IMPROVING BREAST CANCER DETECTION WITH NAIVE BAYES: A PREDICTIVE ANALYTICS APPROACH

Muhammad Garba[1], Muhammad Abdurrahman Usman[1] and
Anas Muhammad Gulumbe[2]

[1]Department of Computer Science, Faculty of Physical Sciences, Kebbi State University of Science & Technology, Aliero. Nigeria.
[2]Computer Science, Faculty of Physical Sciences, Kebbi State University of Science & Technology, Aliero. Nigeria.

ABSTRACT

The study focuses on predicting breast cancer survival using naïve bayes techniques and compares several machine learning models across large dataset of 310,000 patient records. The survival and non-survival classes were the two main categories. The objective of the study was to assess the effectiveness of the Naïve Bayes classifier in the data mining area and to attain noteworthy outcomes for survival classification that were consistent with the body of existing literature.

Naive Bayes achieved an average accuracy of 91.08%, indicating reliable performance but with some variability across folds. Logistic Regression achieved an accuracy of 94.84%, excelling in identifying instances of class 1 but struggling with class 0. Decision Tree model, with an accuracy of 93.42%, showed similar performance trends.

At 95.68% accuracy, Random Forest outperformed Decision Tree. However, all models faced challenges in classifying instances of class 0 accurately. The Naive Bayes algorithm was compared with K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). Future research will enhance prediction models with new methods and address the challenge of accurately identifying instances of class 0.

KEYWORDS

Machine Learning, Data Mininig, Naïve Bayes, Cancer, random survival forest,

1. INTRODUCTION

Breast cancer (BC) is the most common cancer among women and is the first leading cause of cancer-related deaths among women and the second leading cause of cancer deaths worldwide.[1]. According to the 2012 World Health Organization (WHO) classification, breast cancer is primarily categorized into carcinomas and sarcomas. In 2023, it is estimated that 5,400 Canadian women will die from breast cancer, representing 13% of all cancer deaths in women[2]. Breast cancer is caused by the uncontrolled growth of cells in breast tissues, which can be either benign or malignant. It is known as the most common invasive type of cancer among women[3]. The way that stromal cells and tumor cells interact in the tumor microenvironment determines how quickly breast cancer progresses[4]. While most breast cancer patients experience a lower rate of disease

David C. Wyld et al. (Eds): SPTM, AIS, SOENG, AMLA, NLPA, IPPR, CSIT – 2024
pp. 185-196, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141116

recurrence after receiving chemotherapy, therapies like targeted, endocrine and others develop acquired resistance.
The difference in breast cancer mortality between Black and White women has not decreased; Black women still have a 40% higher death rate from the disease despite a lower incidence rate. Death rates among Hispanics, Blacks, Whites, and Asians/Pacific Islanders decreased throughout the last five years, but rates among American Indians and Alaska Natives remained steady.

Studies have leveraged innovative biomedical technologies, high-quality data, and advanced analytical methods to make significant advancements in predicting breast cancer survivability, suggesting time- and cost-effective treatment options for breast cancer patients.

Several factors can affect breast cancer survivability, including:

1. Tumor Stage: One important issue to consider is the degree of cancer at the moment of diagnosis. Early-stage cancers (I and II) generally have higher survivability rates compared to later stages (III and IV) when the cancer has spread to lymph nodes or other organs [8]
2. Tumor Subtype: Breast cancer is classified into various types based on the existence or missing molecular indicators like receptors for hormones and HER2. The subtype can influence the aggressiveness of the cancer and the effectiveness of treatment.
3. Response to Treatment: How well the cancer responds to treatment, such as chemotherapy, hormone therapy, or targeted therapy, can affect survivability.
4. Tumorigenic Cell Population: Research has identified tumorigenic (tumor-initiating) and nontumorigenic breast cancer cells. The ability to prospectively identify and target the tumorigenic cell population may lead to more effective therapies

The survivability of individuals with breast cancer is largely determined by these and other factors. It's important to keep in mind that each patient's situation is unique and that many factors influence survivability; these aspects should all be assessed and managed by medical professionals.

