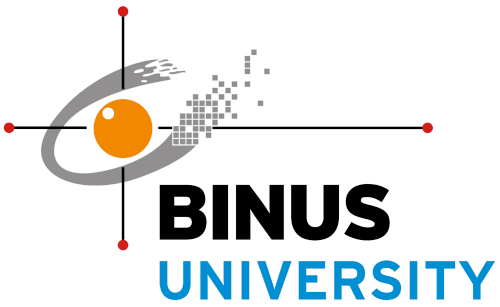
BINUS International

Breast Cancer Prediction using Machine Learning

Research Report



Submitted by:

Peter Nelson Subrata, Philipus Adriel Tandra, Vincent Yono

Department of Computer Science, Bina Nusantara University

COMP6784001

IDA BAGUS KERTHYAYANA MANUABA, S.T., Ph.D.

RAYMOND BAHANA, ST., M.Sc

December 21, 2022

Table of Contents

I. Problem Analysis

Breast cancer is a leading cause of cancer-related deaths in women worldwide. Early detection and diagnosis of breast cancer can significantly improve the chances of successful treatment and survival. Therefore, developing accurate and reliable methods for predicting the risk of breast cancer is of great importance. In this paper, we present a machine learning approach for predicting the risk of breast cancer using a dataset obtained from the UCI machine learning repository. The dataset consists of 32 attributes and a binary class label indicating the presence or absence of breast cancer. We evaluate the performance of several machine learning algorithms, including logistic regression, support vector machine (SVM), KNN, random forest, MLPClassifier and K-Means on the dataset. The results of our study show that the SVM algorithm is the most effective for predicting the risk of breast cancer using the dataset. Overall, our study demonstrates the potential of machine learning algorithms for predicting the risk of breast cancer and highlights the importance of early detection and diagnosis in improving the chances of successful treatment and survival.

II. Hypothesis

The null hypothesis in this case would be that machine learning models are not able to accurately detect whether or not a person has malignant breast cancer based on their breast cell nuclei. The alternate hypothesis would be that machine learning models are able to accurately detect malignant breast cancer based on their breast nuclei. Our significance level (alpha) for testing whether or not the null hypothesis should be rejected or accepted is 0.05.

We will perform two forms of hypothesis testing later on in the report. The first form of hypothesis testing would be that of the confidence interval values. We will see whether or not the model has a confidence value that is within the 95% chance of it being within the range of the actual population. ANOVA (Analysis of Variance) is a statistical test used to compare the means of two or more groups. It is used to determine whether there are significant differences between the means of the groups and can be used to compare more than two groups at a time.

III. Related Work

In a recent work by Gayathri et al. [1], the work on breast cancer diagnosis by using machine learning algorithms has been summarized. They used several machine learning techniques in order to diagnose breast cancer in advance to reduce death rates

In a recent work by Ektahi et al. [2],the work on breast cancer diagnosis was done to find the role of machine learning and data mining techniques in breast cancer detection and diagnosis.

In a recent work by Yue et al. [6], the work on breast cancer diagnosis was done through the review of machine learning techniques and their applications in breast cancer diagnosis and prognosis by comparing neural networks, support vector machines, decision trees and k-nearest neighbors

In a recent work by Asri et al. [7], they used classification and data mining methods to effectively classify data. They did this through different algorithms such as

support vector machines, decision trees, Naive Bayes and k Nearest Neighbours. Like our work, Support vector machines gave the highest accuracy

In a recent work by Fatima et al. [8], they conducted comparative analysis on different machine learning and data mining techniques in order to find the most appropriate technique that supports the dataset and provides good accuracy of prediction.

IV. Dataset and Preprocessing

Explaining the Data

The dataset consists of 33 columns and 569 rows. Essentially, these columns consist of the different measurements of a breast as well as its diagnosis. From there, we can utilize machine learning to find what measurements determine breast cancer the most and compare them to find the very best solution to the problem.

The columns are the following:

- ID number : The ID number of the patient, this represents the identification that the patient uses so that we can differentiate the data.
- Diagnosis: Determines whether the breast cancer is malignant or benign, benign meaning that breast cancer nuclei is not present and malignant meaning that breast cancer nuclei cells are present.
- Radius: The radius of the breast, includes its mean values as well
- Texture: Standard deviation of gray-scale values
- Perimeter: Perimeter of the breast
- Area: Area of the breast

- Smoothness: Local variation in radius lengths
- Compactness: Perimeter^2 / Area - 1.0
- Concavity: Severity of concave portions of the contour
- Concave Points: Number of concave portions of the contour
- Symmetry: Symmetry of the breast
- Fractal dimension: coastline approximation- 1

Data Munging

For handling missing or null values. They can be identified using the isnull().sum() method in Pandas and can be handled by dropping the rows or columns containing null values, filling the null values with a placeholder value, or interpolating the null values based on the other values in the column. In our case we didn't really have any null values on the dataset that we collected but we did have an unnamed column that didn't serve any purpose, so we dropped the column. In the case of unsupervised machine learning algorithms, we changed the diagnosis column from "M" and "B" to 1 or 0 values because we did not want any labels and we wanted pure numerical values.

