Developing Breast Cancer Detection Classifier Using Various Machine Learning Methods

Hopkins Authentic Research Program in Science: B Block, Hopkins School
Albert Yang
May 21, 2021

Abstract:

Breast cancer (BC) is one of the most prevalent and deadly cancers. There is a 1 in 8 chance that a woman in the United States will develop BC and a 1 in 39 chance to die from it (1). Typical diagnosis involves mammograms, x-ray pictures of the breast done by radiologists and sent to oncologists for further examination. Results are usually given within two weeks but can take up to or even longer than 30 days in some cases (2). This delay could provide the tumor an adequate amount of time to grow and inflict greater risk on the patient. One in five women is also misdiagnosed from the mammogram, whether being a false positive or false negative. The advent of machine learning (ML) may provide an avenue for quicker, more accurate results in aiding medical staff. In this study, I developed a breast cancer detection classifier using data from the University of Wisconsin Hospitals, Madison, from Dr. William H. Wolberg (3). Six different supervised machine learning techniques were tested to find the optimal accuracy: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-nearest neighbor (KNN), Support Vector Machines (SVM), and Gaussian Naive Bayes (GNB). The performance of the study was measured with respect to the radius, texture, perimeter, area, and smoothness of the tumor. Results show that the RF provided the best accuracy of 93.71% in predicting whether a tumor was malignant or benign. GNB was second with an accuracy of 93.01%, LR with 92.31%, KNN with 91.61%, DT with 89.51%, and SVM in last at 88.11%. Improving the accuracy of the classifier can lead to widespread clinical use and better detection and treatment for patients.

Introduction:

Cancer is a disease where abnormal cells divide at uncontrollable rates, damaging body tissue as a result (4). It is the second leading cause of mortality in the United States, with an estimated 1.9 million new cases and 608,570 deaths in 2021 (5). Since cancer can occur within almost any cell in the human body, there are over 100 different types. Breast cancer is among the most common, in which cells in the breast grow out of control. It occurs in women and rarely men. Early symptoms include a lump in your breast or underarm, bloody discharge from the nipple, and changes in shape and texture of the nipple and breast (6). Early localized detection of BC leaves the patient with a near 100% survival rate; however, if the cancer becomes regional or distant, the five-year survival rate falls to approximately 86% and 28%, respectively (7). Thus, prompt detection of BC can lead to early treatment and be life-saving.

Machine learning is the application of artificial intelligence (AI) that provides systems the ability to improve automatically through experience and by the use of data instead of being explicitly programmed. This ability for models to independently adapt to new data makes ML a real avenue for future research and discoveries. In ML, classification refers to a model's ability to detect which category a given input belongs to, an everyday example being spam email detection. For this study, I developed a BC classifier that determines if the tumor is malignant or benign based on the radius, texture, perimeter, area, and smoothness of the tumor. The dataset used is from the University of Wisconsin Hospitals, Madison, from Dr. William H. Wolberg. Data is separated into training and testing groups to help develop the models.

Supervised learning (SL) is a ML technique that matches an input to an output based on a set of training data and examples. Six SL algorithms were used in this study: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Gaussian Naive Bayes (GNB). The goal is to find the accuracy of each and also create a confusion matrix (CM), a table that displays the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). TP represents correctly classified malignant tumors, while a TN is a correctly classified benign tumor. FP occurs when the classifier returns malignant for a benign tumor, and FN occurs when benign is returned for a malignant tumor. The CM helps us confirm the accuracy of our classifier.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

LR determines the probability, ranging from 0 to 1, of a data point belonging to a class. Based on this value and the given classification threshold, the data point is assigned to the more probable class, thus making LR a useful tool in solving classification problems. Where points sit on the curve can determine if a tumor will be classified as malignant or benign. However, LR struggles to solve nonlinear problems, as this algorithm relies on the linearity between the dependent and independent variables.