Machine learning (ML) is a significant area of study within artificial intelligence that deals with algorithms that use data to continually grow smarter with experience. The field focuses on prediction based on known properties learned from the training data. The three main categories of techniques are reinforcement learning, supervised learning, and unsupervised learning. ML is used in many different sectors, including bioinformatics, finance, astronomy, medicine, and farming. Supervised learning algorithms include classification algorithms. By applying machine learning to existing data and observations, this technique is able to group and organize fresh data and observations.

In this study, the Naïve Bayesian algorithm is being utilized for categorizing breast cancer data in order to determine the patient's odds of survival. There are numerous techniques for building classifiers, including the Bayesian approach, decision tree approach, artificial neural network approach, support vector machine approach, genetic algorithm approach, rough set method, fuzzy set method, and more.

Many researchers are drawn to the Bayesian method because of its distinctive ability to express uncertain knowledge, its capacity for rich probability expression, and its incremental learning characteristics that involve the integration of prior knowledge. Several studies have employed Bayesian methods for breast cancer prediction, including Bayesian logistic regression. Other machine learning algorithms, such as support vector machines, decision trees, naive bayes, K-nearest neighbors, and ensemble classifiers, have also been used for breast cancer prediction.

The studies produced positive results in terms of sensitivity, specificity, accuracy, precision and F-measure. Furthermore, some studies have used Bayesian optimization techniques to optimize the prediction accuracy of machine learning algorithms.

2. LITERATURE SURVEY

Both the systematic review and individual studies highlight the increasing interest in employing machine learning methods to predict the survival rate and significant prognostic factors of breast cancer. These investigations offer valuable insights into the potential of machine learning to enhance the precision and dependability of models predicting breast cancer survival. Ultimately, this advancement could contribute to more informed medical decision-making and improved patient outcomes. When compared to conventional techniques, machine learning models have demonstrated encouraging outcomes in predicting breast cancer survival. According to a thorough review, the 5-year survival rate of patients with breast cancer has been forecasted using machine learning approaches, particularly decision trees.

Furthermore, a study examined how well machine learning algorithms predicted breast cancer survival when compared to conventional Cox regression. Out of all the models, the study discovered that the random survival forest (RSF) model had the best discriminative performance, suggesting the promise of machine learning algorithms in this situation.

Using SEER data and a Random Forest classifier, researchers were able to estimate breast cancer survival time within a two-year window with up to 72% accuracy, demonstrating the potential of machine learning approaches in predicting survival time[15]. In a different study, the effectiveness of many machine learning classifiers for predicting breast cancer outcomes was examined. These classifiers were Support Vector Machine, Logistic Regression, Random Forest, XGBoost, AdaBoost, k-Nearest Neighbors, and Naive Bayes. The study demonstrated the application of machine learning algorithms to making decisions about treatment and outcome prediction for breast cancer.

3. METHODOLOGY

We downloaded the Surveillance, Epidemiology, and End Results (SEER) dataset on breast cancer. The National Cancer Institute (NCI) created SEER, a comprehensive database of population-based data on cancer incidence and survival in the United States of America. Data science methodologies harbor the potential to bring invaluable contributions to diverse scientific realms, casting fresh insights upon ubiquitous inquiries. The diagnosis of patients poses a formidable challenge, with only a few doctors possessing the ability to accurately predict diseases. Data Mining, encompassing various techniques, aims to unearth information and decision-making knowledge from databases.

