Predicting the Presence of Breast Cancer in an Abnormal Lump Behavior

Eric Adams

Chapman University, CADS

CS 520 Reproducible Research

Abstract

Using Mathematical Modeling to predict breast cancer presence in abnormal lump behavior in Women and identifying the cause and effects that are prevalent to cancer patients. In an attempt to help execute this experimentation, the research was undertaken to investigate, assess and compare models of distinct algorithmic classification methods in diagnosing the presence of cancer effects or not. In checking for the patterns of cancer affect occurrences over time and deciding on the best predictive models for the analysis, data were collected from the Kaggle website as a secondary source. A model developed for the forecasted number of reported effect cases was a compilation of several classification models stacked together with a learning base model of Linear Discriminant Analysis. The stacked classifier came up with an accuracy score of 93.87% better as compared to individual separate algorithms. Python software was used to analyze the data.

Keywords: Keyword; ML; Python, Lump Structures, Diagnosis, Linear Discriminant Analysis

INTODUCTION

Worldwide, breast cancer is the most common type of cancer in women and the second highest in terms of mortality rates. 2.3 million women received a breast cancer diagnosis in 2020, and there were 685,000 fatalities worldwide. Breast cancer has become a prevalent pandemic across the globe with 7.8 million women alive as of the end of 2020 who received a diagnosis in less than a decade. There is, therefore, the need for its early detection which can provide room for early treatment, this will not only help cure it but also prevent its recurrence and thus help to find preventive measures to improve the quality of life of patients and enhancement of their life expectancy. Diagnosis of breast cancer is performed when an abnormal lump is found (from self-examination or x-ray) or a tiny speck of calcium is seen (on an x-ray). After a suspicious lump is found, the doctor will conduct a diagnosis to determine whether it is cancerous and, if so, whether it has spread to other parts of the body. This breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison by Dr. William H. Wolberg. The dataset contains five (5) independent mean features (radius, texture, perimeter, area, and smoothness) and one categorical label (diagnosis). It appears that class 1 indicates a positive outcome (cancer present) and class 0 indicates a negative outcome (no cancer present) for the breast cancer diagnostic test.

METHODOLOGY

This chapter focuses mainly on the definition of machine learning, applications of machine learning, and its review for data analysis. Machine learning is the process of developing intelligent machines which can learn by itself. In, 1959 Arthur Samuel, a computer scientist at IBM and a pioneer in AI and computer gaming, coined the term machine learning. He defined it as a "field of study that gives computers the ability to learn without being explicitly programmed.

Supervised Machine Learning

Machines are trained under the supervision of humans. These machines are trained using well "labeled data and the goal is to predict outcomes for the new variables. Datasets are referred to as labeled if the column and target variable are fully labeled. SML is very useful in risk assessment, image and object detection, fraud detection, etc. SML can be grouped into Regression and Classification. In the Regression algorithm, the output is expected to be continuous numeric data. Regressions algorithms include SLR, DTR, RFR, SVR, etc. In the Classification algorithm, the output is expected to be categorical data. This output can be grouped into Binary classification and Multi-Class Classification. Classification algorithms include KNN, DTC, Logistic Regressions, Naïve-Bayes, SVC, etc.

Unsupervised Machine Learning

Machines are trained using unlabeled data without any supervision. This mechanism generally focuses on understanding relationships within datasets. Unlabeled data refers to a dataset with no column and target variables. Sample algorithms include but are not limited to Clustering Dimensionality Reductions such as PCA, K-means, DBSCAN, etc.

Classification Algorithms

There are several classification algorithms that were considered in the project. Sample algorithms like Gradient Boost Classifier, logistic regression, SVC, Decision Tree Classifier, Random Forest Classifier, etc.

Decision Trees A decision tree could be a flow-chart-like tree structure that uses a branching methodology to illustrate each outcome of a decision. Every node inside the tree represents a check on a particular variable, and every branch is the outcome of that test.

Gradient Boost Classifier It is a tree-based supervised learning classification algorithm that uses an ensemble boosting technique for classification. It operates on combining weak learners to make strong learners. These weak learners are arranged sequentially to reduce the error (Boosting). It uses a Fully grown Decision Tree as a weak learner or base estimator with maximum depth from 8 to 32. Its base estimator is non-changeable. The learning happens by optimizing the loss functions (MSE, MAE, RMSE).