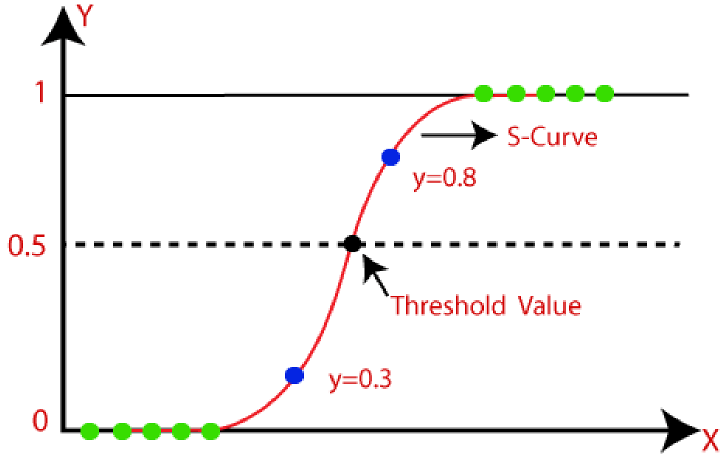


Fig. 1: Logistic regression graph for better visualization (8)

DT are flow chart-like structures containing nodes and branches that repeatedly split data into smaller groups based on a feature. The DT recursively evaluates how well it is able to split the data that comes into its separate categories. Each node gets updated when the model runs until it can accurately split the data into its proper classes. DT are easy to visualize but typically suffer from high variance.

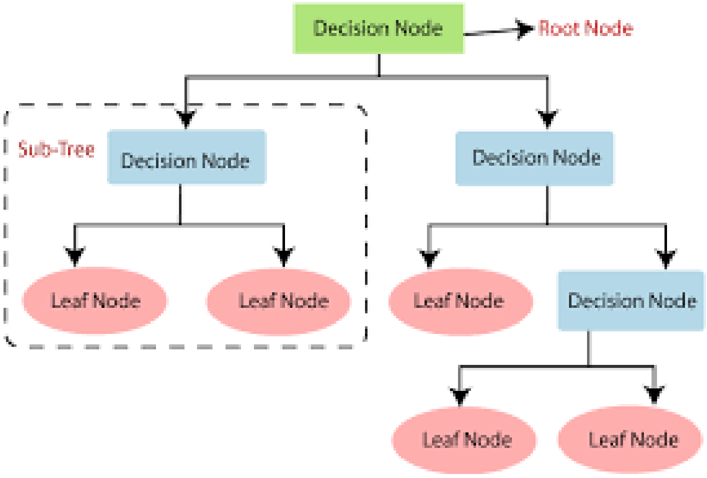


Fig. 2: Decision Tree depiction and labeling of key parts (9)

RF is made up of many DT all working together to classify new points. Each tree reports its classification, and the RF returns the most popular. Having multiple DT cast a vote is effective as it ensures that different patterns of classification arrive at the same decision. Though overfitting may still have an effect, a large number of trees will limit its impact.

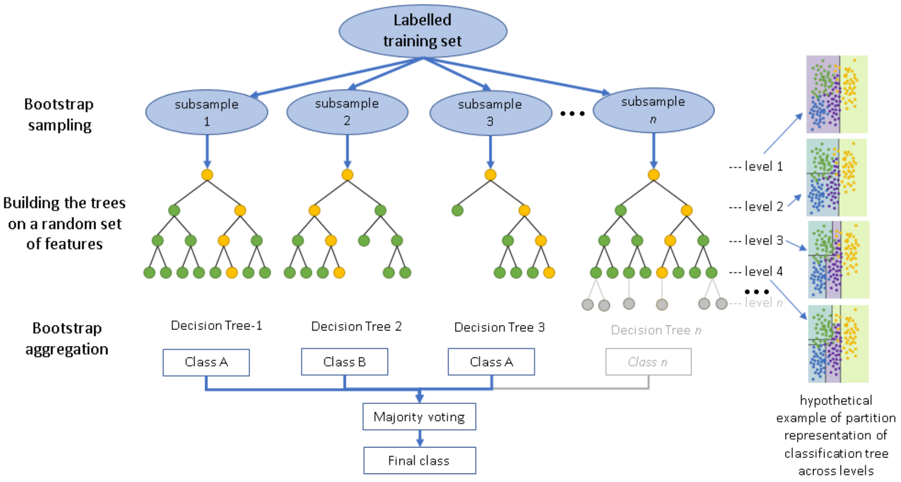


Fig. 3: RF diagram for better understanding (10)

KNN looks at the proximity of similar data points. A new datapoint will be classified based on what the closest points around it have already been determined as. However, the algorithm is contingent on the user determining how many points to analyze around the one that needs to be classified, allowing for possible manual errors.