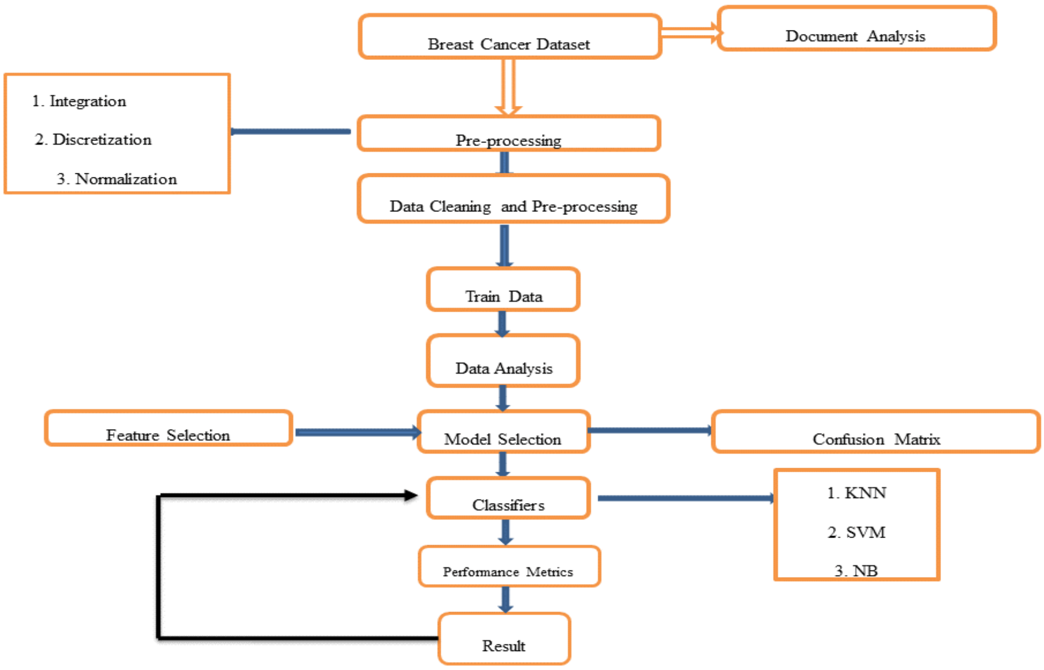

Figure 1: Proposed architecture diagram

This extracted knowledge finds practical application in decision support, predictions, forecasting, and estimation. Data mining is an essential step in the process of knowledge discovery in databases, in which intelligent methods are applied in order to extract patterns. While the incidence of breast cancer rises with increasing wealth across all age groups, women in the poorest countries experience a disproportionately high death rate from breast cancer, especially those under 50.

```
df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Avr_age_range	53771.0	53.981365	6.666705	42.0	47.0	57.0	62.0	62.0
AgeAbove50_1_or_2	53771.0	1.279984	0.448995	1.0	1.0	1.0	2.0	2.0
er_status_recode_yes1_no2	53771.0	1.169515	0.375210	1.0	1.0	1.0	1.0	2.0
malignant1_negative2	53771.0	1.110004	0.312897	1.0	1.0	1.0	1.0	2.0
no_lymph_2_lymph1	53771.0	1.986610	0.114939	1.0	2.0	2.0	2.0	2.0
Death_byCancer0_byOther1.	53771.0	0.941809	0.234107	0.0	1.0	1.0	1.0	1.0
tumor_size	53771.0	50.985215	164.179656	0.0	11.0	18.0	30.0	999.0
regional_nodes_examined	53771.0	7.920533	17.379687	0.0	1.0	3.0	6.0	99.0
nodes_positive	53771.0	11.198955	29.831885	0.0	0.0	0.0	1.0	99.0
