Predicting The Presence Of Breast Cancer in Gene Expression

DATA SCIENCE /MACHINE LEARNING (DS/ML)

GROUP 27

ACTIVIE MEMBERS

Eric Adams (Group Leader)

Oladotun Usiola (Assistant)

Oshikoya Oluwatosin (Assistant)

Oyeniran Oyeniyi

BOOTCAMP 2022

CHAPTER 1

INTRODUCTION & ABSTRACT

1.0 BACKGROUD OF STUDY

Cancer refers to a collection of diseases characterized by abnormal and uncontrolled growth, it is caused by alterations or mutations in the genetic code and can be induced in somatic chemicals, radiation, viruses and can more so be inherited.

Figure 1a: From normal to cancerous cells

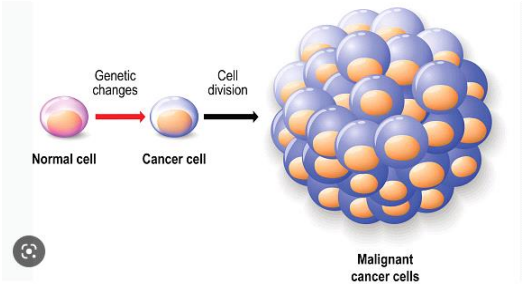
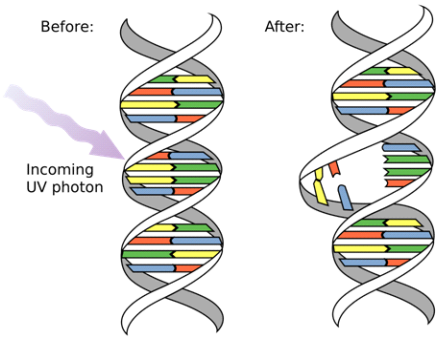


Figure 1b: Mutation



Breast cancer can therefore be defined as a disease in which cells in the breast grow out of control. It consists of a group of biologically and molecularly heterogeneous diseases originated from the breast. It spread outside the breast through blood vessels and lymph vessels(metastasis).

One of the most typical cancers discovered in women is breast cancer. 2.3 million women received a breast cancer diagnosis in 2020, and there were 685,000 dead globally. Breast cancer is the most common cancer in the world, with 7.8 million women alive as of the end of 2020 who had received a diagnosis within the past years. 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 life expectancy.

A gene expression dataset with 3000 observations and two attributes was used. The dataset distribution was normally distributed. The model is trained to understand the characteristic features of the gene expression dataset by cleaning and visualizing to identify descriptive statistics with tools like Sweetviz, dtale (visualization tools) and discard outliers using a boxplot. We verify the linear combinations correlation factor of the two independent variables with the dependent variable, to identify the best predictors for the analysis. Gene One and Gene Two reported 55% and 69% correlation respectively. The dataset is balanced with 1500(patients with cancer)/1500(patients without cancer). Using a pair plot in the seaborn package, we indicated that there was somewhat overlapping at the interactions between the independent variables.

The objective of the project is to develop an ML model that could diagnose the synthesis of this

cancerous behavior in a typical genotype and predict the presence of cancer-based on gene expression.

CHAPTER 2

METHODOLOGY

2.0 INTRODUCTION

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.

2.1 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.

2.2 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.

2.3 CLASSIFICATION ALGORITHMS

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

2.3.1 DECISION TREE

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 that the outcome of that test.

2.3.2 GRADIENTBOOSTCLASSIFIER

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).

2.4 DATA COLLECTION

The Gene Expression data were gathered from online sources (secondary data) with two feature variable (Gene One & Gene Two) and a target variable (Cancer Present).

CHAPTER 3

DATA ANALYSIS

3.0 INTRODUCTION

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 conclusion and necessary recommendations.

3.1 DATA PRESENTATION

Online sources were used to gather the data for the gene expression study, which determines if a patient has cancer 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.

3.1 Preliminary Analysis I (Descriptive Statistics)

Table 3.1.1: Summary Statistics

	Count	Mean	STD	Min	Max	25%	50%	75%	Missing Values	Data Balancing
Gene 1	3000.0	5.600133	1.828388	1.0	10.0	4.3	5.6	6.9	0	1500
Gene 2	3000.0	5.410467	1.729081	1.0	10.0	4.0	5.4	6.7	0	1500
CP	3000.0	0.500000	0.500083	0.0	1.0	0.0	0.5	1.0	0	

From table 3.1.1, the summary statistics of 3000 observations were conducted for the experiment. The standard deviation of 1.828388 is an indication of the large dispersion of cancerous detections in Gene One about the mean. Gene One has the highest mean value of 5.6 indicating a somewhat strong presence and effect of cancerous existence. The maximum and minimum values depict the highest and

lowest swings of recorded cancerous cases respectively. There are no missing values present in the dataset. Our records show balanced cancerous counts of 1500 presence in both Gene One and Gene Two.

