or "shadow," from the most informative perspective. Univariate analysis was conducted to find credible independent variables, which was done using a time-dependent Cox regression model. Following the Akaike Information Criterion (AIC) determination, we utilized the backward selection technique to choose the final multivariable model used in the study (20). Starting with the least significant variable, the model of independent variables was re-run using the AIC technique until no variables were eliminated. The model was re-run using the reverse selection method. In all studies, the value of 0.05 was used. In all testing, R version 3.6 was utilized (Khourdifi & Bahaj,2018). For the final multivariable model, a small number of clinically irrelevant variables were carefully selected from a large number of candidates. Algorithms that can maintain their state This is due to the thorough training received. So that the number of recurrences in each of the three groups remained the same, we formed mutually exclusive groupings of 60/20/20 components.

These and other characteristics, which are time-dependent, contribute to developing the fundamental idea of survival (lab test results, mammography, and so on). System A uses the MNK-size matrix A. M represents the total number of patient tracking records and N and k represent the dimensions of the monitoring window, respectively. A grid search was performed to identify the best window size for time-stamped data, and the results were analyzed. There were 32 hidden cells during the first layer, while the second layer featured 20 hidden cells. Activation layers with hyperbolic tangents were added once the recurrent layers with gated units had been successfully established. Initially, Adam's rate was set at zero, but after numerous rounds of optimization, the model was decreased to a value between ten and fifteen. To train the model, a 100-piece method was utilized. Expulsions accounted for 0.25 percent of all students. The Python network structure

was created using the Keras and TensorFlow frameworks, respectively "(Python 3.5, Keras 2.1.2, TensorFlow 1.4.0)". Logistic regression, random forest, and gradient optimization are some of the machine learning approaches that have been frequently utilized in medical applications to create our models. Grid search was used to determine the range of hyperparameters available for each model. This study examined the relationship between predicted and actual repetition times using the Harrell match index in the Python lifeline module (C-Index) (22, 23).

An evaluation of recurrence estimates overtime periods of two, five, and seven years was conducted using the AUC method. These genes were utilized to evaluate the performance of the model in respect of the key BC characteristics. Please remember that we did not compare each group's expected and actual values after two, five, or seven years of follow-up. The median absolute and maximum errors were used to evaluate the specificity and accuracy of the model, respectively. Once the follow-up period surpassed expectations and there were no repetitions, the data sets were deleted, replaced with fresh ones, and then destroyed.

Preliminary Requirements

The Madison Breast Cancer University of Wisconsin Hospitals Database data were analyzed in the publicly available research. Each sample has eleven distinct characteristics. Each occurrence was assigned some attributes ranging from two to ten. There are 699 instances in all. Specific instances are destroyed due to a lack of functionality (Islam et al.,2017).

Along with the single class attribute, there are nine more characteristics. Each occurrence is likely to result in one of two outcomes: benign or malignant. Another numerical-valued variable is the instance ID. We gather two types of data. Tumors have been identified as benign as well as malignant (M). Thirty of the 569 variables were generated during the data analysis. Breast cancer

screening is a classification or clustering problem that machine learning methods such as neural networks may help address.

Our method depends on a massive database of continuously updated dangerous and good file information to accelerate the classification process. For example, clustering methods may be more efficient when dealing with limited categories, such as breast cancer data. It provides the theoretical framework for future study. After collecting and selecting attributes, the newly obtained data may be submitted to machine learning techniques. Naive Bayes, Random Forest, and Vector Support Machines are just a few of the methods that may be utilized to understand how to use the program.