survival_months	53771.0	43.863384	11.806778	0.0	39.0	45.0	52.0	59.0
surv_4years_1_or_2	53771.0	1.577374	0.493982	1.0	1.0	2.0	2.0	2.0

Figure 2. Result of Descriptive Statistics

From the Figure 2, The Pandas describe() method generates descriptive statistics that provide an overview of the distributional shape, dispersion, and central tendency of a dataset. When applied to a Data Frame, the statistical summary of the numerical data is given. This summary includes the count, mean, standard deviation, minimum value, 25th percentile, 50th percentile (median), 75th percentile, and maximum value.

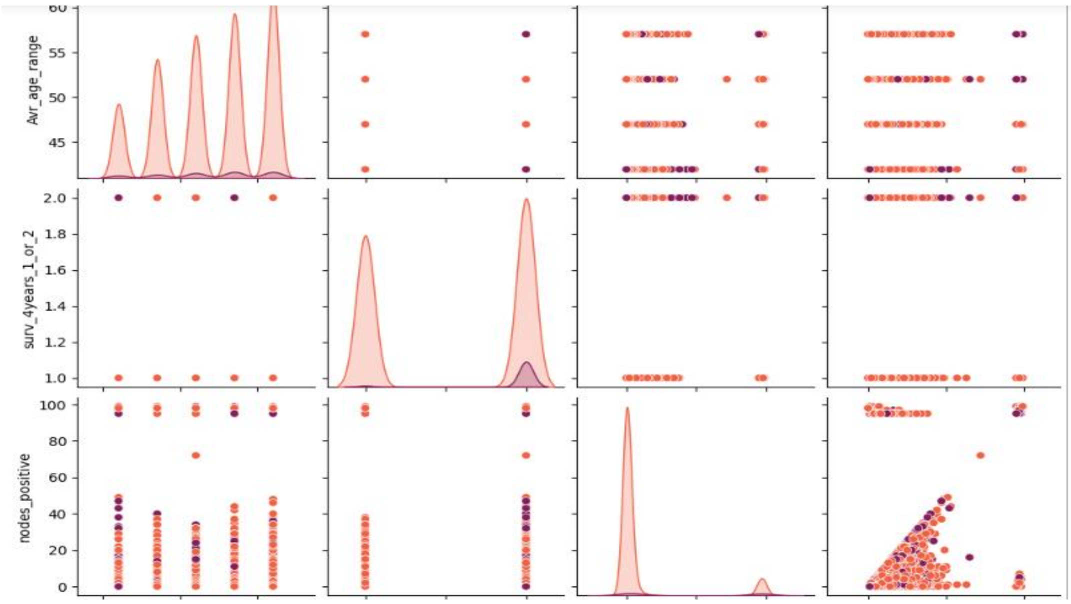

Figure 3. Graphs of Scatter Matrix

From the Figure 3, above, the scatter matrix plotted is used to visualize the relationship between multiple variables and survival time. Displaying the pairwise correlations between all the variables in a dataset, including the survival time, is done through a grid of scatter plots. A density plot, or histogram, is typically displayed on the diagonal of the matrix to display the distribution of each variable. By examining the scatter plot matrix, researchers can identify any patterns or trends in the data and determine which variables are most strongly associated with survival time.