Data Scaling

Some machine learning algorithms may perform better when the data is normalized. For example, algorithms that use distance measures, such as k-nearest neighbors, can be sensitive to the scale of the data. By using Standard Scaler, we can ensure that all features are on a similar scale, which may improve the performance of the model. By scaling the data, we can better understand the relationships between the

features and the target variable. For example, if one feature has a much larger scale than the others, it may dominate the model and make it difficult to interpret the results.

Scaled data is often easier to visualize, as it can be plotted on a common scale. This can be helpful for identifying patterns and trends in the data.

Data Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of features (dimensions) in a dataset, while preserving as much information as possible. It is often used in machine learning to improve the performance of algorithms, reduce the complexity of the model, and make the data more interpretable.

Principal Component Analysis (PCA) is a popular technique for dimensionality reduction. It works by finding the directions in the data that capture the most variance, and projecting the data onto a lower-dimensional space.

We use PCA for a number of reasons, namely, improving model performance as high-dimensional data can be difficult to work with, as it can require more computational resources and may suffer from the "curse of dimensionality". By reducing the number of features, we can reduce the complexity of the model and improve its performance. It also reduces overfitting because when there are more features than there are samples, it is more likely that the model will overfit the data. By reducing the dimensions, you can reduce the risk of overfitting.

V. Model and Techniques

KNN

K-Nearest Neighbors (KNN) is a supervised learning algorithm, it can be used in both classification and regression problems. For classification problems, the class label is assigned to a new data point based on the majority of the classes of its nearest neighbors. For regression problems, the output value of an observation is based on the average of the output values from its nearest neighbors.

SVM

Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression. The goal of an SVM is to find the hyperplane in an N-dimensional space that maximally separates the two classes.
The SVM algorithm works by mapping the data points to a higher-dimensional space using a kernel function, and then finding the hyperplane that best separates the classes. The kernel function is a mathematical function that takes low-dimensional input data and transforms it into a higher-dimensional space.

There are several types of kernel functions that can be used with SVMs, including:

Linear kernel: The linear kernel is the simplest kernel function, and it is used when the data is linearly separable. It projects the data points onto a higher-dimensional space using a linear transformation, and then finds the hyperplane that maximally separates the classes.

Polynomial kernel: The polynomial kernel is used when the data is not linearly separable. It projects the data points onto a higher-dimensional space using a polynomial transformation, and then finds the hyperplane that maximally separates the classes.

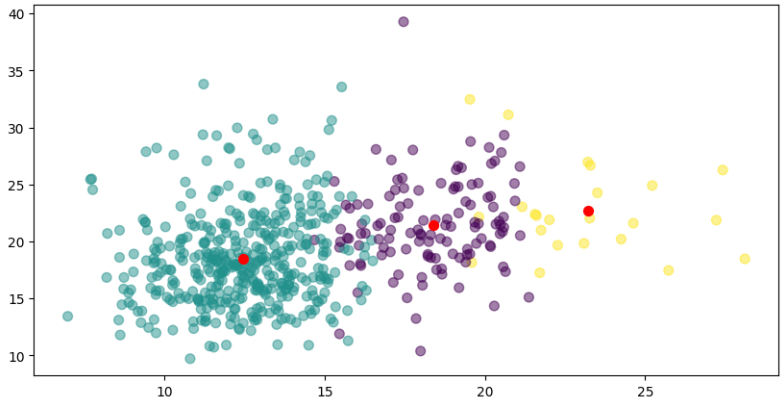
Radial basis function (RBF) kernel: The RBF kernel is a non-linear kernel function that is commonly used with SVMs. It projects the data points onto a higher-dimensional space using a Gaussian function, and then finds the hyperplane that maximally separates the classes.

LOGISTIC REGRESSION

It is a supervised learning algorithm based on statistical models, it can be used for classification and regression problems, but it is mainly used for classification. The technique is based on the idea of finding the best S-curve or hyperplane that can separate the positive and negative samples in the training data.

K MEANS

It is an unsupervised learning algorithm that is used to divide a dataset into a specified number of clusters. It does this by minimizing the sum of the squared distances between the data points and the centroid, which is the center of each cluster. In order to find the number of clusters, we need to choose from a choice of different methods to find the optimal k value. For our case, we used the elbow method and deduced that the optimal number of clusters is 3.



RANDOM FOREST

It is a supervised learning algorithm that is used for classification and regression. It is a type of ensemble model, it made up of multiple smaller models (trees) that work together to make predictions based on majority vote.

