ML Algorithms for Breast Cancer Prediction

Tanishq Gautam

Abstract

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society.

The goal of this project is to classify whether the tumor mass is benign or malignant in women residing in the state of Wisconsin, USA. This will help in understanding the important underlying importance of attributes thereby helping in predicting the stage of breast cancer depending on the values of these attributes.

Github

1. Introduction

Of the 184 major countries in the world, breast cancer is the most common cancer diagnosis in women in 140 countries (76%) and the most frequent cause of cancer mortality in 101 countries (55%). Breast cancer ratios are statistically higher in women in more developed countries as compared to other diseases. But it is also globally increasing day by day. Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body. Excluding skin cancer, breast cancer is the most common type of cancer in women in the United States, accounting for one of every three cancer diagnoses. Breast cancer ranks second among cancer deaths in women.

Breast cancer prediction has been the subject of extensive study. Below are some notable studies that have been conducted: - The utilization of current technology breakthroughs to create breast can- cer prediction models is discussed by Chaurasia V. He develops a prediction model using Naive Bayes, RBF Network, and J48 and employs the 10-fold cross-validation approach to measure the unbiased estimate of these models' performance. Verma D used five classification algorithms Naive bayes, SMO, REP Tree, J48 and MLP upon breast cancer dataset. Ojha. U. places emphasis on the choice of parameters for estimating the likelihood of breast cancer recurrence using data mining techniques. He shows how clustering and classification methods are used. According to the author, for the experimental data set, classification techniques performed better than clustering. K-Means, EM, PAM, Fuzzy c-mean, Mean, and KNN were used for clustering, while Naive Bayes, SVM, and Mean were used for classification. A classifier that can distinguish between benign and malignant breast tumors is created by comparing two machine learning algorithms, according to B.L. Rodrigues. The majority of studies appear to only achieve accuracy rates of 93– 94

2. Dataset

The dataset utilized in this study was produced by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital in Madison, Wisconsin, in the United States, and is openly accessible. On July 15, 1992, Olvi Mangasarian gave it away. Dr. Wolberg used fluid samples obtained from patients with solid breast masses[10] and Xcyt, a user-friendly graphic computer program that can do the analysis of cytological features based on a digital scan, to construct the dataset.

Ten real-valued features are computed 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$\hat{2}$ / 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" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recorded with four significant digits.

3. Experiments

When we calculate mean, variance, standard deviation, number of sample (count) or max min values, it helps us to understand what is going on with the data. We do need standardization or normalization before visualization, feature selection, feature extraction or classification. Also, we can see that area_mean feature's max value is 2500 and

smoothness_mean features' max 0.16340 which implies that we need to do standardization or normalization before visualization, feature selection, feature extraction or classification

Pre-processing will manage the missing attributes, the unbalanced data, and the quantity of attributes used to train the classifier in light of the dataset that was accepted. Two approaches are suggested to handle the 16 missing values: the first is to utilize the "replacemissingvalues" filter. The means from the training data will be used as a replacement for all missing values for characteristics in the dataset by this filter. The revised dataset will contain 683 occurrences if the cases with missing values are all removed.

The first impression is that substituting the missing attributes with the mean value from the training set is not a good idea because the size of a single cell is not related to the mean size of the other cells given the nature of the missing characteristics (all of them are bare nuclei size).

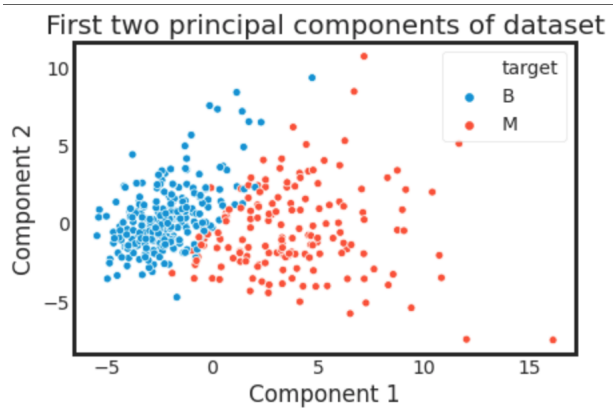
4. Methodology

We have implemented 7 different machine learning algorithms to compare the performance.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear dimensionality reduction technique that can be utilized for extracting information from a high-dimensional space by projecting it into a lower-dimensional sub-space. It tries to preserve the essential parts that have more variation of the data and remove the non-essential parts with fewer variation.

One important thing to note about PCA is that it is an Unsupervised dimensionality reduction technique, you can cluster the similar data points based on the feature correlation between them without any supervision.



K Nearest Neighbours

knn is essentially classification by finding the most similar data points in the training data, and making an educated guess based on their classifications. K is number of nearest neighbors that the classifier will use to make its prediction. KNN makes predictions based on the outcome of the K neighbors closest to that point. One of the most popular choices to measure this distance is known as Euclidean.

We ran a Knn pipeline with standard scaler and pca. We then implemented GridSearchCV which returns the best parameters of KNN and PCA for our dataset.

Using this we get knn n_neighbors' = 11, pca n_components = 9.

KNN gave us a pretty decent performance with a training accuracy of 97% and CV accuracy of 97% and a Test performance of 95%.

