|  |  |
| --- | --- |
| **000** | **054** |
| **001** | **055** |
| **002** | **056** |

**003**

**004**

**005**

ML Algorithms for Breast Cancer Prediction 057

**058**

**059**

**006** Tanishq Gautam

|  |  |
| --- | --- |
| **007** | **061** |
| **008** | **062** |
| **009** | **063** |
| **010** | **064** |
| **011** | **065** |

**060**

**012**

**013**

**014**

**015**

**016**

**017**

**018**

**019**

**020**

**021**

**022**

**023**

**024**

**025**

**026**

**027**

# 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 im- portant underlying importance of attributes thereby helping in predicting the stage of breast cancer depending on the values of these attributes.* ***Github*** *:-* [*Link*](https://github.com/tgautam98/CS584-Machine-Learning-Project)

# Introduction

|  |  |  |
| --- | --- | --- |
| **028** | Mangasarian gave it away. Dr. Wolberg used fluid sam- | **082** |
| **029** Of the 184 major countries in the world, breast cancer is | ples obtained from patients with solid breast masses[10] and | **083** |
| **030** the most common cancer diagnosis in women in 140 coun- | Xcyt, a user-friendly graphic computer program that can do | **084** |
| **031** tries (76%) and the most frequent cause of cancer mortality | the analysis of cytological features based on a digital scan, | **085** |
| **032** in 101 countries (55%). Breast cancer ratios are statisti- | to construct the dataset. | **086** |
| **033** cally higher in women in more developed countries as com- | Ten real-valued features are computed for each cell | **087** |
| **034** pared to other diseases. But it is also globally increasing | nucleus: • radius (mean of distances from center to points | **088** |
| **035** day by day. Breast cancer is a malignant cell growth in the | on the perimeter) • texture (standard deviation of gray-scale | **089** |
| **036** breast. If left untreated, the cancer spreads to other areas of | values) • perimeter • area • smoothness (local variation in | **090** |
| **037** the body. Excluding skin cancer, breast cancer is the most | radius lengths) • compactness (perimeter2ˆ / area - 1.0) • | **091** |
| **038** common type of cancer in women in the United States, ac- | concavity (severity of concave portions of the contour) • | **092** |
| **039** counting for one of every three cancer diagnoses. Breast | concave points (number of concave portions of the contour) | **093** |
| **040** cancer ranks second among cancer deaths in women. | * symmetry • fractal dimension (”coastline approximation” | **094** |
| **041** Breast cancer prediction has been the subject of exten- | - 1) | **095** |
| **042** sive study. Below are some notable studies that have been |  | **096** |
| **43** conducted: - The utilization of current technology break- | The mean, standard error and ”worst” or largest (mean | **097** |
| **44** throughs to create breast can- cer prediction models is dis- | of the three largest values) of these features were computed | **098** |
| **045** cussed by Chaurasia V. He develops a prediction model us- | for each image, resulting in 30 features. For instance, field | **099** |
| **046** ing Naive Bayes, RBF Network, and J48 and employs the | 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst | **100** |
| **047** 10-fold cross-validation approach to measure the unbiased | Radius. All feature values are recorded with four significant | **101** |
| **048** estimate of these models’ performance. Verma D used five | digits. | **102** |
| **049** classification algorithms Naive bayes, SMO, REP Tree, J48 |  | **103** |
| 1. and MLP upon breast cancer dataset. Ojha. U. places em- 2. phasis on the choice of parameters for estimating the like- | **3. Experiments** | **104**  **105** |
| **52** lihood of breast cancer recurrence using data mining tech- | When we calculate mean, variance, standard deviation, | **106** |
| **53** niques. He shows how clustering and classification meth- | number of sample (count) or max min values, it helps us | **107** |

ods are used. According to the author, for the experimen- tal 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, ac- cording to B.L. Rodrigues. The majority of studies appear to only achieve accuracy rates of 93– 94

# Dataset

The dataset utilized in this study was produced by Dr. William H. Wolberg, a physician at the University of Wis- consin Hospital in Madison, Wisconsin, in the United States, and is openly accessible. On July 15, 1992, Olvi

**066**

**067**

**068**

**069**

**070**

**071**

**072**

**073**

**074**

**075**

**076**

**077**

**078**

**079**

**080**

**081**

**108**

**109**

**110**

**111**

**112**

**113**

**114**

**115**

**116**

**117**

**118**

**119**

**120**

**121**

**122**

**123**

**124**

**125**

**126**

**127**

**128**

**129**

**130**

**131**

**132**

**133**

**134**

**135**

**136**

**137**

**138**

**139**

**140**

**141**

**142**

**143**

**144**

**145**

**146**

**147**

**148**

**149**

**150**

**151**

**152**

**153**

**154**

**155**

**156**

**157**

**158**

**159**

**160**

**161**

to understand what is going on with the data. We do need standardization or normalization before visualization, fea- ture 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 visu- alization, feature selection, feature extraction or classifica- tion

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 at- tributes 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 miss- ing characteristics (all of them are bare nuclei size).