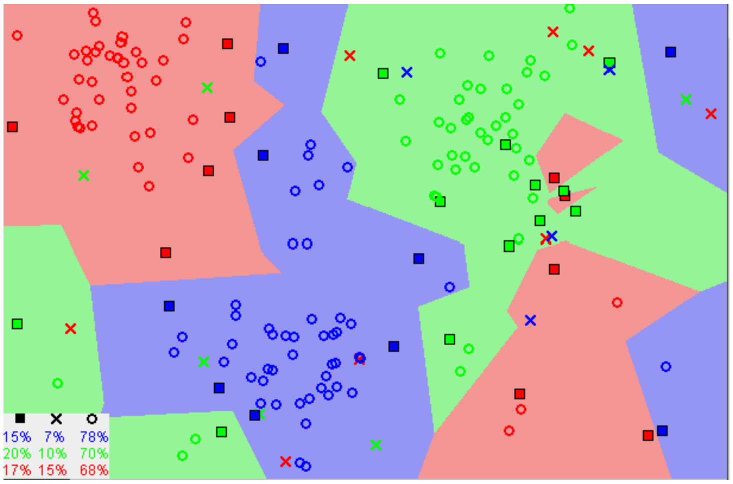


Fig 4: KNN depiction of proximity classification (11)

SVM is based on finding a hyperplane that divides the data into two classes. A hyperplane is a line that separates the data, and whatever side data points lie on will determine the classification. The distance between the hyperplane and the nearest point from either set is called the margin. The goal is to maximize the margin and any point within the training set, giving a greater chance of new data being classified correctly. Problems may arise when a line can not clearly split the data, and there are many outliers.

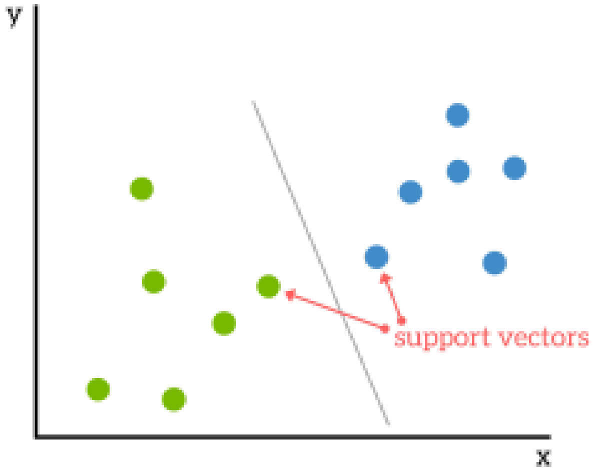


Fig. 5: SVM diagram with hyperplane (12)

GNB is a classification technique based on Bayes Theorem: the probability of A given B is true is equal to the probability of B given A is true times the probability of A over B. It uses this algorithm to give the likelihood of given class labels to belong to a class. However, GNB assumes that a particular feature in a class is independent of other features. Even if features are interdependent, they are still considered independently, an assumption that simplifies computations and explains the naive nature.

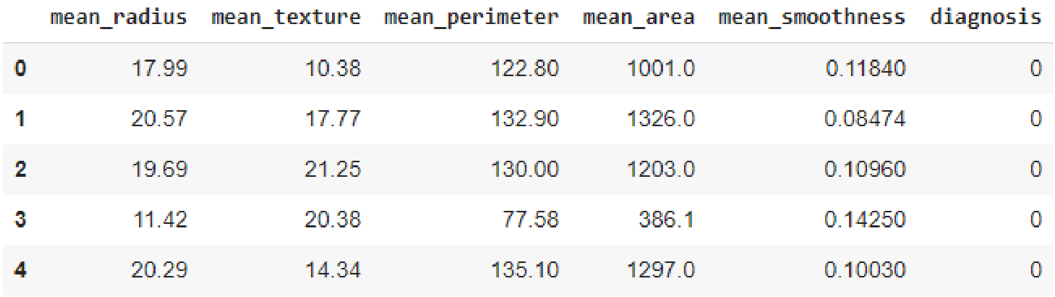
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Material and Methods:

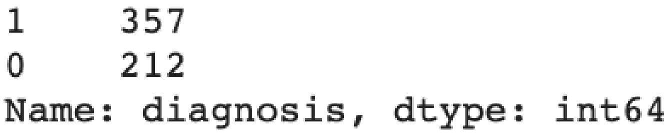
All code was done in python and within the Google Colab notebook. The first step to building our classifier is loading in the necessary libraries, including numpy as np, pandas as pd, matplotlib.pyplot as pt, and seaborn as sns. Next, we must load the data, which is stored in a file called "Breast_cancer_data.csv." In order to analyze the data set, we can use the .shape and .value_counts() functions to find how many tumors are in this set and the number of malignant and benign ones there are. This data can be visualized with the sns.countplot() command. It is also important to check for any empty columns or rows through the .isna().sum() function. It is crucial that the data is cleaned to ensure optimal conditions for training and testing. Next, using sns.pairplot and sns.heatmap, we can visualize the correlation between variables by creating a pairplot chart and heatmap. After taking a look at the data and the correlations, we can split the data into a training set and testing set. I've decided to go with the conventional 0.75, 0.25 split, respectively. Next, we fit data to each of our models and find the accuracy along with the CM. With the accuracies and CM for each model, we can try to understand which is superior in breast cancer detection, why, and the clinical implications.

Results:

The data contains the mean radius, texture, perimeter, area, and smoothness of each patient's tumor and their respective diagnosis, either 1, malignant, or 0, benign. Of the 569 tumors in this study, 357 are malignant, and 212 are benign.



	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
0	17.99	10.38	122.80	1001.0	0.11840	0
1	20.57	17.77	132.90	1326.0	0.08474	0
2	19.69	21.25	130.00	1203.0	0.10960	0
3	11.42	20.38	77.58	386.1	0.14250	0
4	20.29	14.34	135.10	1297.0	0.10030	0



```
1    357
0    212
Name: diagnosis, dtype: int64
```

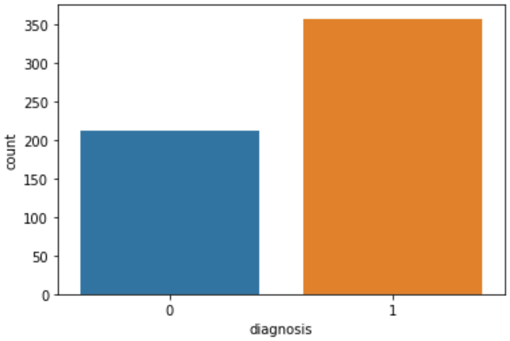


Fig. 6: 5 rows of the data table, number of malignant and benign tumors, visualization of said numbers

The pairplot shows the correlation between the different measurable variables, and it is clear that "mean_area," "mean_radius," and "mean_perimeter" are closely related to one another. Benign tumors are typically greater in size and are smoother, as well. The heatmap also shows how the mean area, radius, and perimeter are related, having correlations of 0.99 and 1.

