# Cancer Prediction and Analysis Using Supervised Machine Learning Algorithms

**Hany Raza, Sriya Venkatesh**
Northeastern University
raza.h@northeastern.edu, venkatesh.s@northeastern.edu

**Abstract**

Breast cancer is the most common type of cancer faced by women worldwide. Accurate and early diagnosis is crucial in treatment of breast cancer. The proposed system in this paper will be used to predict if the tumor of each patient is "Benign" or "Malignant" using data visualization and various supervised machine learning techniques. The supervised machine learning techniques used in this system, include logistic regression, naïve Bayes, support vector machine, k-nearest neighbors, decision tree, random forest. Python was used to create this system. In this paper, a comparative analysis on each ML model's ability to predict if the tumor was "malignant" vs " benign" was performed. It was found that Random Forest showed the highest classification accuracy of 95.32%. As a result of these findings, it is evident that new methods for breast cancer screening could be opened. Machine learning techniques can impact cancer detection and provide significant benefits during the classification and decision-making process.

## Introduction

Breast cancer is the most prevalent form of cancer in women and one of the leading causes of death in women. Breast cancer screening and diagnosis at an early stage will greatly minimize the increasing number of deaths. Breast cancer is the fifth leading cause of death worldwide, according to Globocan 2018 statistics, and it accounts for one out of every four cancer cases diagnosed in women worldwide. It was also discovered that the death rate due to breast cancer was 6.8 per 100,000, and the age of incidence of breast cancer was 23.7 per 100,000 in 2018.

Cancer is caused mostly by mutations or modifications to DNA or RNA, which allow cells to develop into cancer cells. In this paper, we used public data about breast cancer tumors from Breast Cancer Wisconsin (Diagnostic) Data Set. The primary goal of this research is to develop a Healthcare-appropriate model by disclosing the predictive variables of early-stage breast cancer patients from a broader viewpoint and comparing the model's strength with accuracy tests. The following are the key contributions of this paper:

- Creating an appropriate model by disclosing the predictive variables of early-stage breast cancer patients from a wider context and comparing the model's robustness using consistency tests.
- To test the model, a more detailed comparison and analysis using data visualization and machine learning tools for breast cancer diagnosis and visibility is performed.
- Observe which characteristics are most important in forecasting breast cancer and learn about general patterns.

The database used for breast cancer stage estimation in this article is called "Breast Cancer Wisconsin (Diagnostic) Data Set." When analyzing the dataset properties, the two groups in this dataset that are used to predict breast cancer are "Malignant (M)" and "Benign (B)." The alternative features reflect various aspects of breast cancer incidence that can be used to categorize whether a particular condition induces breast cancer or not. The feature "Diagnosis" is used to predict the stages are 0(B) and 1(M) values, where 0 represents "benign" and 1 represents "malignant."

## Background

This paper describes the application of various supervised machine learning techniques and optimization techniques to the task of classifying breast cancer biopsy data into two binary classes: "malignant" or "benign". The following parts include a thorough review of the supervised machine learning methods used in this paper.

## Support Vector Machine

Support Vector Machines's goal is to find an N-dimensional hyperplane that can classify data points. The maximal margin hyperplane in the p-dimensional feature space that distinguishes the classes is the most important aspect of this classification technique.

SVMs seek a hyperplane $w.x + b = 0, x_i \in R^n$ that divides the data points $x_i$ and corresponds to a given decision rule: $g(x) = sign(w.x + b)$. SVMs choose the dividing hyperplane $w.x + b = 0$ with the greatest margin that is farthest away from the data points.

The fundamental theory is that placing a hyperplane far away from any observed data points reduces the possibility of making incorrect assumptions when classifying new data. The path to the nearest data points are optimized. With SVM, it is possible to compare two features without difficulty. When there are many attributes to classify, SVM is not always that simple. When the margin is maximized, the results predicted are more reliable. Various Kernel Functions are used in Support Vector Machines to project the non-linear non-separable input space into a higher dimensional linear separable space. The Python scikit-learn library's SVC function was used to match SVM models to our oversampled training data.

## Naïve Bayes Algorithm

The naïve Bayes algorithm is a simple and fast classification algorithm. It is represented by:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Its operation is based on Bayes' theorem. The basic assumptions of this algorithm are that each variable contributes equally and independently to the output.