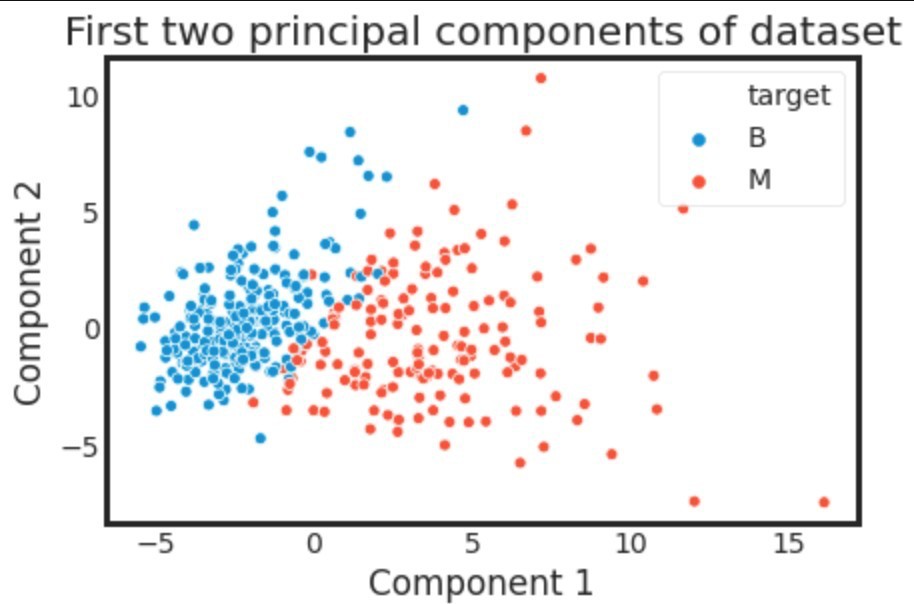
# Methodology

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

## Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a linear dimension- ality 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 re- move 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 simi- lar data points in the training data, and making an educated guess based on their classifications. K is number of near- est neighbors that the classifier will use to make its predic- tion. 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 pa- rameters 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 fol- lows Gaussian normal distribution and supports continuous data. We have explored the idea behind Gaussian Naive Bayes along with an example. Gaussian Naive Bayes sup- ports continuous valued features and models each as con- forming 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 stan- dard 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 pa- rameters 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 pa- rameters of PCA and LR for our dataset.

Using this we get pca n components = 8 and LR C value

**162**

**163**

**164**

**165**

**166**

**167**

**168**

**169**

**170**

**171**

**172**

**173**

**174**

**175**

**176**

**177**

**178**

**179**

**180**

**181**

**182**

**183**

**184**

**185**

**186**

**187**

**188**

**189**

**190**

**191**

**192**

**193**

**194**

**195**

**196**

**197**

**198**

**199**

**200**

**201**

**202**

**203**

**204**

**205**

**206**

**207**

**208**

**209**

**210**

**211**

**212**

**213**

**214**

**215**

**216**

**217**

**218**

**219**

**220**

**221**

**222**

**223**

**224**

**225**

**226**

**227**

**228**

**229**

**230**

**231**

**232**

**233**

**234**

**235**

**236**

**237**

**238**

**239**

**240**

**241**

**242**

**243**

**244**

**245**

**246**

**247**

**248**

**249**

**250**

**251**

**252**

**253**

**254**

**255**

**256**

**257**

**258**

**259**

**260**

**261**

**262**

**263**

**264**

**265**

**266**

**267**

**268**

**269**

= 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 pre- dictive models for both classification and regression prob- lems. Ensemble methods use multiple learning models to gain better predictive results - in the case of a random For- est, the model creates an entire forest of random uncorre- lated decision trees to arrive at the best possible answer.

The random Forest starts with a standard machine learn- ing 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 mar- gin 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 frame- work. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outper- form all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, deci- sion 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 gra- dient descent architecture.

We ran a XGBoost pipeline with standard scaler. We then implemented GridSearchCV which returns the best pa- rameters 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 prob- lem, 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 uncorre- lated 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, XG-

Boost. 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%.

# Conclusion

This project investigates different models for breast can- cer prediction. Three different types of Machine Learning methods including Random Forest, Support Vector Classi- fier (SVM), XGBoost are compared and analyzed for opti- mal 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

**270**

**271**

**272**

**273**

**274**

**275**

**276**

**277**

**278**

**279**

**280**

**281**

**282**

**283**

**284**

**285**

**286**

**287**

**288**

**289**

**290**

**291**

**292**

**293**

**294**

**295**

**296**

**297**

**298**

**299**

**300**

**301**

**302**

**303**

**304**

**305**

**306**

**307**

**308**

**309**

**310**

**311**

**312**

**313**

**314**

**315**

**316**

**317**

**318**

**319**

**320**

**321**

**322**

**323**

**324**

**325**

**326**

**327**

**328**

**329**

**330**

**331**

**332**

**333**

**334**

**335**

**336**

**337**

**338**

**339**

**340**

**341**

**342**

**343**

**344**

**345**

**346**

**347**

**348**

**349**

**350**

**351**

**352**

**353**

**354**

**355**

**356**

**357**

**358**

**359**

**360**

**361**

**362**

**363**

**364**

**365**

**366**

**367**

**368**

**369**

**370**

**371**

**372**

**373**

**374**

**375**

**376**

**377**

did not give good test performance. SVM with best param- eters gave the highest test performance and also an equally comparable training performance as XGBoost. For Random Forest with univariate feature selection, we observed a bet- ter accuracy than implementing the feature selection with correlation. Default data included 33 features but after fea- ture 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 tech- niques. 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 a˜ 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

**378**

**379**

**380**

**381**

**382**

**383**

**384**

**385**

**386**

**387**

**388**

**389**

**390**

**391**

**392**

**393**

**394**

**395**

**396**

**397**

**398**

**399**

**400**

**401**

**402**

**403**

**404**

**405**

**406**

**407**

**408**

**409**

**410**

**411**

**412**

**413**

**414**

**415**

**416**

**417**

**418**

**419**

**420**

**421**

**422**

**423**

**424**

**425**

**426**

**427**

**428**

**429**

**430**

**431**