```
df.corr()
```

	Avr_age_range	AgeAbove50_1_or_2	er_status_recode_yes1_no2	malignant1_negative2	no_lymph_2_lymph1	Death_byCancer0_byOtl
Avr_age_range	1.000000	-0.834103	-0.029741	0.101594	-0.004330	0.00
AgeAbove50_1_or_2	-0.834103	1.000000	0.012138	-0.081559	0.003455	0.00
er_status_recode_yes1_no2	-0.029741	0.012138	1.000000	0.015099	-0.046552	-0.16
malignant1_negative2	0.101594	-0.081559	0.015099	1.000000	-0.009720	-0.02
no_lymph_2_lymph1	-0.004330	0.003455	-0.046552	-0.009720	1.000000	0.23
Death_byCancer0_byOther1.	0.000294	0.006736	-0.162941	-0.021022	0.239211	1.00
tumor_size	-0.015582	0.008645	0.055686	-0.003441	-0.102264	-0.13
regional_nodes_examined	-0.043049	0.035538	0.078711	-0.023053	-0.148793	-0.16
nodes_positive	0.000977	-0.001868	0.107863	0.084134	-0.242728	-0.27
survival_months	0.012765	-0.007280	-0.101396	-0.008195	0.148975	0.40
surv_4years_1_or_2	-0.009366	0.008857	0.051598	0.012132	-0.053814	-0.18

Figure 4. Correlation Analysis

The corr() method in Pandas is used to calculate the correlation between columns in a DataFrame. Correlation is a statistical measure that describes the strength and direction of a relationship between two variables. The correlation coefficients between each pair of columns in the original DataFrame are included in the new DataFrame that is produced by the corr() method. Multicollinearity in survival analysis, refers to the presence of near-linear relationships between independent variables in the model. This can lead to estimation instability and difficulties in the interpretation of the model's parameters.

The purpose of selecting and evaluating models, the sklearn.model_selection module has several functions for dividing datasets into training and testing sets, as well as for cross-validation. Specifically helpful for dividing the dataset into training and testing subgroups in the framework of cancer survival analysis is the train_test_split function from sklearn.model_selection. We divide up our data into train and test sets using the train_test_split() function. First, We divide our data into features (X) and labels (y). The dataframe is split up into four sections: y_train, y_test, X_train, and X_test. The model has been fitted and trained using the X_train and y_train sets. To check if the algorithm is correctly predicting the outputs or labels, utilize the X_test and y_test sets. We are able to test the train and test set sizes explicitly.

The arrays created are split into train and test sets. A train set comprises 70% of the dataset, with the remaining 30% going into the test set. Features in the training and testing sets are standardized using the StandardScaler. Making sure that all characteristics are on the same scale through standardization is a crucial preprocessing step in machine learning that can enhance the model's performance. The standardized training set is then used to train a machine learning model, and the standardized testing set is used to test the model's performance.

```
from sklearn.linear_model import LogisticRegression
lr=LogisticRegression()

model1=lr.fit(X_train,y_train)
prediction1=model1.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix

cm=confusion_matrix(y_test,prediction1)
cm
```

```
array([[  233,   679],
       [  153, 15067]], dtype=int64)
```

Figure 5. Logistic Regression Selection

From the Figure 5, The linear model in sklearn.The scikit-learn package contains a class called LogisticRegression that carries out the statistical technique known as logistic regression, which is used to predict binary classes.A logistic regression model is trained using the features in the cancer survival dataset using the LogisticRegression class. The model can then be used to predict the probability of survival for new data. The function sklearn.metrics.confusion_matrix is utilized to assess the effectiveness of a trained model by contrasting the expected and actual class labels. The confusion matrix's result is displayed below.

Array ([[233, 679],
[153, 15067]], dtype = int64)

The confusion matrix counts the number of True and False predictions in order to assess the degree to which the classification system predicts the future. This deduces the following:

a) True positives (TP) = 233 i.e. Meaning 233 case are correctly identified and analyzed.
b) False positives (FP) = 679 i.e. Meaning 679 cases are incorrectly identified.
c) True negatives (TN) = 15,067 i.e. Meaning 15,067 case are correctly rejected.
d) False negatives (FN) = 153 i.e. Meaning 153 case are incorrectly rejected.

```
from sklearn.metrics import accuracy_score

accuracy_score(y_test,prediction1)