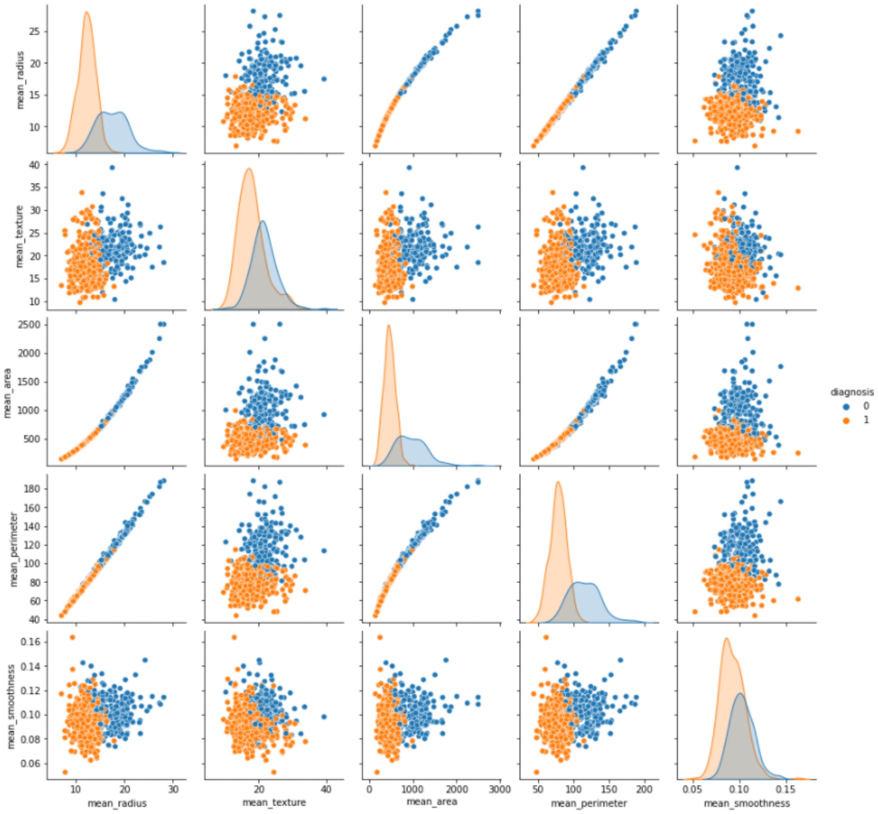


Fig. 7: Pairplot graphs between the different measurable variables

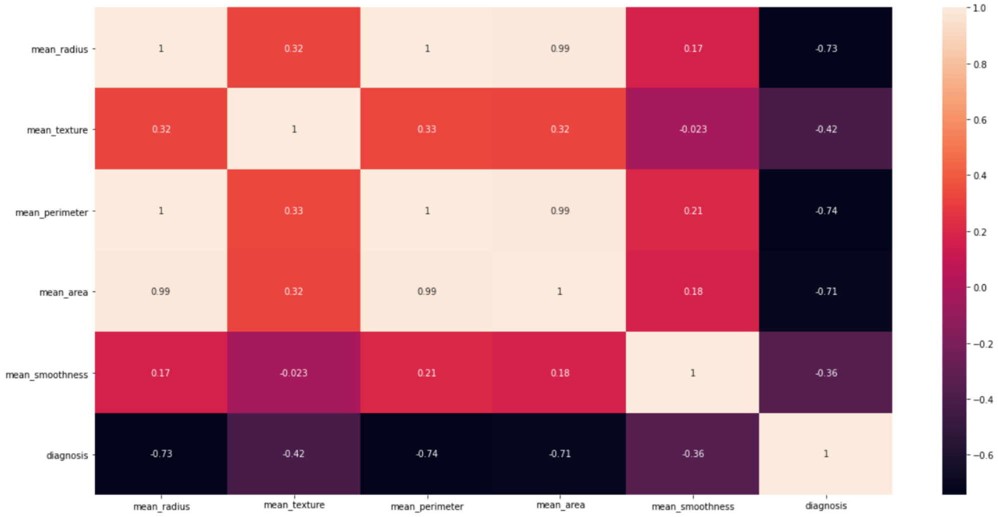
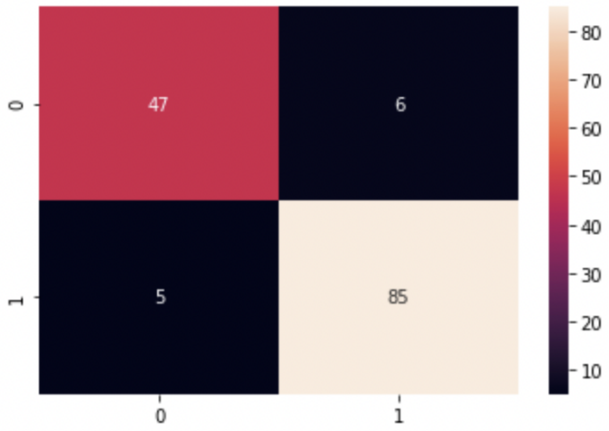