3.2 Preliminary Analysis ll

3.2.1 Correlation & Outlier Detection

Table 3.2.1(a): Correlation

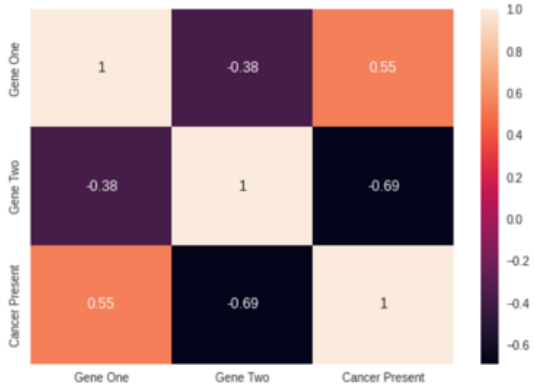
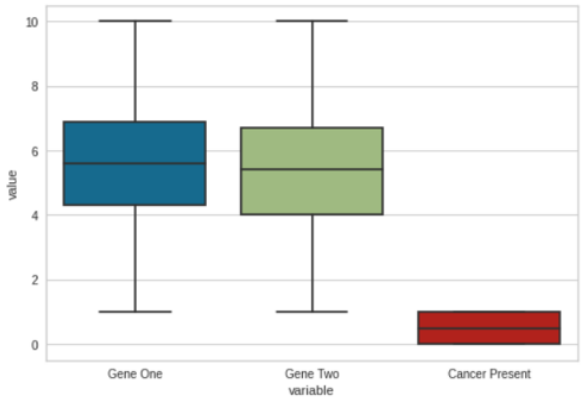


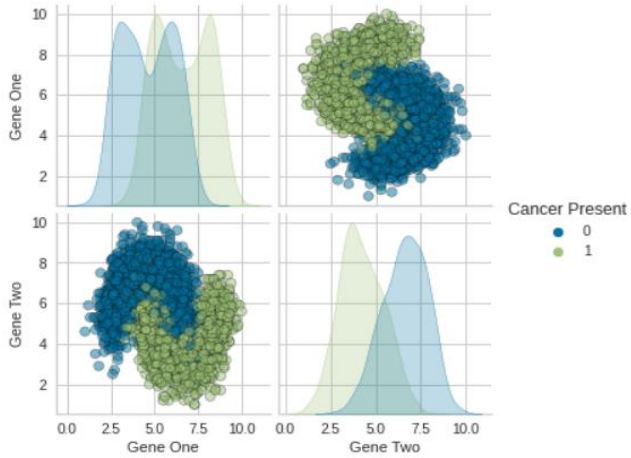
Table 3.2.1(b): Boxplot



According to Table 3.2.1(a), the correlation between Gene One with the Cancer Present is 55% (moderately positive) whiles the correlation between Gene Two with Cancer Present is 69% (moderately Negative). Based on table 3.2.1(b), a boxplot for graphically demonstrating the locality and outlier detection indicates that there is no outlier on the dataset. The distribution of the dataset shows the normal distribution.

3.2.2 Which Algorithm to Choose

Table 3.2.2: Algorithmic Choice



From Table 3.2.2, the functionality of "pair plot" in seaborn (visualization tool) enables the combinatorial interaction between two independent variables. 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. 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.

3.3 MODEL DEVELOPMENT

Table 3.3(a): Metric Performance in each Algorithm

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.9345	0.9794	0.9337	0.9387	0.9359	0.8689	0.8694	0.139
knn	K Neighbors Classifier	0.9285	0.9660	0.9326	0.9286	0.9304	0.8569	0.8573	0.175
ada	Ada Boost Classifier	0.9285	0.9795	0.9198	0.9407	0.9293	0.8571	0.8587	0.211
lightgbm	Light Gradient Boosting Machine	0.9243	0.9745	0.9384	0.9168	0.9271	0.8485	0.8495	0.138
rf	Random Forest Classifier	0.9226	0.9720	0.9291	0.9216	0.9249	0.8450	0.8459	0.709
et	Extra Trees Classifier	0.9172	0.9483	0.9174	0.9213	0.9189	0.8343	0.8352	0.512
dt	Decision Tree Classifier	0.9065	0.9125	0.9023	0.9148	0.9081	0.8129	0.8139	0.026
lr	Logistic Regression	0.8529	0.9389	0.8535	0.8595	0.8560	0.7056	0.7065	0.681
nb	Naive Bayes	0.8517	0.9387	0.8488	0.8610	0.8544	0.7033	0.7043	0.024
qda	Quadratic Discriminant Analysis	0.8517	0.9388	0.8500	0.8602	0.8545	0.7033	0.7043	0.034
ridge	Ridge Classifier	0.8511	0.0000	0.8500	0.8591	0.8540	0.7021	0.7030	0.027
lda	Linear Discriminant Analysis	0.8511	0.9388	0.8500	0.8591	0.8540	0.7021	0.7030	0.022
svm	SVM - Linear Kernel	0.8249	0.0000	0.8465	0.8372	0.8275	0.6494	0.6714	0.046