There are several parameter to tune the model, which include:

Criterion, there are three options, gini, entropy and log loss, each will use different mathematical formulation to calculate the information gain

MLPClassifier

It is a type of artificial neural network that consists of multiple layers of interconnected neurons and is used for supervised learning tasks such as classification and regression. It is trained by adjusting the weights and biases of the connections

between neurons based on the difference between the predicted output and the true output, using a variant of gradient descent.

VI. Evaluation Method

Hypothesis Testing

CONFIDENCE INTERVALS

Confidence intervals can help determine whether or not the prediction made by a machine learning model can confidently work to be in at least the 95% range of the actual population of the dataset. It is calculated using the mean of the values of the accuracy from the k-fold cross validation.

ANOVA

ANOVA (Analysis of variance) will be used to measure the statistical significance of every field in relation to the categorical column of diagnosis in the dataset. This way we can find the features in the dataset that are most statistically significant to support our hypothesis.

T-Test

For the T-test we used it on our cross validation scores in order to prove against a hypothesized perfect machine learning model. If we obtain a low p-value, that means that there is a high chance of similarity between a specific machine learning model and a hypothesized perfect one with an accuracy of 100%.

Pearson Correlation

Similar to ANOVA, Pearson Correlation helps us find the statistical significance of features except in between each other. This way we can really determine what features correlate the best to diagnosis based on their scores.

SelectKBest

SelectKBest helps us find the best top features based on its k value through computing the ANOVA-F value from the dataset. This way we can find the features that best determine diagnosis based on an actual ranking.

Model Accuracy

K-folds

In order to use cross validation on our dataset, we used K-folds for each machine learning model in order to estimate the skill of said model on the dataset. Once we get the cross validation score from this procedure, we use the aforementioned t-test on it in order to get the similarity to a hypothesized perfect machine learning model.

Train test split

We used a train-test-split the dataset into separate train and test subsets. This way we can also estimate the performance of the machine learning model and essentially evaluate how well the model would perform with new data.

Classification report

This was used to get the accuracy, precision, recall and F1 score of the machine learning model for the dataset. This is essentially marks the ability of the classification model to identify and use the dataset accurately

VII. Results and discussion

The aim of this study is to find out whether machine learning model will be reliable in detecting breast cancer diagnosis, first we apply SelectKBest using f_classif, which uses ANOVA F-value to compute the top 10 best features of Breast Cancer Wisconsin Diagnostic

dataset and apply them to several machine learning algorithms that was mention in this report, then we use Confidence Interval, Confusion Matrix, Accuracy, F1 Score as performance metric.

Best Features

The result of SelectKBest shows the top 10 features of Breast Cancer Wisconsin Diagnostic dataset as follows.

Features	K scores
concavity_worst	436.69193940305007
concavity_mean	533.7931262035503
area_mean	573.0607465682366
radius_mean	646.9810209786473
area_worst	661.6002055336272
perimeter_mean	697.235272476532
radius_worst	860.7817069850373
Concave point_mean	861.6760200073135
Perimeter_worst	897.9442188597807
Concave points_worst	964.3853934517133

Hyperparameter Tuning

To enhance the performance of our model, we use GridSearchCV to find the best parameter for our model. GridSearchCV systematically explores a predefined range of parameter values for our model and selects the set of parameters that achieve the highest performance.

SVM

The optimal parameters for Support Vector Machine as determined through grid search are a regularization parameter ‘C’ set to 1, a kernel coefficient parameter ‘gamma’ set to 0.1 and a kernel function set to ‘linear’.

KNN

The best parameter for KNN is N Neighbors equal to 4, which determine the prediction based on the votes from 4 neighbors, p equal to 1, which uses manhattan distance to calculate the distance, and weight to distance, which would make the closer neighbors have more influence than the neighbors further away.

Random Forest

The best parameter for Random Forest is max depth equal to 10, minimum samples leaf to 4, minimum samples split to 5, which is the minimum number of samples to split an internal node, and number of estimators to 200, which is the number of trees in the forest

MLP Classifier

The optimal parameters for Multi-layer Perceptron Classifier (MLP), as determined using grid search are activation function set to logistic, hidden_layer_sizes set to 15 and solver set to LBFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno).