0.9484254897098934
```

Figure 6. Testing Accuracy

The above function on Figure 6 accepts the true labels and the predicted labels as parameters and returns the accuracy of the predictions. After passing the testing accuracy value, we arrived same value as the confusion matrix which is 0.9484254897098934.

Random Forest Classifier is a class in the scikit-learn library that implements a random forest algorithm, which is an ensemble method used for classification and regression tasks. The model result outcome is 0.956793949913216 which is higher than the Decision tree (0.9341681130671956). Based on the model out there, the Random Forest classifier is not a good model for this analysis but performs better than Decision tree.

3.1 Comparison between Random Forest, Logistic Regression and Decision Tree Algorithms

- Both models perform well in identifying instances of class 1, but they struggle with class 0.
- Decision trees perform worse than logistic regression in most cases, particularly when it comes to precision and recall for class 0.
- The choice between the two models may depend on the specific goals and requirements of the problem, as well as considerations of interpretability and computational efficiency. Logistic Regression may be preferred when the emphasis is on precision and recall balance.

Table 1. Machine Learning Models Comparison

S/no	Models	Precision, Recall, and F1-Score	Accuracy	Overall
1	Random Forest	performs better for both classes compared to Logistic Regression and Decision Tree	95.68%	most balanced and accurate model among the three
2	Logistic Regression	lower precision, recall, and F1-score for class 0	94.84%	accurate but less balanced, especially for class 0.
3	Decision Trees	lower precision, recall, and F1-score for both classes compared to Random Forest.	93.42%	least accurate and balanced, especially for class 0.

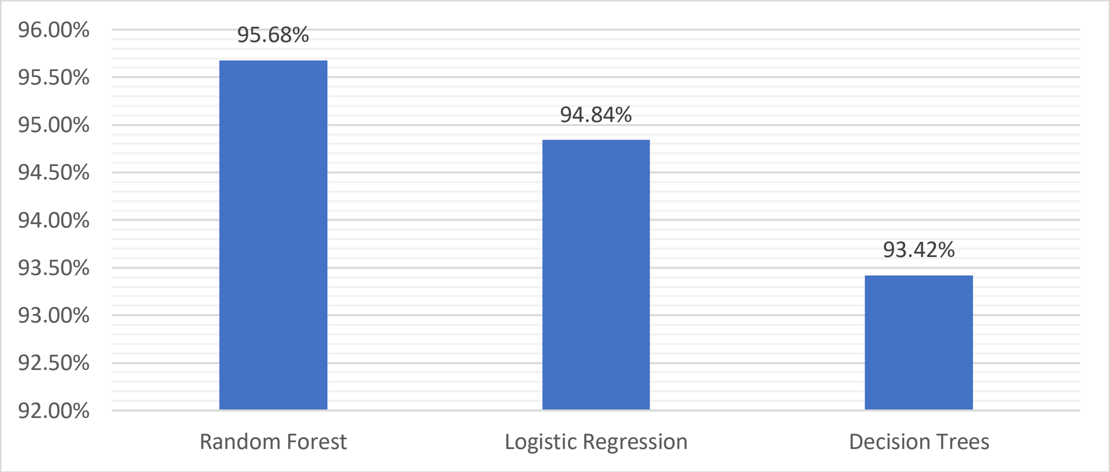

Figure 8. Machine Learning Models Comparison

The K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes algorithms are all popular machine learning algorithms used for classification tasks. The particulars of the dataset will determine which algorithm is best for you.

Based on their attributes, objects can be categorized using the KNN algorithm approach. A majority vote of an unclassified point's k-nearest neighbors, where k is a positive integer, determines the class to which it belongs. To find the closest neighbors, the algorithm uses Euclidean distance metrics.