Based on table 3.3(a), the output result of various displays. It can be identified that the gradient boost classifier demonstrated outstanding metric performance with a high accuracy score value of 93.45% at 0.139 secs. This algorithmic metric performance without hyperparameter tuning was then followed by KNN, ADA, Lightgbm, and RFC respectively, etc. Nevertheless, from table 3.3.1(b), due to the fact that the top five black box algorithms had similarly great metric performance with less time of iterations, a stacking classifier with hyperparameter tuning was considered. GradientBoostClassifier has considered the main estimator while the remaining four algorithms (KNN, ADA, Lightgbm, and RFC) were considered base learners.

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9405	0.9790	0.9535	0.9318	0.9425	0.8808	0.8811
1	0.9405	0.9787	0.9419	0.9419	0.9419	0.8809	0.8809
2	0.9405	0.9841	0.9419	0.9419	0.9419	0.8809	0.8809
3	0.9226	0.9726	0.9302	0.9195	0.9249	0.8451	0.8452
4	0.9524	0.9911	0.9651	0.9432	0.9540	0.9047	0.9049
5	0.9345	0.9820	0.9070	0.9630	0.9341	0.8692	0.8707
6	0.9524	0.9766	0.9302	0.9756	0.9524	0.9048	0.9058
7	0.9226	0.9796	0.9186	0.9294	0.9240	0.8452	0.8453
8	0.9345	0.9736	0.9535	0.9213	0.9371	0.8689	0.8694
9	0.9102	0.9709	0.8837	0.9383	0.9102	0.8205	0.8220
Mean	0.9351	0.9788	0.9326	0.9406	0.9363	0.8701	0.8706
Std	0.0127	0.0057	0.0231	0.0166	0.0128	0.0254	0.0253

Table 3.3(b)

3.3.1 Confusion Matrix & Decision Boundary

Table 3.3.1(a): Confusion Matrix

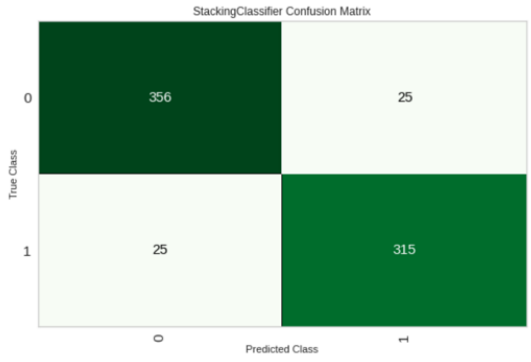
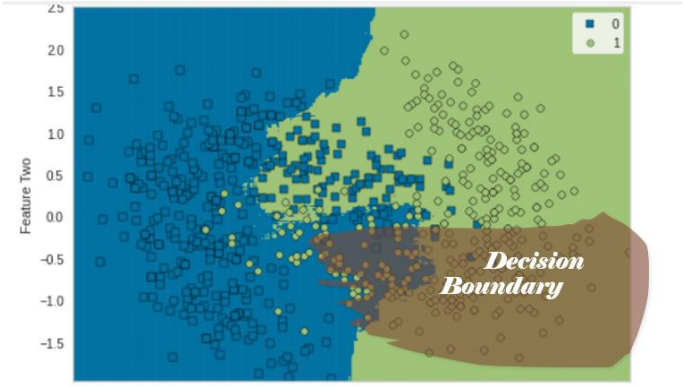


Figure 3.3.1(b)Decision Boundary



From table 3.3.1(a), it illustrates a special kind of contingency table with two dimensions ("actual" and "predicted"). The resulted analysis shows 25 misclassifications arising from both type I and type II errors. A totality value of 356 & 315 was correctly classified representing TP and FN respectively. Based on Figure 3.3.1(b), shows the occurrence of misclassification by assigning patients from "no cancer group" to a difference category of "patient with cancer" and vice versa.