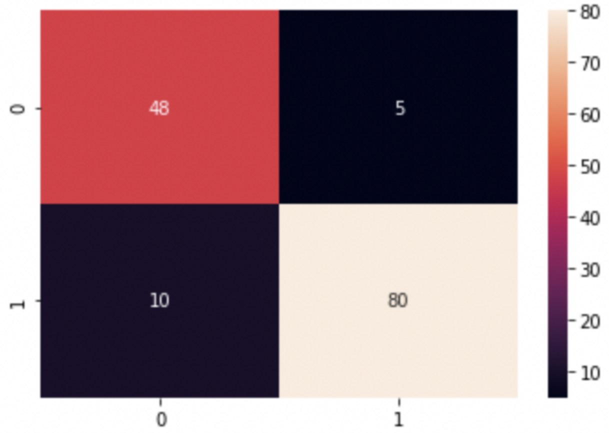
Fig. 8: Heatmap of different measurable variables and their correlations to one another

The LR model produced an accuracy of 92.31%, predicting 47 TN, 85 TP, 5 FN, and 6 FP. The DT had an accuracy of 89.51%, with 48 TN, 80 TP, 10 FN, and 5 FP. The RF classifier was the most accurate at 93.71% and had 49 TN, 85 TP, 5 FN, 4 FP. The KNN had 45 TN, 86 TP, 4 FN, 8 FP and 91.6%. This was higher than the SVM at 88.11% with 38 TN, 88 TP, 2 FN, 15 FP, but lower than GNB with 46 TN, 87 TP, 3 FN, 7 FP at 93.01%.

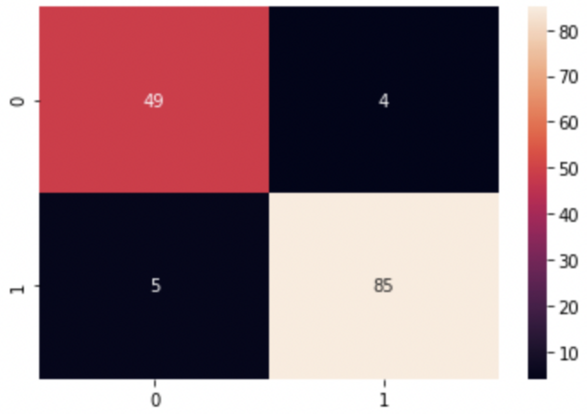
```
Accuracy score using Logistic Regression: 92.3076923076923
```