SVM divides the input space into classes by drawing a hyperplane, then classifies observations according to where they lie on the hyperplane. Because the classification method is defined by a small number of training points (the support vectors), it is memory-efficient and requires fewer computational resources when inferring the class of fresh observations.

Naive Bayes is a simple and fast supervised machine learning algorithm that can be used for classification tasks. It is based on Bayes' theorem and assumes that the features are conditionally independent given the class. Naive Bayes can be used for both binary and multiclass classification problems.

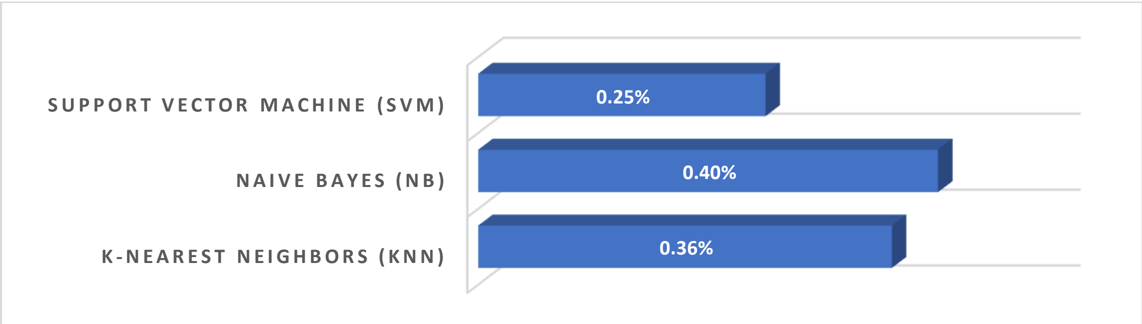

Figure 9. Mean accuracy of KNN, NB, and SVM

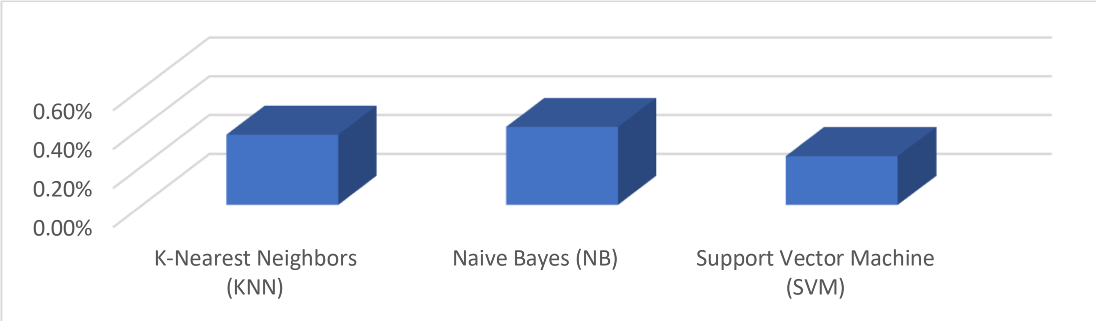

Figure 10. Standard Deviation of KNN, NB, and SVM

From the Figure 9 and 10 above, The K-Nearest Neighbors (KNN), Naive Bayes (NB), and Support Vector Machine (SVM). The results are presented in terms of mean accuracy and standard deviation over the 10 folds. Here's an interpretation of the results:

i. K-Nearest Neighbors (KNN):

   a. Mean Accuracy: 95.29%
   b. Standard Deviation: 0.36%
   c. Interpretation: The KNN model achieved an average accuracy of approximately 95.29%, with a relatively low variability indicated by the standard deviation of 0.36%.

ii. Naive Bayes (NB):

   a. Mean Accuracy: 91.08%
   b. Standard Deviation: 0.40%
   c. Interpretation: The Naive Bayes model demonstrated an average accuracy of around 91.08%, with a standard deviation of 0.40%. This suggests a moderate level of variability in performance across different folds.

iii. Support Vector Machine (SVM):

   a. Mean Accuracy: 95.41%
   b. Standard Deviation: 0.25%
   c. Interpretation: The SVM model performed reasonably consistently across multiple scales, as demonstrated by its low standard deviation of 0.25% and average accuracy of roughly 95.41%.

Based on mean accuracy, the SVM model appears to perform the best among the three algorithms, followed by KNN, and then Naive Bayes.

4. IMPROVING NAÏVE BAYES ALGORITHMS EFFICIENCY AND PERFORMANCE

To improve the performance of Naive Bayes using AdaBoost, we use the A daBoost Classifier in scikit-learn. One way to build an ensemble of weak Naive Bayes classifiers is to use the AdaBoost algorithm. Using various weighted copies of the data used for training, AdaBoost iteratively trains weak classifiers, combining their predictions to produce a strong classifier.