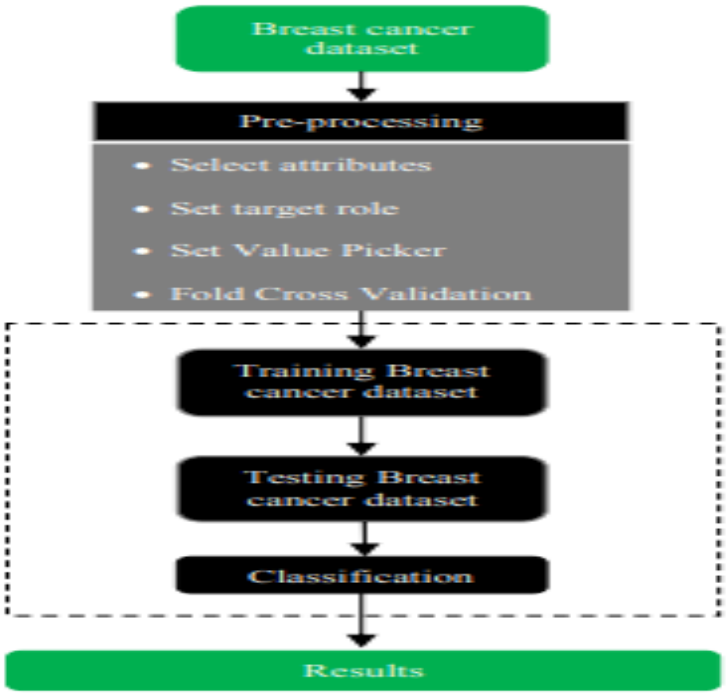
Technical Approach

The parameters for this section are described, and the evaluation results of the machine learning algorithms employed in this section. The fraction of correctly identified occurrences serves as a proxy for detection accuracy (Islam et al,.2020). We may compute this, for example, by dividing the number of correct forecasts by the total number of events in your data collection. As previously stated, accuracy is highly dependent on classifier thresholds and hence varies between experiments. While comparing classification algorithms is unsuccessful, it provides a representative class image. We may use the following equation to determine the precision:

$$Accuracy = \left(\frac{(TP + TN)}{TP + FP + TN + FN}\right) \quad (1)$$

Where: TP = True positive; FN= False negative; FP= False positive; TN = True negative.

P and N denote the positive and negative populations associated with the incidence of malignant and benign tumors, respectively. The figure below shows the architectural design of the proposed structure



Analysis

We will look through the dataset for possible patterns for those blood donors readmitted within the 30-day timeframe. We intend to look at age, gender, race, diagnosis type, and all the different types of medication to check for any possible relationships. We will need to put together tests that will look at each value and help us determine whether there is a relationship at a sample level, and if that passes, we can move forward to working with the more extensive data.

Requirement Development

To develop a prediction model for breast cancer, we must use the most up-to-date data available. Consequently, external data sources must be incorporated into your model as early in the development process as is reasonably practicable. Create auditing, monitoring, and retraining systems that are effective and efficient (Islam et al.,2017). When we rely on the same helpful data repeatedly, we run the danger of a model straying from the route we planned. To do this, the process must be simplified and automated to the greatest extent feasible while also depending on the most recent, relevant, and high-quality data available.

Model Deployment, Testing, and Evaluation

When it comes to using public policy machine learning models, governance, trust, use costs, and precision are all important issues to consider. The Charlotte-Mecklenburg and Nashville Metropolitan Police Departments, for example, have both adopted early intervention systems that are based on machine learning and artificial intelligence (Dhahri et al.,2019). The improvement of the system is the next objective of this continuing endeavor. When supervisor feedback is included in the models, the accuracy of the models should gradually improve over time. We want to develop and test more techniques for assessing the importance of officer-level traits in future research. Supervisors can assign the most appropriate treatment choices based on the performance of machine learning treatments in real-time. If we have used machine learning models, the next step is to determine how well they performed on your datasets. Therefore, models are developed on the test data set as a result of this process. The data from the test set accounted for around 30 percent of the total data set utilized to estimate the risk of breast cancer. To evaluate the risk of breast

cancer, a 10-fold cross-validation test was also carried out. It was necessary to use various metrics to determine and compare the performance of the multiple techniques.

A patient is categorized as having no difficulties by the model, but in reality, they are experiencing challenges due to the model's false prediction (0). Adverse error is a type of error that occurs when something goes wrong (Islam et al,.2020). The prediction may be incorrect if the anticipated class is valid (1) but the actual class is false (-0). The results are more clearly expressed when the standardized confusion matrix is used (NCM). When compared to the confusion matrix, this is less confusing. In computing the accuracy of a model, the percentage of the total number of predictions made by the model is multiplied by 100.

In other words, it is a measure of how many patients experienced difficulties concerning how many others did not have trouble. In other words, this is the fraction of patients experiencing problems compared to those experiencing model-dependent issues. The ability of a classifier to predict bad outcomes is directly proportional to the degree to which it is specific. In many ways, it is the polar opposite of the concept of reminder. This represents the proportion of patients who have not had any difficulties.

Expected Results

With the use of the tenfold cross-validation test, we may apply and assess our classifiers. To establish the distribution of effectiveness and efficiency values, we attempt to visually examine the data using our pre-treatment and preparation methods to ascertain their visual evaluation (Dhahri et al.,2019. When evaluating efficiency, consider how much time it takes to create the model, how many examples are successfully categorized, how many cases were wrongly classified, and what degree of precision is achieved. The table below classifies the performance

The standard evaluation criteria:

Considering below factors to evaluate the models:

- Time to build the model
- Correctly classified instances
- Incorrectly classified instances
- Accuracy
- TP Rate
- FP Rate
- Recall
- Precision

And I am expecting the highest rate of accuracy using SVM, followed by K-NN. I plan to execute the K-NN, SVM, RF, NB classifiers and choose the optimal classifier for my use case.