3.3.2 Statistical Report AND ROAUC.

Figure 3.3.2: AUC and Hypothesis

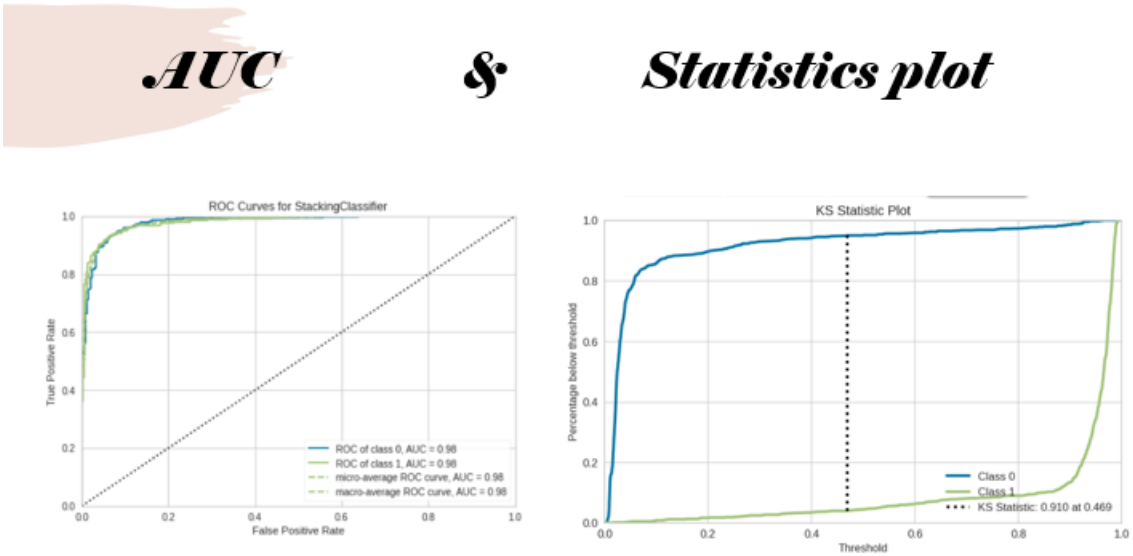


Figure 3.3.1 represents the receiving operating characteristics curve performance measurement at various threshold settings. Upon using the stack classifier algorithm, the area under the curve shows a metric score of 98% which is greater than the threshold set value (50%). Hence, this indicates a great sign of a good model. The statistical plot shows a strong test statistic, and this indicates a statistical significance in our variable.

3.4 Test Data & Prediction

Table 3.4: PREDICTION

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Stacking Classifier	0.9317	0.9779	0.9333	0.9302	0.9318	0.8633	0.8633

	Gene One	Gene Two	Cancer Present	Label	Score
0	7.4	3.4	1	1	0.9723
1	6.3	4.9	1	0	0.7150
2	6.6	2.9	1	1	0.9855
3	6.6	4.5	0	0	0.5858
4	5.4	7.2	0	0	0.9673
5	4.2	9.3	0	0	0.9946
6	6.6	3.7	0	1	0.9577
7	5.4	2.9	1	1	0.9864
8	2.5	4.8	0	0	0.9748
9	6.2	6.5	0	0	0.9600
10	6.7	6.1	0	0	0.9243
11	6.1	7.7	0	0	0.9788
12	4.6	7.3	0	0	0.9726
13	5.4	5.1	1	1	0.8802
14	5.0	4.6	1	1	0.9528
15	2.7	6.2	0	0	0.9855
16	3.9	8.8	0	0	0.9938
17	4.1	4.0	1	1	0.9591

CHAPTER 4

RESULT, CONCLUSION & RECOMMENDATION

Based on the results obtained from this research work, the presence of certain gene can be used in predicting or determining the presence of Cancer in a patient's body.

From the results displayed above, it can be concluded that Stacking Classifier with GradientBoostClassifer as the main estimator coupled with other base learners came up with great accuracy score for this project work. A feature importance extraction is necessary to be investigated on the GBC for further statistical analysis in determining the relative recurrence of cancer spreads on a particular genotype. The advanced study of black box algorithms about explainability and interpretability could help reduced 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. They divided 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 a disease. 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 relationship of variable interactions was used, however, an advanced study with "causal inference" could vividly aid in identifying the true parametric variable that could cause a change on the ecosystem of the cancer existence.

REFERENCE

(1) D. Hanahan, R.A. Weinberg
Hallmarks of cancer: the next generation J.A. Cruz, D.S. Wishart

(2) Applications of machine learning in cancer prediction and prognosis
Cancer Informat, 2 (2006), p. 59 S. Michiels, S. Koscielny, C. Hill

(3) *Prediction of cancer outcome with microarrays: a multiple random validation strategy*
Lancet, 365 (2005), pp. 488-492 I.H. Witten, E. Frank

(4) *Data mining: practical machine learning tools and techniques* Morgan Kaufmann (2005)
K. Park, A. Ali, D. Kim, Y. An, M. Kim, H. Shin

(5) *Robust predictive model for evaluating breast cancer survivability* Engl Appl Artif Intell, 26 (2013), pp. 2194-2205 A.T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A.R. Razavi, L.G. Ahmad