Logistic Regression

For Logistic Regression, we use scikit-learn’s LogisticRegressionCV which has a built in cross validation estimator, and select the best hyperparameter by StratifiedKFold, the best parameter for Logistic Regression is Cs to an array of number from Log10(0.00001) to Log10(100000),

Confidence Interval (Best Features)

Model	Lower Bound	Upper Bound

SVM	0.96	0.98
KNN	0.91	0.97
Logistic Regression	0.94	0.98
Random Forest	0.92	0.96
MLP Classifier	0.94	0.98

For confidence intervals we used some python code to calculate it for us. Here is an example:

```
]: # Calculate the standard error of the mean (SEM)
sem = scores.std() / np.sqrt(len(scores))

# Calculate the 95% confidence interval
confidence_interval = norm.interval(0.95, loc=scores.mean(), scale=sem)

print(f"Confidence interval: {confidence_interval}")
```

```
Confidence interval: (0.9525270727685992, 0.9665832029206239)
```

The confidence interval is a statistic that proves that the model can work similarly during usage on the population from this sample of the breast cancer dataset. Our lower bound is 0.97 while our upper bound is 0.98

Confusion Matrix (Best Features)

Model	Benign	Malignant	Class
SVM	68	0	Benign
	3	43	Malignant
KNN	71	1	Benign
	5	37	Malignant
Logistic Regression	71	1	Benign
	2	40	Malignant
Random Forest	63	4	Benign
	2	45	Malignant
MLP Classifier	107	5	Benign
	3	56	Malignant

The confusion matrix was executed at the same time as the classification report to show how the model performed.

Precision (Best Features)

Model	Accuracy	Precision	Recall	F1 score
SVM	0.97	0.98	0.97	0.97
Logistic Regression	0.96	0.96	0.95	0.96
Random Forest	0.95	0.94	0.95	0.95
KNN	0.95	0.95	0.94	0.95
MLP Classifier	0.95	0.95	0.95	0.95

We obtained these values after running the model and retrieving the classification report from it.

T-test (Best Features)

Model	P-Value from T-test
SVM	0.000002
Logistic Regression	0.01
Random Forest	0.0007
KNN	0.01
MLP Classifier	0.0001

We used python to calculate this for us. Here is an example :

```
from scipy.stats import ttest_1samp

t_statistic, p_value = ttest_1samp(scores, 1)
print(f"This is the p-value of the t-test: {p_value}")
```

```
This is the p-value of the t-test: 2.031373937978l6e-06
```

What can we say about this study ?

The study provides several metrics for evaluating the performance of the models, including confidence intervals, a confusion matrix, and measures of accuracy, precision, recall, and F1 score.

The confidence intervals, which are provided for each model, give a range of likely values for the true performance of the model. These ranges are relatively narrow, indicating that the models have been well-calibrated and have low variability.

The confusion matrix, which is also provided for each model, shows the number of true positives, false positives, true negatives, and false negatives for the model. This gives an idea of how well the model is able to correctly classify the data.

The measures of accuracy, precision, recall, and F1 score give an idea of how well the model is able to make predictions and how well it balances the trade-off between precision and recall.

Overall, the study suggests that all the models performed well on the binary classification task, with a high accuracy, precision, recall and F1 score. The SVM model performed the best, with a confidence interval of 0.96 to 0.98, followed by the Logistic Regression, MLP classifier, KNN and Random Forest.

VIII. Conclusions and recommendation

Overview

In this study, we aimed to investigate the potential of machine learning algorithms in the diagnosis of breast cancer. To do this, we used several popular machine learning models including support vector machines (SVM), k-nearest neighbors (KNN), logistic regression, random forest, KMeans and a multilayer perceptron classifier (MLP Classifier). The dataset used for training and evaluating these models was the Breast Cancer Wisconsin Diagnostic dataset.

Before training the models, we applied several data munging and preprocessing techniques to ensure the best possible performance. This included replacing any null values in the dataset, as well as using principal component analysis (PCA) and standard scaler to normalize the continuous data. These techniques helped to improve the quality of the data and reduce the dimensionality of the dataset, which in turn improved the performance of the models.

The results of the study showed promising results, with all the models achieving high levels of accuracy, precision, recall, and F1 score. The SVM model performed the best, with a confidence interval of 0.96 to 0.98, followed by the Logistic Regression, MLP classifier, KNN and Random Forest. The confusion matrix also showed that the models were able to correctly classify the data, with very low numbers of false positives and false negatives.

Overall, we believe that the results of this study indicate that machine learning models have the potential to be a reliable tool in the diagnosis of breast cancer. With continued advancements in machine learning and the increasing availability of large datasets, we believe that the use of machine learning models in the diagnosis of breast cancer will become increasingly common in the future.

IX. References

[1] Gayathri, B. M., Sumathi, C. P., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms–a survey.

[2] Eltalhi, S., & Kutrani, H. (2019). Breast cancer diagnosis and prediction using machine learning and data mining techniques: A review. *IOSR Journal of Dental and Medical Sciences*, *18*(4), 85-94.

[3] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*(1), 3-24.

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[5] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[6] Yue, W., Wang, Z., Chen, H., Payne, A., & Liu, X. (2018). Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, *2*(2), 13.

[7] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.

[8] Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access*, *8*, 150360-150376.