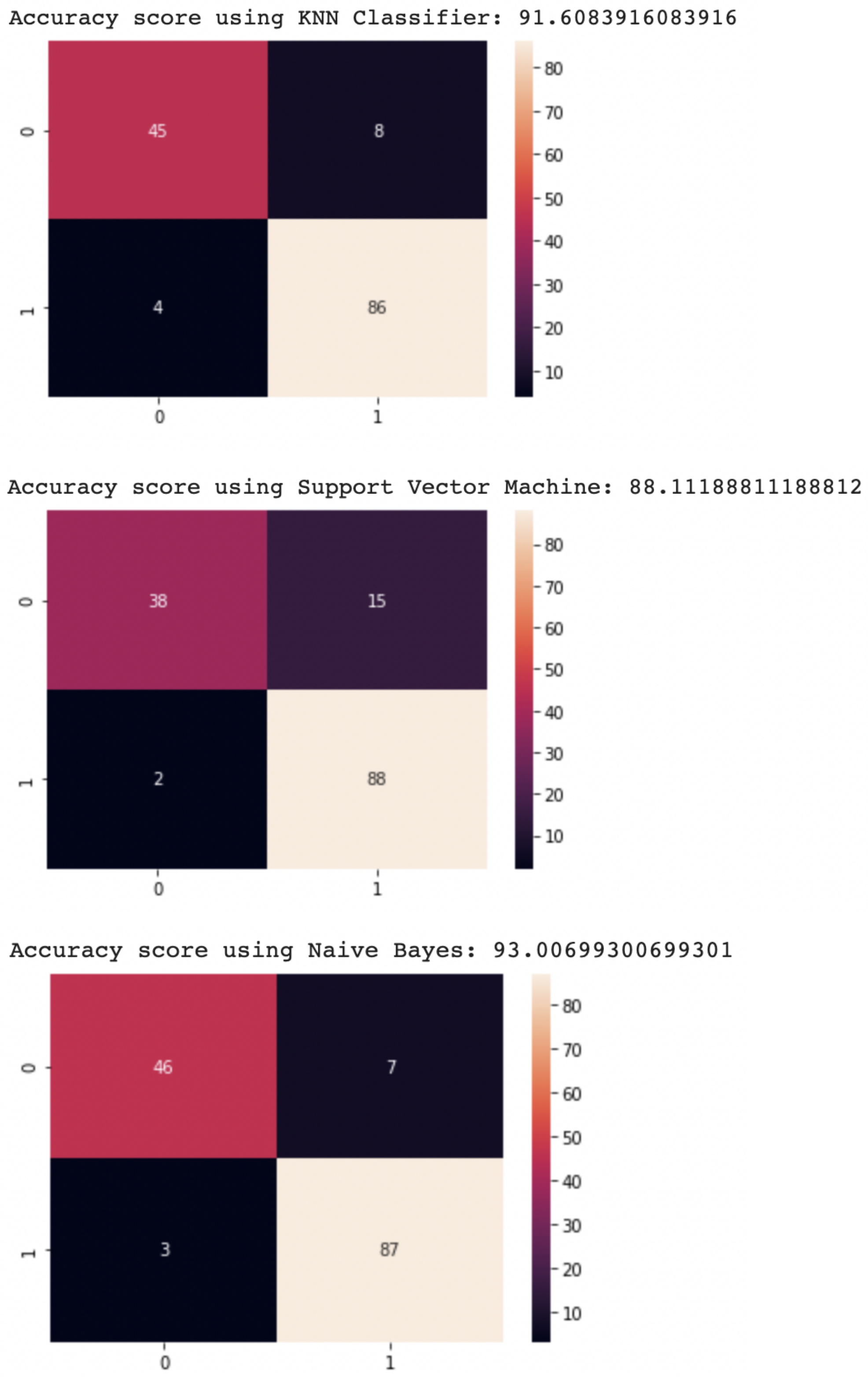


```
Accuracy score using Decision Tree Classifier: 89.5104895104895
```



```
Accuracy score using Random Forest Classifier: 93.7062937062937
```





```
Accuracy score using KNN Classifier: 91.6083916083916
```

```
Accuracy score using Support Vector Machine: 88.11188811188812
```

```
Accuracy score using Naive Bayes: 93.00699300699301
```

Fig. 8: Accuracy and CM of LR, DT, RF, KNN, SVM, and GNB classifiers

Discussion:

The performance of six different supervised ML-based classification algorithms was tested: LR, DT, RF, KNN, SVM, and GNB. The results of this study help us better understand the overall effectiveness of using ML and which algorithms are best suited for BC classification. BC is a common and deadly disease, allowing for limited room for error when detecting. Conventional detection uses mammograms which have a two-week turnaround time, if not 30 days in some cases, significantly longer than the time it would take for a computer with a given ML model to produce results. Mammograms are 80 to 98 percent effective in detecting BC in women with non-dense breast tissue. However, this percentage falls to around 50 with women who have dense breast tissue (13). If the models we developed can exceed these percentages and reliably provide accurate results, the prospect of standardized clinical use and practice is not out of reach.