Each feature will be independent of the others and will have the same impact on the performance. As a result, the naïve Bayes theorem does not extend to real-world problems, and using this algorithm will result in low accuracies.

One kind of naïve Bayes implementation is Gaussian naïve Bayes:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma^2 y}} exp\, \frac{-(x_i - \mu y)^2}{2\sigma^{2y}}$$

It is presumptively assumed that features have a natural distribution. The probability of features being present is Gaussian and it follows a conditional probability.

## Logistic Regression

Logistic regression is a method that was developed in the early twentieth century for biological research. It has also been common in social studies. Logistic regression is another form of predictive analysis. When there is a single binary dependent variable and multiple independent variables, logistic regression is appropriate. Linear and logistic regressions vary in terms of the dependent variable. For continuous variables, linear regression is a better method.

Logistic regression is divided into two stages: forward and backward propagation. Weights are multiplied with features in the first step of forward propagation. Since the weights are unspecified at first, randomized values may be allocated. A probability between 0 and 1 is generated by the sigmoid function. The function is given by:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The prediction is carried out in accordance with a threshold value. Following prediction, the predicted value is applied to the actual values, and a loss function is calculated. The loss function shows how much the expected value differs from the actual value. Backward propagation is used where the loss function value is significant. Backward propagation's goal is to update weight values based on the cost function by taking the derivative.

## Decision Tree

One of the most commonly used supervised learning techniques is the decision tree (DT). The core applications are regression and classification. It tries to solve problems by constructing a tree. The features are labeled as decision nodes, and the outputs are known as leaf nodes. The decision tree algorithm recognizes feature values as categorical.

At the start of this algorithm, it is critical to choose the best attribute and position it at the top of the tree figure before splitting the tree. The Gini index and information gain are two tools for function collection. Entropy or randomness can be calculated by:

$$H(x) = E_x[I(x)] = -\sum p(x) log\, p(x)$$

For each variable Entropy values are estimated, and information values are computed by subtracting these values from one. A higher information gain improves an attribute and elevates it to the top of the tree. The Gini index is a measure of how often a randomly selected variable is incorrectly detected. As a result, a lower Gini index value indicates improved attributes. Gini index is given by:

$$G = \sum p_i * (i - p_i)\ for\ i = 1...n$$

A decision tree is simple to comprehend. However, if the data contains a variety of features, it can result in overfitting issues. As a result, knowing when to avoid growing trees is critical. Methods for preventing the model from

overfitting: Pre-pruning, which stops rising early but is difficult to select a stopping point, and post-pruning, which is a cross-validation used to verify whether extending the tree can boost or contribute to overfitting.

A DT structure is made up of a root node, a separating node, a decision node, a terminal node, a sub-tree, and a parent node. The DT induction process is divided into two stages: the growth phase and the pruning phase. During the growth process, the training data is recursively partitioned, resulting in a decision tree with a natural "if", "then", and "else" construction that fits conveniently into a program.

### Random Forest

Random forest is a type of ensemble learning model that is used for classification as well as regression. A random forest, in reality, is made up of several decision trees. As a result, in some situations, a random forest is preferable to a decision tree. The rotation forest algorithm works by constructing a classifier based on attribute extraction. The attribute set is divided into K distinct subsets at random. Its aim is to develop classifiers that are both correct and relevant. The biggest challenge in the decision tree is function collection, and there are many approaches to it.

Rather than looking for the most popular feature when separating nodes, random forest looks for the best feature within a random subset of features. It is possible to make it much more unpredictable by using randomized thresholds for each attribute rather than aiming for the optimal one.

Any built-in feature parameters can speed up or improve the accuracy of the model. To improve prediction power, max functions, n estimators, and min sample leaf are used. N jobs and random state are commonly used to make models run faster. In this analysis, n estimators are used to calculate the number of trees to rise, as well as random state parameters, to improve the model's accuracy and speed.

### K-Nearest Neighbor

KNN identifies the labels of the data before making a prediction. It can be used for clustering and regression. K is a numerical value representing the closest neighbors. There is no training process in the KNN algorithm. The Euclidean distance to the k-nearest neighbors is used to make predictions. It is given by:

$$Dis(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2}$$

This approach is used to predict breast cancer in a dataset that already has marks such as malignant and benign.