Linear Discriminant Analysis

LDA is a supervised ML algorithm, that reduces dimensions by constructing new features which are linear combinations of the original features. LDA uses eigen decomposition for achieving its task and this offers data compression, better visualization, and efficiency.

Formalizing the Objective of LDA To ensure maximum class separability, the two criteria must be met: We want to minimize variability within a class (Inter class scatter) We want to increase the between-class variability (Between class scatter) From Figure

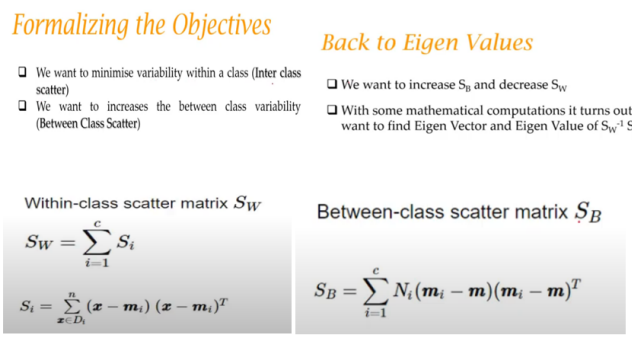

Figure 1 LDA mathematical intuition

1, considering the two classes in the target variable of the dataset. For each individual record, the mean difference is calculated, and this results in computing the individual classes' covariance matrix S(W). To determine S(B), we assign the mean of each class (m1,m2) in a vector form, and find the overall mean of each class and covariance matrix based on the means vector of the classes. We later apply eigen computations based on the matrix product of the inverse of S(W) (minimize) and S(B) (maximize) to design a direction that can give maximum separability. To calculate new variables, we use the input metric and multiply it by a weight vector, which comes from the eigen computations values outcome.

Data Collection

The Hearing test data were gathered from online sources (Kaggle) with five mean feature variables (radius, area, texture, perimeter, smoothness) and a target variable (diagnosis).

DATA ANALYSIS

In this chapter, we shall look at the information that can be derived from the data collected and make inferences based on our outcome of results and then give a conclusion and necessary recommendations.

Data Representation

Online sources were used to gather the data for the hearing test study, which determines if a patient has hearing defect or not. NumPy, Pandas, Seaborn, Matplotlib, and other Python libraries were imported to clean, analyze, preprocess, and assess the data. The dataset was loaded into Jupyter Notebook, and information about its characteristics (such as its shape and description) was retrieved. The dataset was also examined for balance and missing values.

	count	mean	std	min	25%	50%	75%	max
mean_radius	569.0	14.127292	3.524049	6.98100	11.70000	13.37000	15.7800	28.1100
mean_texture	569.0	19.289649	4.301036	9.71000	16.17000	18.84000	21.8000	39.2800
mean_perimeter	569.0	91.969033	24.298981	43.79000	75.17000	86.24000	104.1000	188.5000
mean_area	569.0	654.889104	351.914129	143.50000	420.30000	551.10000	782.7000	2501.0000
mean_smoothness	569.0	0.096360	0.014064	0.05263	0.08637	0.09587	0.1053	0.1634
diagnosis	569.0	0.627417	0.483918	0.00000	0.00000	1.00000	1.0000	1.0000

Figure 2 Data Summary

From Figure 2, the summary statistics of 569 observations were conducted for the experiment. The standard deviation of 351.914129 is an indication of the large dispersion of cancerous diagnostic detections about the mean area. The mean area has the highest descriptive value among all features and hence can pose a skew distribution with a false outcome. The other features have strong dependencies with the mean area, and this can indicate vehement multicollinearity. The mean feature is removed from the analysis. The maximum and minimum values depict the highest and lowest swings of recorded cancerous effect cases respectively. There are no missing values present in the dataset. The records initially indicated an imbalance structure of "1: <Cancer Present>: 357" and "0: <No Cancer Present>: 212", however; using Fix Imbalance Methods like SMOTE to rebuild "sudo" dataset to balance the dataset. Hence, the dataset shows balanced diagnostic counts of 569 cancerous existence.

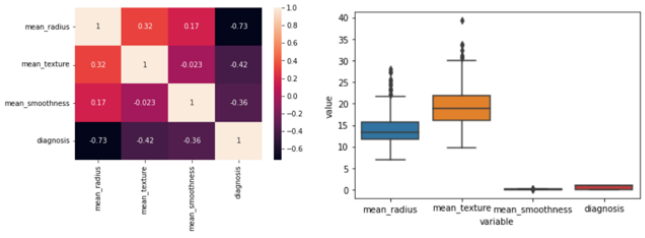

Figure 3 Data Summary

Preliminary Analysis II - Correlation Outlier Detection According to Figure 3, the correlation between Age with the test result is 68% (moderately positive) whiles the correlation between Physical Score and with test result is 79% (moderately Negative). Based on Figure 3, a boxplot for graphically demonstrating the locality and outlier detection indicates that there is an outlier. However, the outlier situation was fixed before the modeling

development. The distribution of the dataset shows the normal distribution.