It is necessary to take simulation faults into account to properly evaluate the effectiveness of classifiers in this study. As a result, we use the following criteria to assess the effectiveness of our classifier. The word "absolute error" refers to how forecasts or projections accurately predict the result of a particular situation. Absolute Error and Mean Squared Error are terms that are used to describe errors. I would be considering the "confusion matrix" to evaluate the performance of the classifier. And finally, end with the R-squared (ROC) curve to represent the performance of multiple classifiers over time. Thus, we may choose the best models from the plot and discard others depending on the most accurate grading. As a result, each row in the table above shows the actual rates for each category, while each column represents the projections.

Execution and Management of Project

Project Plan

All that is left is to carry out and monitor the master project plan. Processes allow us to keep track of time, money, quality, change, risk, and concerns. The primary goal of building a machine learning model to predict breast cancer is to foresee results that may be used to plan for and prevent similar occurrences in the future (Dhahri et al.,2019). I found the primary dataset and will be working on data cleansing methodologies to clear the outliers, duplicates, and missing values. There is also an explanation of why the project's broader transformation approach should help it progress. Critical projects will be delivered efficiently by project leaders who express their visions to their teams and work with them. It may be beneficial to relate your vision and strategy to your performance to bridge the gap between viewpoint and performance. Remember that the model's execution will be influenced by the style, timeline, and money. During the project planning phase, the project performance plan outlines the project's results and acceptance criteria. Acceptance criteria, in other words, are standards that a user utilizes to determine whether or not a specific service is acceptable. This technique simplifies the process of procedure improvement. I would be communicating in Teams with the groups and incorporating the comments received via peer reviews. R and Python notebooks will be worked on individually, and I would like to check-in the code in the Git Hub repository. The details of the GIT repo will be published in Teams and Blackboard for visibility.

Project Risk

It is necessary to develop a new set of risk models that include both current and long-term risk evaluations. We want to build and test a risk prediction system for our business. Radiologists use the BI-RADS scale to evaluate screening mammography. 123,869 women and 447,640 mammograms were included in the research (Islam et al,.2020). Self-reported family history and previous breast biopsies were used to collect data. Diseases that do not spread or do not spread rapidly are classified as such. The absolute probability of breast cancer diagnosis was calculated using women's monitoring, prevision time, and risk profile. The findings indicate that all predictors are compatible with the proportionate risk hypothesis. Internal validation was used to assess the model's predictive capabilities.

Breast cancer was found in 2 068 of 121,969 females studied after an average of 8,52 years of follow-up. 18.34 percent of breast cancer patients had a family history, and many had ambiguous biopsies (13,86 percent versus 13,86 percent). The difference is 23.76 percent vs. 22.72 percent. Females were aged 55-59 years, and obese women had a higher risk of breast cancer (E/O ratio of 1,19; 95 percent confidence interval: 1.08-1,45). (Islam et al,.2020). Mammography has been added to the list of diagnostic procedures for such cancers, calcifications, and architectural abnormalities. Thus, the true nature of the danger was never overstated or misunderstood.

References

Bakr, M. A. H. A., Al-Attar, H. M., Mahra, N. K., & Abu-Naser, S. S. (2020). Breast Cancer Prediction using JNN. *International Journal of Academic Information Systems Research (IJAISR)*, *4*(10).

Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated breast cancer diagnosis based on machine learning algorithms. *Journal of healthcare engineering*, *2019*.

Islam, M. M., Haque, M. R., Iqbal, H., Hasan, M. M., Hasan, M., & Kabir, M. N. (2020). Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, *1*(5), 1-14.

Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2017, December). Prediction of breast cancer using support vector machine and K-Nearest neighbors. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 226-229). IEEE.

Khourdifi, Y., & Bahaj, M. (2018, December). Applying best machine learning algorithms for breast cancer prediction and classification. In *2018 International conference on electronics, control, optimization and computer science (ICECOCS)* (pp. 1-5). IEEE.

Dataset:

https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

William H. Wolberg and O.L. Mangasarian: "Multisurface method of pattern separation for medical diagnosis applied to breast cytology", Proceedings of the National Academy of Sciences, U.S.A., Volume 87, December 1990, pp 9193-9196.

O. L. Mangasarian, R. Setiono, and W.H. Wolberg: "Pattern recognition via linear programming: Theory and application to medical diagnosis", in: "Large-scale numerical optimization", Thomas F. Coleman and Yuying Li, editors, SIAM Publications, Philadelphia 1990, pp 22-30.

K. P. Bennett & O. L. Mangasarian: "Robust linear programming discrimination of two linearly inseparable sets", Optimization Methods and Software 1, 1992, 23-34 (Gordon & Breach Science Publishers).