The classification is based on its closest neighbor's class label and the class labels ("malignant" or "benign") of its neighbors. Using KNN values of 1, 3, and 5, the following class selection of the training sample identification is made:

- Since there is only one square within the inner circle, K = 1 indicates that it is reserved to first-class.
- Since there are two triangles and just one square inside the inner circle, K = 3 indicates it is assigned to the second class.
- K= 5 is allocated to the first class since it contains three squares as opposed to two triangles outside the outer circle.

In general, selecting "smaller values for K" can be noisy and have a greater effect on the outcome. "Larger K values" would result in finer decision limits, which means lower variance but increased bias and computational cost. Cross-validation is one of the easiest ways to select K for a better accuracy score. One method for determining the cross-validation dataset from the training dataset.

Take a small portion of the training dataset and label it a validation dataset, and then use it to test various values of K. This way, we will estimate the mark for each instance in the validation set using with K = 1, K = 2, K = 3, and so on, and then we will look at the value of K gives us the best result on the validation set, and then we will take that value and use it as the final set of our algorithm to minimize the validation or misclassification error. As a result, it is clear that the precision and performance of this algorithm are heavily dependent on the value for "K" or the number of neighbors.

## Related Work

Naive Bayes, support vector machine, and decision trees to characterize a Wisconsin breast cancer dataset, with support vector machine (SVM) producing the highest results with an accuracy score of 96.99 percent (P Aruna et. al. 2011). A Wisconsin breast cancer dataset to compare the success of supervised learning classifiers using naive Bayes, SVM, neural networks, and decision tree approaches. SVM produced the most reliable results, with a score of 96.84 percent, according to the study findings (Chaurasia et. al. 2014). The same results to compare the efficiency of machine learning algorithms such as SVM, decision tree, naive Bayes, and k-nearest neighbors. The thesis aimed to characterize data in terms of quality and efficacy by measuring each algorithm's accuracy, precision, sensitivity, and specificity. The experimental results revealed that SVM had the highest accuracy of 97.13 percent (Asri et. al. 2016).

Investigated Breast cancer data was investigated and prediction was made using 202,932 patient records (Delen et. al. 2005). The dataset was split into two groups: those who survived (93,273) and those who did not survive (109,659), and then the naive Bayes, neural network, and

c4.5 decision tree algorithms were used. The obtained results demonstrated that the c4.5 decision tree outperformed the other strategies.

To obtain the best results for classifying the Diabetic disease dataset, (Ou et. al. 2011) compared naive Bayes, judgment tree, and random tree. Based on the results of this report, nave Bayes was determined to be the best classifier, with 76.3% score. A dependence enhanced naive Bayes classifier and a nave creedal classifier to forecast heart attacks using patient profiles such as age, gender, blood pressure, and blood sugar was analyzed (Srinivas et. al. 2010). According to the findings of the report, naive Bayes produced better results.

Clinical reports from medical intensive care units in their study were used (Bernal et al. 2011). To forecast the decline in patients within the hospital over 24 hours, machine learning methods such as logistic regression, neural networks, decision trees, and k- nearest neighbors were used. Among training results, logistic regression and the k-nearest neighbor (KNN)-5 methodology yielded the highest accuracy ratings. In order to achieve higher precision, parameters must be used rather than the algorithm (Bernal et. al. 2011). The best approach for breast cancer estimation using data mining techniques on a large number of documents was analyzed (Wang et. al. 2015). They used the support vector machine (SVM), the artificial neural network (ANN), the naive Bayes classifier, and the Ada-Boost Tree. The topic of reducing the feature space was addressed, and then Principle Component Analysis (PCA) was used with the aim of reducing the feature space.

(Williams et. al. 2016) conducted research on breast cancer risk modeling using data mining classification techniques. Breast cancer is the most prevalent form of cancer in women in Nigeria. There are few services available to detect breast cancer until it is too late to help. As a result, they wanted to find an effective way to predict breast cancer. In their analysis, they used several data mining techniques: naive Bayes and J48 decision trees.

According to (Nithya et. al. 2011), the biggest issue of breast cancer is labeling the tumour. Cancer has been detected and characterized using computer-aided diagnosis (CAD). Their primary idea was to use data mining techniques to improve breast cancer diagnosis. Bagging, multiboot, random subspace, and multilayer perceptron were used to improve the classification efficiency of naive Bayes, support vector machine-sequential minimal optimization (SVM-SMO), and multilayer perceptron.