Independent Variable Interactions

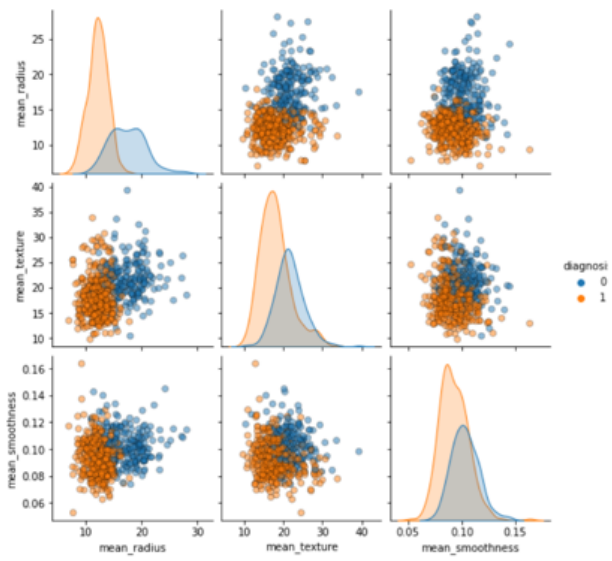

Figure 4 Data Summary

From Figure 4 the functionality package of "pair plot" in seaborn (visualization tool) enables the combinatorial interaction between two independent variables. The mean perimeter and mean area were removed from the analysis due to the heavy linear dependencies among each other. These two feature removals were made to avoid the presence of the multicollinearity effect and redundancy in the dataset for the analysis. The observed figure shows mild overlapping in its distribution, and this indicates the possibility of using tree base algorithms for its analysis due to its non-linearity behavior. However, tree base algorithm renders computational cost. Linear Quadratic Discriminant Analysis would be a good choice since the overlapping between the independent feature interactions isn't completely wild. If the overlapping intensity is wild, a KNN algorithm would have been the best option since it uses the concept of Euclidean/Manhattan distances to find the similarity of data points for its classification.

Model Building

Based on Figure 5, the output result of various algorithms is displayed. It can be identified that the linear discriminant analysis demonstrated outstanding metric performance with a high accuracy score value of 92.46% at 0.020 secs. This algorithmic metric performance without hyperparameter tuning was then followed by QDA ET, RF, Lightgbm, and GBC respectively, etc. Nevertheless, from figure 5, since the top five algorithms had similarly great metric performance with less time of iterations, a stacking classifier with hyperparameter tuning was considered (figure 6). Linear Discriminant Analysis has considered the main estimator while the remaining four algorithms (gbc, ada, lda, and knn) were considered base learners.

Confusion Matrix Decision Boundary Figure 7, illustrates a special kind of contingency table with two dimensions ("actual" and "predicted"). The resulting analysis shows 9 misclassifications arising from both type I and type II errors. A totality

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lda	Linear Discriminant Analysis	0.9246	0.9805	0.9850	0.9088	0.9436	0.8309	0.8443	0.024
qda	Quadratic Discriminant Analysis	0.9183	0.9818	0.9600	0.9191	0.9374	0.8202	0.8276	0.032
et	Extra Trees Classifier	0.9183	0.9752	0.9650	0.9164	0.9380	0.8187	0.8279	0.408
rf	Random Forest Classifier	0.9120	0.9749	0.9500	0.9225	0.9323	0.8061	0.8198	0.449
lightgbm	Light Gradient Boosting Machine	0.9119	0.9725	0.9447	0.9245	0.9319	0.8067	0.8173	0.218
gbc	Gradient Boosting Classifier	0.9089	0.9688	0.9397	0.9276	0.9295	0.7999	0.8149	0.156
ada	Ada Boost Classifier	0.9027	0.9628	0.9300	0.9197	0.9233	0.7901	0.7949	0.195
nb	Naive Bayes	0.9026	0.9751	0.9650	0.8950	0.9267	0.7826	0.7943	0.027
lr	Logistic Regression	0.8901	0.9476	0.9400	0.8954	0.9152	0.7593	0.7682	0.658
dt	Decision Tree Classifier	0.8869	0.8800	0.9100	0.9185	0.9106	0.7557	0.7661	0.027
ridge	Ridge Classifier	0.8650	0.0000	0.9500	0.8567	0.8994	0.6967	0.7098	0.021
knn	K Neighbors Classifier	0.8427	0.9097	0.9042	0.8576	0.8788	0.6550	0.6617	0.049
svm	SVM - Linear Kernel	0.7033	0.0000	0.6313	0.9063	0.6657	0.4455	0.5033	0.026