In this study, RF proved to be the best model with an accuracy of 93.71%, which to no surprise was higher than the 89.51% from DT. Random forests are typically more accurate since they combine multiple trees, limiting overfitting in the process. The advantage that DT has is that they are easier to interpret and understand since a sole tree is not as complicated as multiple. LR, with an accuracy of 92.31%, performs relatively well due to the use of percentages and odds in mapping data points, leading to less bias and lower variance results. However, LR is dependent on what classification threshold is determined, affecting the potential accuracy. The KNN provided an accuracy of 91.61% but relied on determining the amount of nearest neighbors to take into account. The standard is five, but another value could have provided even better results. SVM was the least accurate at 88.11% meaning the data points could not be split effectively by the hyperplane. A different dataset with easier split clusters would help. GNB had a higher accuracy of 93.01% due to the algorithm's ability to compute conditional probabilities. The model is easy to estimate due to only needing the mean and standard deviation.

Though the models performed well, an accuracy of around 99% would be ideal for clinical use. Accuracy was inherently limited by the data set used, which only had 569 different patients whose tumors could be analyzed; however, a larger sample size typically leads to better results. The number of parameters tested could also be expanded. More can be measured than just mean radius, texture, perimeter, area, and smoothness. Examples that come to mind are compactness, symmetry, and concavity. Taking greater factors into account can help increase accuracy. For future studies, more supervised ML techniques could have been used. Unsupervised, deep learning techniques could even be tested. These methods do not rely on a given training and data set; rather, the model is fully independent in discovering patterns and information and classifying data.

Conclusion:

RF was the most accurate model tested, with an accuracy of 93.71%. GNB came in second with 93.01%, followed by LR with 92.31%, KNN with 91.61%, DT with 89.51%, and SVM in last at 88.11%. The final results help show that RF is the superior choice, comparatively speaking; however, none of these percentages are high enough for reliable clinical use, as accuracy of 99% is ideal. Testing a model on a larger data set can help determine whether the model needs finite tweaking. Other supervised and unsupervised methods could be used to optimize the accuracy of the classifier. For example, Google developed a deep learning tool to detect metastasized breast cancer with an accuracy of 99% (14). Optimizing the classifier has the potential to assist more BC patients by prompting them to the right treatment through early and accurate detection.

Acknowledgments:

I would like to thank Dr. E, Mr. Taylor, and Ms. Abraham for providing me with a Codecademy pro account where I was able to learn the basics of ML in python and for their tremendous guidance throughout this difficult year.

References/Bibliography:

1. American Cancer Society: How Common Is Breast Cancer? https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=Current%20year%20estimates%20for%20breast%20cancer&text=About%20281%2C550%20new%20cases%20of,will%20die%20from%20breast%20cancer

2. CDC: What Is a Mammogram? https://www.cdc.gov/cancer/breast/basic_info/mammograms.htm

3. Kaggle Dataset: University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. https://www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset

4. CDC: Leading Causes of Death https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm

5. CDC: Cancer Facts and Figures 2021 https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2021.html#:~:text=Estimated%20numbers%20of%20new%20cancer,deaths%20in%20the%20United%20States.)

6. American Cancer Society: Breast Cancer Signs and Symptoms https://www.cancer.org/cancer/breast-cancer/about/breast-cancer-signs-and-symptoms.html

7. American Cancer Society: Survival Rates for Breast Cancer https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html

8. Javatpoint: LR Diagram https://www.javatpoint.com/logistic-regression-in-machine-learning

9. Decision Trees, Inductive Bias and Hyperparameters: DT Diagram https://www.niser.ac.in/~smishra/teach/cs460/lectures/lec3/

10. Breiman, Leo. 2001. Random Forests. Machine Learning. Vol-45, p.5-32: RF Diagram https://www.pcigeomatics.com/geomatica-help/concepts/focus_c/oa_classif_intro_rt.html

11. Harrison, Onel. 2018. Towards Data Science: Machine Learning Basics with the K-Nearest Neighbors Algorithm https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

12. Bambick, Noel . KDNuggets: SVM Diagram https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

13. Yale Medicine: Dense Mammogram Statistics https://www.yalemedicine.org/conditions/dense-breasts#:~:text=Research%20shows%20that%20mammograms%20can,women%20with%20dense%20breast%20tissue.

14. Jessica Kent. Health Analytics: Google Deep Learning Tool 99% Accurate at Breast Cancer Detection https://healthitanalytics.com/news/google-deep-learning-tool-99-accurate-at-breast-cancer-detection