Research on breast cancer biopsy predictions based on a mammogram diagnosis was conducted (Oyewola et al. 2018). In their research, they used classifications such as logistic regression (LR), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), random forest (FR), and support vector machine (SVM). SVM, LR, multilayer perceptron, KNN, softmax regression, SVM, and Gated Recurrent Unit (GRU) SVM techniques were used by (Agarap et. al. 2018). A multilayer perceptron generated the most accurate results, with an accuracy score of 99.4 percent. Various machine learning methods for predicting breast cancer cells was throughly analyzed (Westerdijk et. al. 2018). She evaluated the models' efficiency by examining their accuracies, sensitivities, and specificities. We contrasted the accuracy scores of the LR, random forest, SVM, neural network, and ensemble models. The accuracy score can be used to boost breast cancer prediction.

A reliable tool for forecasting eight types of cancer, including breast cancer, lung cancer, and ovarian cancer was analyzed (Vard et al. 2015). They used Particle Swarm Optimization to normalize datasets and computational feature selection approaches to isolate features on a normalized dataset in their study. For classification, they used decision trees, help vector machines, and a multilayer perceptron neural network. The classification of cancer patients' risk categories as low or large was done (Kourou et. al. 2016). To provide a model for cancer risks or patient outcomes, ANN, Bayesian networks (BNs), SVM, and decision tree (DT) techniques were used. According to (Pratiwi et. al. 2018), breast cancer is the leading cause of death in women. To diagnosis breast cancer, machine learning methods were favoured. Java was used to create intelligent breast cancer prediction, demonstrating that all functionalities performed well and quickly.

A rigorous data analytical approach that could be applied to breast cancer datasets was analyzed (Shukla et. al. 2018). The viability of patients and cancers was factored into the model. They identified using the Epidemiology, Surveillance, and End Results (SEER) software, and clustered using the Self-Organizing Map (SOM) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

## Project Description and Experiments

The experiments were carried out in the order of : Exploratory Data Analysis, Handling Data Imbalance, Analysis and comparison of performance. the supervised machine learning methods used in this paper are K-Nearest Neighbors, Support Vector Machine, Random Forest, Decision Tree, Naive Bayes and Logistic Regression techniques were used.

## Exploratory Data Analysis

Exploratory data analysis quantitatively explains or summarizes aspects of a series of data, as well as the method of compressing main features of the dataset into basic numerical measurements. Mean, standard deviation, and correlation are often used as metrics. In this paper we have performed univariate analysis, bivariate analysis and box plot for exploratory data analysis.

### Univariate Analysis

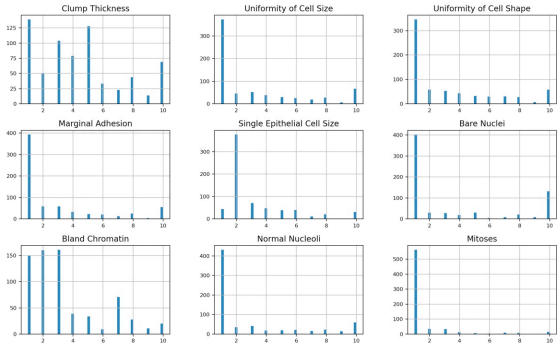


Figure 1: Univariate analysis on features.

Univariate analysis doesn't deal with causes or relationships and its major purpose is to describe the dataset; It takes data, summarizes that data and finds patterns in the data. Here we can see how each features is distributed in the biopsy data and we can get a glimpse of how the dataset is organized.

### Bivariate Analysis

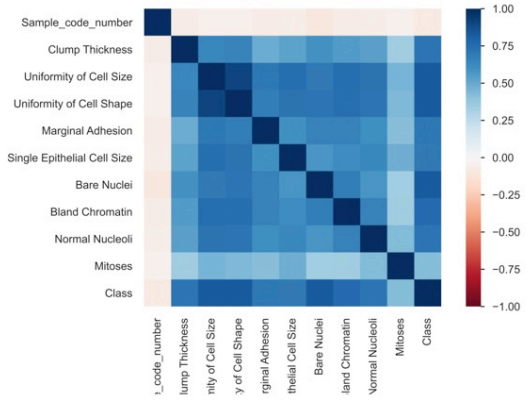


Figure 2: Bivariate analysis on features.