Figure 5 Model Comparisons Without Hyperparameter Tuning

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9375	0.9833	1.000	0.9091	0.9524	0.8621	0.8704
1	1.0000	1.0000	1.000	1.0000	1.0000	1.0000	1.0000
2	0.9688	0.9917	0.950	1.0000	0.9744	0.9344	0.9364
3	0.9375	0.9833	1.000	0.9091	0.9524	0.8621	0.8704
4	0.9062	0.9833	0.950	0.9048	0.9268	0.7966	0.7984
5	0.8750	0.9667	0.950	0.8636	0.9048	0.7241	0.7311
6	0.9375	0.9958	0.950	0.9500	0.9500	0.8667	0.8667
7	0.7812	0.9417	0.950	0.7600	0.8444	0.4909	0.5270
8	0.9677	0.9955	1.000	0.9524	0.9756	0.9281	0.9305
9	0.8710	0.9693	1.000	0.8261	0.9048	0.7103	0.7421
Mean	0.9182	0.9811	0.975	0.9075	0.9386	0.8175	0.8273
Std	0.0598	0.0168	0.025	0.0716	0.0427	0.1389	0.1285

Figure 6 Model Comparisons With Hyperparameter Tuning

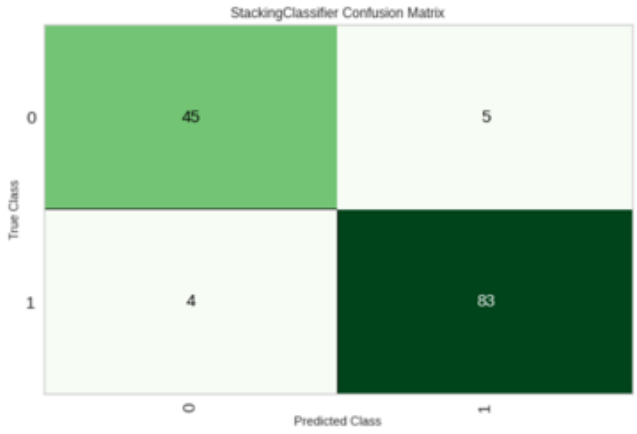

Figure 7 Confusion Matrix

Figure 8 Class Boundary

value of 45 & 83 were correctly classified representing TP and FN respectively. Based on Figure 8, shows the occurrence of misclassification by assigning individuals from " 0: patient diagnosed with no cancerous effect group" to a different category of " 1: patient diagnosed with cancerous " and vice versa.

Statistical Report AND ROAUC

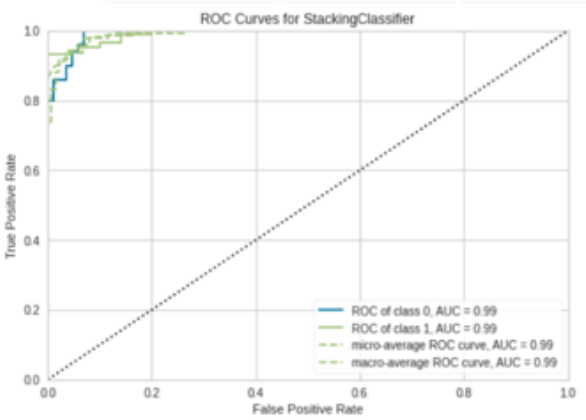

Figure 9 ROAUC Curve

Figure 9 represents the receiving operating characteristics curve performance measurement at various threshold settings. Upon using the stack classifier algorithm, the area under the curve shows a metric performance score of 99% which is greater than the threshold set value (50%). Hence, this indicates a great sign of a good model. The statistical plot from Figure 10 shows a strong test statistic, and this indicates a statistical significance in our variable.