Gaussian Naive Bayes

Gaussian Naive Bayes is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. We have explored the idea behind Gaussian Naive Bayes along with an example. Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all what is needed to define such a distribution.

We ran a GB pipeline with standard scaler and pca. We then implemented GridSearchCV which returns the best parameters of PCA for our dataset.

Using this we get pca n_components = 7.

GB gave us a not so good performance with a training accuracy of 93% and CV accuracy of 91% and a Test performance of 91%.

Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1".

We ran a LR pipeline with standard scaler and pca. We then implemented GridSearchCV which returns the best parameters of PCA and LR for our dataset.

Using this we get pca n_components = 8 and LR C value = 1.66.

LR gave us a good performance with a training accuracy of 99% and CV accuracy of 98% and a Test performance of 97%.

Random Forest

Random Forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results - in the case of a random Forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

The random Forest starts with a standard machine learning technique called a "decision tree" which, in ensemble terms, corresponds to our weak learner. In a decision tree, an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. The random Forest takes this notion to the next level by combining trees with the notion of an ensemble. Thus, in ensemble terms, the trees are weak learners and the random Forest is a strong learner.

We ran a Random Forest pipeline with standard scaler . We then implemented GridSearchCV which returns the best parameters of RF for our dataset.

Using this we get max_depth'= 3, n_estimators'= 200.

RF gave us a good performance with a training accuracy of 99% and CV accuracy of 95% and a Test performance of 94%.

Support Vector Machine

SVM depends on supervised learning models and trained by learning algorithms. A SVM generates parallel partitions by generating two parallel lines. For each category of data in a high-dimensional space and uses almost all attributes. It separates the space in a single pass to generate flat and linear partitions. Divide the 2 categories by a clear gap that should be as wide as possible. Do this partitioning by a plane called hyperplane.

An SVM creates hyperplanes that have the largest margin in a high-dimensional space to separate given data into classes. The margin between the 2 classes represents the longest distance between closest data points of those classes.

We ran a SVM pipeline with standard scaler and pca . We then implemented GridSearchCV which returns the best parameters of SVM, pca for our dataset.

Using this we get pca n_components: 8, svc_C: 100.0, svc_gamma: 0.001, svc_kernel: rbf.

SVM gave us a good performance with a training accuracy of 99% and CV accuracy of 99% and a Test performance of 97%.

XGBOOST

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture.

We ran a XGBoost pipeline with standard scaler. We then implemented GridSearchCV which returns the best parameters of XGBoost for our dataset.

Using this we get gamma = 0.30000000000000004, learning_rate: 0.1, max_depth: 4, n_estimators: 100, reg_lambda: 2.154434690031882.

XGBoost gave us a good performance with a training accuracy of 99% and CV accuracy of 97% and a Test performance of 96%.

Stacking

Stacking (sometimes called Stacked Generalization) is a different paradigm. The point of stacking is to explore a space of different models for the same problem. The idea is that you can attack a learning problem with different types of models which are capable to learn some part of the problem, but not the whole space of the problem. So, you can build multiple different learners and you use them to build an intermediate prediction, one prediction for each learned model.

The best estimators for each are used to make uncorrelated predictions which in turn are concatenated and fed into a secondary Support Vector Machine estimator by stacking.

We ran a pipeline with KNN, GB, LR, RF, SVM, XGBoost. We then implemented GridSearchCV which returns the best parameters of Stacking for our dataset.

Using this we get 'C': 1.0, 'gamma': 0.1, 'kernel': 'rbf'.

Stacking gave us a good performance with a training accuracy of 100% and CV accuracy of 100% and a Test performance of 99%.

5. Conclusion

This project investigates different models for breast cancer prediction. Three different types of Machine Learning methods including Random Forest, Support Vector Classifier (SVM), XGBoost are compared and analyzed for optimal solutions. Even though all of those methods achieved desirable results, different models have their own pros and cons. The best performance for training is in XGBoost but it did not give good test performance. SVM with best parameters gave the highest test performance and also an equally comparable training performance as XGBoost. For Random

Forest with univariate feature selection, we observed a better accuracy than implementing the feature selection with correlation. Default data included 33 features but after feature selection, the optimum number of features were 5 with accuracy 95

References

1. Chaurasia V., Pal., S, Tiwari., BB.: Prediction of benign and malignant breast cancer using data mining techniques. Journal of Algorithms Computational Technology, Vol.12(2), pp. 119–126.

2. Verma, D., Mishra., N.: Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques. In Proceedings of the International Conference on Intelligent Sustainable Systems (ICISS), pp 533-538, (2017).

3. Ojha, U., Goel., S.: A study on prediction of breast cancer recurrence using data mining techniques. In 7th International Conference on Cloud Computing, Data Science Engineering – Confluence, pp 527-530, (2017).

4. Rodrigues., B.L.: Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection. In: Proceedings of XI Workshop de Vis ã o Computational, pp 15-19, (2015).

5. UCI Machine Learning Repository: Breast Cancer Wisconsin Dataset

6. Vivek Kumar, Brojo Kishore Mishra,: Prediction of Malignant Benign Breast Cancer: A Data Mining Approach in Healthcare Applications