Bivariate analysis observes two variables at the same time. One variable is dependent on the other, while the other is independent. We can examine the changes that occur between the variables and the degree to which they occur. In our dataset we can observe how each feature is correlated to other features. Highly correlated features are later removed as they reduce the accuracy of the algorithm. Thus bivariate analysis helps us identify correlated and insignificant features and how they affect our dataset.

### Box Plots

Box plots provide a glimpse into the basic statistics of the outliers and data. Outliers are values that are extreme and deviate from the normal data in the dataset. Different classes of tumors were labeled, and box plots were built for every feature. The box plot findings were used to choose features that could easily differentiate different tumor types. Box plots are important cause they show the outliers in the data set. Outliers can significantly affect the accuracy of algorithms cause they can mislead the algorithms during the training process.

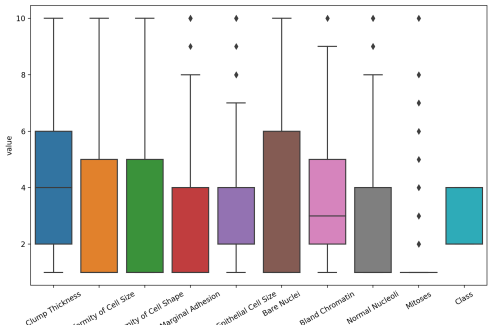


Figure 3: Box plots of features.

## Data Preprocessing

It is critical that we preprocess our data before feeding it into our model. Preprocessing is an important step in Machine Learning because the accuracy of data and the valuable knowledge that can be extracted from it directly influences our model's capacity to learn. Our data has 10 Numeric Features out of which 9 are taken into consideration and 1 Categorical Feature, which is then converted to numeric for model training purposes.

### Data Split

We divided our dataset into two sections before fitting classification models, namely training and test sets. The training set was selected at random from the dataset and comprises 80% of the original set, while the evaluation set contains the remaining 20%.

### Data Standardization

The StandardScaler is used in data standardization to eliminate strongly correlated and insignificant features.

Correlation simply refers to a shared relationship involving two or more things. In mathematics, the above phenomenon is quantified by a suitable equation known as the correlation coefficient. The correlation formula is as follows:

$$\frac{\sum_{i=1}^{n}(x_i - x^-)(y_i - y^-)}{\sqrt{\sum_{i=1}^{n}(x_i - x^-)^2 \sum_{i=1}^{n}(y_i - y^-)^2}}$$

When the correlation coefficient is less than 0, x and y are said to be negatively correlated. If it is greater than zero, they are strongly connected. The correlation coefficient ranges from -1 to 1. When correlation coefficient is < 0, we say that x and y are negatively correlated. If it is > 0, they are positively correlated. Correlation coefficient varies between -1 and 1. When there are several strongly correlated functions, the variance of weights is high. The above variance should be minimal for the model to be robust enough. If the variance is high, the model will be very sensitive to data. If the variation is high, the weights vary significantly from the training results. This indicates that the model will not do well with test results. This is why highly correlated and insignificant features are removed.

## Machine Learning Models

**Naive Bayes**

Its operation is based on Bayes' theorem. The basic assumptions of this algorithm are that each variable contributes equally and independently to the output.

Each feature will be independent of the others and will have the same impact on the performance. As a result, the naïve Bayes theorem does not extend to a lot of real-world problems. We were able to get an accuracy of 94.15% for GaussianNB and an accuracy of 94.15% again for BernoulliNB. Naive Bayes for our application gave us one of the highest accuracy of classification. The performance analysis is again made using precision, recall, F1-score and accuracy of prediction.

| Naive Bayes Model | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Gaussian NB | 0.94 | 0.96 | 0.95 | 0.94 |
| Bernoulli NB | 0.97 | 0.93 | 0.95 | 0.94 |

Table 1 : NB

The implementation of the supervised ML-Models used in this paper are stated below.