Test Data Prediction

Below is the illustration of Prediction in Figure 11

RESULT, CONCLUSION RECOMMENDATION

Based on the results obtained from this research work, the presence of varied lump structures can be used in predicting or deter-

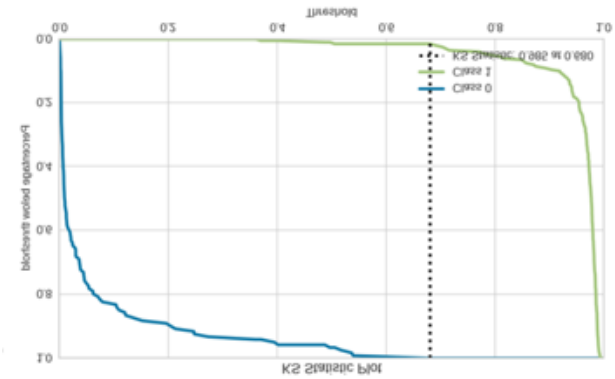

Figure 10 Statistical Hypothesis

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Stacking Classifier	0.9386	0.9767	0.9718	0.9324	0.9517	0.8675	0.8689

	mean_radius	mean_texture	mean_smoothness	diagnosis	Label	Score
0	17.990	10.38	0.11840	0	0	0.9439
1	20.570	17.77	0.08474	0	0	0.9934
2	20.290	14.34	0.10030	0	0	0.9874
3	18.250	19.98	0.09463	0	0	0.9926
4	13.710	20.83	0.11890	0	0	0.9448
5	15.780	17.89	0.09710	0	0	0.8467
6	19.170	24.80	0.09740	0	0	0.9958
7	14.540	27.54	0.11390	0	0	0.9725
8	11.840	18.70	0.11090	0	1	0.8876
9	17.020	23.98	0.11970	0	0	0.9935
10	16.740	21.59	0.09610	0	0	0.9862
11	13.030	18.42	0.08983	1	1	0.9707

Figure 11 Test Data Prediction

mining the presence of Cancer in a female patient's body. From the results displayed above, it can be concluded that Stacking Classifier with Linear Discriminant Analysis as the main estimator coupled with the other four (4) base learners came up with a great accuracy score for this project work. A feature importance extraction is necessary to be investigated on the LDA for further statistical analysis in determining the relative recurrence of cancer spreads on a particular feature based on the feature important extraction of the model. The advanced study of black box algorithms about explainability and interpretability could help reduce the effect of computational cost. It should be noted that practically all the research discussed here ran validation tests to gauge how well their learning algorithms performed. Intuitively, their internal mechanism divides the initial datasets into subsets using well-known evaluation procedures. As previously indicated, significant and independent features that could lead to greater validation are necessary to achieve correct findings for their prediction models. These investigations included internal and external validation to enable the extraction of more precise and trustworthy predictions while minimizing any bias. The small number of data samples is one of the most typical constraints identified in the studies analyzed in this study. The

size of the training datasets must be sufficient, which is a fundamental criterion for employing classification systems to simulate disease behavior. An adequate division into training and testing sets is made possible by a relatively big dataset, which results in good estimator validation. A limited training sample compared to the dimensionality of the data might cause misclassifications, and the estimators can create models that are unstable and biased. It goes without saying that a larger patient population used to forecast patients' survival can improve the predictive model's generalizability. Even though correlation as a base tool for determining the relationship of variable interactions was used, however, an advanced study on "causal inference" could vividly aid in identifying the true parametric variable that could cause a change in the ecosystem of the cancer existence.

REFERENCE

Ahmad, L. Ghasem, A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, and A. R. Razavi. "Using three machine learning techniques for predicting breast cancer recurrence." J Health Med Inform 4, no. 124 (2013): 3.

Yue, Wenbin, Zidong Wang, Hongwei Chen, Annette Payne, and Xiaohui Liu. "Machine learning with applications in breast cancer diagnosis and prognosis." Designs 2, no. 2 (2018): 13.

Fatima, Noreen, Li Liu, Sha Hong, and Haroon Ahmed. "Prediction of breast cancer, comparative review of machine learning techniques, and their analysis." IEEE Access 8 (2020): 150360-150376.

Obaid, O. Ibrahim, et al. "Evaluating the performance of machine learning techniques in the classification of Wisconsin Breast Cancer." International Journal of Engineering Technology 7.4.36 (2018): 160-166.

Ganggayah, Mogana Darshini, et al. "Predicting factors for survival of breast cancer patients using machine learning techniques." BMC medical informatics and decision making 19.1 (2019): 1-17.