**Support Vector Machine**

Our main aim is to find the plane that maximizes the margin between the malignant and benign tumors. SVMs seek a hyperplane $w.x + b = 0, x_i \in R^n$ that divides the data points $x_i$ and corresponds to a given decision rule: $g(x) = sign(w.x + b)$. SVMs choose the dividing hyperplane $w.x + b = 0$ with the greatest margin that is farthest away from the data points. We were able to get an accuracy of 93.56% for Support Vector Machine with linear kernel and hyperparameters being C=0.1, gamma = 0.1, an accuracy of 93.56% for RBF kernel with hyperparameters being C=0.1, gamma = 0.1 and an accuracy of 94.73% for when polynomial kernel is used with Support Vector Machine with hyperparameters being C=10, gamma = 0.1. The performance analysis is made using precision, recall, F1-score and accuracy of prediction. Where precision is the measurement of how accurate the data is, recall is the true positive rate of the day and the harmonic average of precision and recall is F1 - Score.

| SVM Model | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Linear | 0.94 | 0.95 | 0.95 | 0.94 |
| RBF | 0.95 | 0.94 | 0.95 | 0.94 |
| Poly | 0.94 | 0.98 | 0.96 | 0.95 |

Table 2 :SVM

**Logistic Regression**

This algorithm gives us the probability of falling into a particular class in this case its malignant or benign. We were able to get an accuracy of 92.98% after pre-processing of data, hyperparameter tuning and applying l2 penalty over it with parameter C = 0.025, gamma = 1 and an accuracy of 94.73% before tuning and pre-processing. When there is a single binary dependent variable and multiple independent variables, logistic regression is appropriate. Significant features like Cell Size and Nuclei are used to predict the outcome: malignant or benign, which is a dependent variable.

| Best C | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| C = 0.025 | 0.95 | 0.96 | 0.96 | 0.95 |
| C = 0.1 | 0.93 | 0.96 | 0.94 | 0.93 |

Table 3 : LR

### Decision Tree

We got an accuracy of 91.81% when Criterion = gini, accuracy of 93.56% when Criterion = gini, max_depth = 5, min_impurity_split = 1e-07, min-samples-split =2and did further tuning. When we changed the criterion = entropy, we got an accuracy of 95.32%. Below table displays some more details.

| Criterion | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Criterion = Entropy | 0.95 | 0.97 | 0.96 | 0.95 |

| Criterion | Maximum Depth | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|---|
| Gini | Plain Gini | 0.91 | 0.96 | 0.94 | 0.92 |
| | max_d = none | 0.90 | 0.96 | 0.93 | 0.91 |
| | max_d = 5 | 0.94 | 0.96 | 0.93 | 0.94 |
| | max_d = 32 | 0.90 | 0.96 | 0.93 | 0.91 |

Table 4 : Decision Tree

### Random Forest

Random forest is just another variation of decision tree. Multiple decision trees are used in Random Forest algorithm. We were able to get an accuracy of 95.32% when the number of decision trees was 5 which is the highest accuracy of all the algorithms whom we used and whose hyperparameters we tuned. We got an accuracy of 94.73% when the number of decision trees was 10, We were able to get an accuracy of 94.15% when the number of decision trees was 50, and got an accuracy of 94.15% when the number of decision trees was 100.

| Number of DT's | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| DT = 5 | 0.95 | 0.97 | 0.96 | 0.95 |
| DT = 10 | 0.95 | 0.97 | 0.96 | 0.95 |
| DT = 50 | 0.94 | 0.97 | 0.95 | 0.94 |
| DT = 100 | 0.94 | 0.97 | 0.95 | 0.94 |

Table 5 : RF

### KNN

The optimum k value is 3 as we got the highest accuracy of 94.15% with it. We were able to get an accuracy of 92.39% when k = 2, 92.98% when k = 5 and the table below displays more such values.

| Value of k | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| k = 2 | 0.91 | 0.97 | 0.94 | 0.92 |
| k = 3 | 0.94 | 0.96 | 0.95 | 0.94 |
| k = 5 | 0.94 | 0.95 | 0.94 | 0.93 |
| k = 7 | 0.94 | 0.95 | 0.95 | 0.94 |
| k = 9 | 0.94 | 0.95 | 0.94 | 0.93 |

Table 6 : KNN

## Conclusion

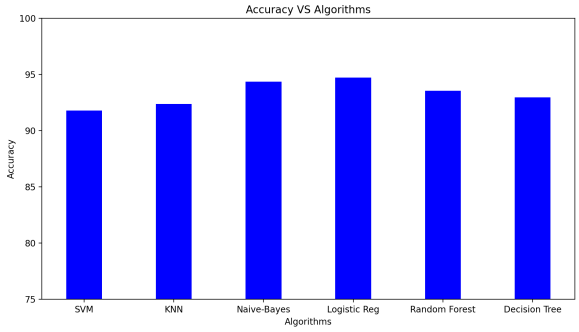


Figure 4: Accuracy of algorithm comparison.

Random Forest gave us accuracy value of 95.32% after fine tuning which is the highest of all algorithms that we implemented with tuned parameters. We understood the importance of removing insignificant and highly-correlated features from our training and also found Cell size and Nuclei to be very significant features directly affecting our output. Before removing highly correlated and insignificant features the accuracies of algorithms were almost similar to each other, but after removing them, there was some difference in the accuracies of the algorithms, which the above figure 4 represents. We tuned all the algorithms and took the best performing model of each of them for comparison. However, with further tuning of all algorithms we get the accuracy of algorithms to be higher and again very close to each other with Random Forest and Decision Tree

taking a lead by a small percentage as displayed by the figure 5 below.

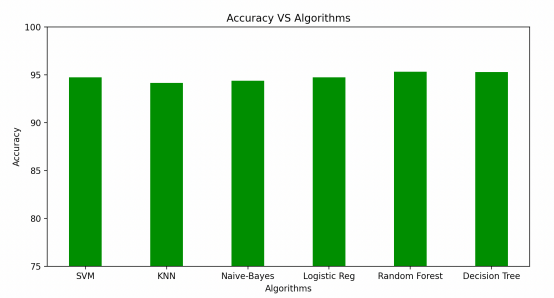


Figure 5 : Accuracy of algorithms before removing highly correlated features.

For future work, these built models of the machine learning classification algorithms can be used to forecast or identify other similar diseases as well, with plain training of data. This can be expanded for other binary classification applications as well and also this knowledge can be useful for implementing other machine learning algorithms for similar cases of automation of breast cancer diagnosis.

## References


Van der Aalst, W. Process Mining: Data Science in Action; Springer: Berlin/Heidelberg, Germany, 2016.

Romero, C.; Ventura, S. Educational data science in massive open online courses. Wires Data Min. Knowl. Discov. 2017, 7, 1–12.

Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. Health Inf. Sci. Syst. 2014, 2, 3.

Sohail, M.N.; Jiadong, R.; Uba, M.M.; Irshad, M. A comprehensive looks at data mining techniques contributing to medical data growth: A survey of researcher reviews. In Proceedings of the 35th IEEE International Conference on Computer Design, ICCD 2017, Boston, MA, USA, 5–8 November 2017; pp. 21– 26.

Petri, I.; Kubicki, S.; Rezgui, Y.; Guerriero, A.; Li, H. Optimizing energy efficiency in operating built environment assets through building information modeling: A case study. Energies 2017, 10, 1167.

Liao, S.H.; Chen, Y.J.; Deng, M.Y. Mining customer knowledge for tourism new product development and customer relationship management. Expert Syst. Appl. 2010, 37, 4212–4223.,

F.; Ferlay, J.; Soerjomataram, I. Global cancer statistics 2 018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 2018, 68, 394–424.

Piñeros, M.; Znaor, A.; Mery, L. A global cancer surveillance framework within noncommunicable disease surveillance: Making the case for population-based cancer registries. Epidemiol. Rev. 2017, 39, 161–169.

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Jothi, N.; Rashid, N.A.; Husain, W. Data mining in healthcare—A review. Procedia Comput. Sci. 2015, 72, 306–313.

Bray, F.; Ferlay, J.; Soerjomataram, I. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 2018, 68, 394–424.

Piñeros, M.; Znaor, A.; Mery, L. A global cancer surveillance framework within noncommunicable disease surveillance: Making the case for population-based cancer registries. Epidemiol. Rev. 2017, 39, 161–169.

Ma, X.; Yu, H. Global burden of cancer. Yale J. Biol. Med. 2006, 79, 85–94.

Sharma, P.; Klemp, J.R.; Kimler, B.F. Germline BRCA mutation evaluation in a prospective triple-negative
breast cancer registry: Implications for hereditary breast and/or ovarian cancer syndrome testing. Breast
Cancer Res. Treat. 2014, 145, 707–714.

Nechuta, S.J.; Caan, B.; Chen, W.Y. The after breast cancer pooling project: Rationale, methodology, and
breast cancer survivor characteristics. Cancer Causes Control 2011, 22, 1319–1331.

Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; Byers, A. Big Data: The Next Frontier for Innovation, Competition, and Productivity; Technical Report; McKinsey Global Institute: Washington, DC,
USA, 2011.

Dhar, V. Data science and prediction. Commun. ACM 2013, 56, 64–73.

Dai, H.; Cheng, Z.; Baii, J. Breast cancer cell line classification and its relevance with breast tumor subtyping.

Blake, C.L.; Merz, C.J. UCI Repository of Machine Learning Databases. 1998. Available online: http://www.ics.uci.edu/mlearn/MLRepository.html (accessed on 20 March 2020).

Alickovic, E.; Subasi, A. Breast cancer diagnosis using GA feature selection and rotation forest. Neural Comput. Appl. 2017, 28, 753–763.

Mangasarian, O.L.; Street, W.N.; Wolberg, W.H. Breast cancer diagnosis and prognosis via linear programming. Oper. Res. 1995, 43, 570–577.

Dubey, A.K.; Gupta, U.; Jain, S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. Int. J. CARS 2016, 11, 2033–2047.

Bazazeh D.; Shubair R. Comparative study of machine learning algorithms for breast cancer detection and diagnosis. 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), 2016, pp. 1-4.

Aalaei, S.; Shahraki, H.; Rowhanimanesh, A.; Eslami, S. Feature selection using genetic algorithm for breast cancer. 16 Computational and Mathematical Methods in Medicine diagnosis: An experiment on three different datasets. Iran. J. Basic Med. Sci. 2016, 19, 476.

Aruna, S.; Rajagopalan, S.; Nandakishore, L. Knowledge based analysis of various statistical tools in detecting breast cancer. Comput. Sci. Inf. Technol. 2011, 2, 37–45.

Chaurasia, V.; Pal, S. Data mining techniques: To predict and resolve breast cancer survivability. Int. J. Comput. Sci. Mob. Comput. 2014, 3, 10–22.

Asri, H.; Mousannif, H.; Al Moatassime, H.; Noel, T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. 2016, 83, 1064–1069.

Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

Qu, Z. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 2011, 9, 515.

Srinivas, K. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In Proceedings of the 5th International Conference on Computer Science &
Education, Hefei, China, 24–27 August 2010; pp. 1344–1349.

Bernal, J.L.; Cummins, S.; Gasparrini, A. Interrupted time series regression for the evaluation of public
health interventions: A tutorial. Int. J. Epidemiol. 2017, 46, 348–355.

Wang, H.; Yoon, W.S. Breast cancer prediction using data mining method. In Proceedings of the 2015
Industrial and Systems Engineering Research Conference, Nashville, TN, USA, 30 May–2 June 2015.

# Appendix

**Dataset statistics**

| | |
|---|---|
| Number of variables | 11 |
| Number of observations | 683 |
| Missing cells | 0 |
| Missing cells (%) | 0.0% |
| Duplicate rows | 8 |
| Duplicate rows (%) | 1.2% |
| Total size in memory | 58.8 KiB |
| Average record size in memory | 88.2 B |

**Variable types**

| | |
|---|---|
| Numeric | 10 |
| Categorical | 1 |

Figure 6 : Dataset Statistics

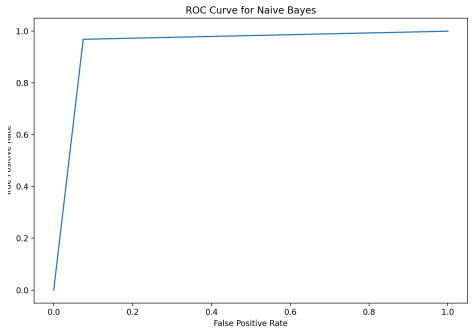


Figure 7 : ROC Curve Naive Bayes

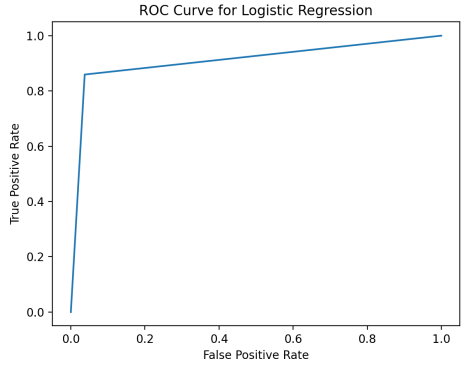


Figure 7 : ROC Curve Logistic Regression