```
from sklearn.ensemble import AdaBoostClassifier
```

```
from sklearn.naive_bayes import GaussianNB
nb_classifier = GaussianNB()
```

```
adaboost_classifier = AdaBoostClassifier(base_estimator=nb_classifier, random_state=10)
```

```
kfold = KFold(n_splits=10, shuffle=True, random_state=10)
cv_results = cross_val_score(adaboost_classifier, X_train, y_train, cv=kfold, scoring='accuracy')
```

```
print(f'AdaBoost ith Naive Bayes: {cv_results.mean():.6f}, ({cv_results.std():.6f})')
```

```
AdaBoost ith Naive Bayes: 0.311692, (0.323618)
```

Figure 11. AdaBoost classifier

Interpretation:

- Mean Accuracy: 0.311692 (31.17%)
    - The mean accuracy of the AdaBoost classifier with Naive Bayes as the base estimator is approximately 31.17%. This shows that around 31.17% of the dataset's occurrences correspond to the class labels that the model, on average, properly predicts.
- Standard Deviation: 0.323618 (32.36%)
    - The relatively high standard deviation of 32.36% indicates a considerable variability in performance across different folds during the cross-validation process. This variability may suggest that the model's performance is inconsistent or that it struggles with certain subsets of the data.

Summary:

- The low mean accuracy suggests that the AdaBoosted Naive Bayes model, as currently configured, does not perform well on the given dataset.
- The high standard deviation indicates inconsistency in the model's performance across different folds, which might be due to the complexity of the dataset or limitations in the base Naive Bayes model.

5. Conclusion

The problems, algorithms, and strategies for the problem of breast cancer survivability prediction in the SEER database were examined and resolved in this study. A number of data mining strategies and tactics were used to address the issue of breast cancer survival. To improve survival analysis for breast cancer, in this research we proposed that support vector machine algorithms, it’s the best suitable for breast cancer survival analysis, it clearly shows the very good promising result.
Future research will focus on incorporation of new methods into the current model of prediction survival along with extending the research in other dimensions. There's room for improvement, especially in accurately identifying instances of class 0.

REFERENCE

[1] Arzanova E, Mayrovitz HN. The Epidemiology of Breast Cancer. In: Breast Cancer. Exon Publications; 2022. p. 1–20.

[2] Dong Q, Huang B. Evaluation of Influence Factors on Crack Initiation of LTPP Resurfaced-Asphalt Pavements Using Parametric Survival Analysis. Journal of Performance of Constructed Facilities. 2014 Apr;28(2):412–21.

[3] Pourmand M. Breast cancer: Causes and prevention. Journal of Cellular Immunotherapy. 2017 Mar;3(1):15.

[4] Mohamed EA, Rashed EA, Gaber T, Karam O. Deep learning model for fully automated breast cancer detection system from thermograms. PLoS One. 2022 Jan 14;17(1):e0262349.

[5] Dong C, Wu J, Chen Y, Nie J, Chen C. Activation of PI3K/AKT/mTOR Pathway Causes Drug Resistance in Breast Cancer. Front Pharmacol. 2021 Mar 15;12.

[6] DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. CA Cancer J Clin. 2019 Nov;69(6):438–51.

[7] Gupta S, Gupta MK. A Comparative Analysis of Deep Learning Approaches for Predicting Breast Cancer Survivability. Archives of Computational Methods in Engineering. 2022 Aug 16;29(5):2959–75.

[8] Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019 Jan 22;321(3):288–300.

[9] Jain P. Detection of Breast Cancer Using Machine Learning Algorithms. Int J Res Appl Sci Eng Technol. 2022 Jun 30;10(6):3484–7.

[10] Junath N, Bharadwaj A, Tyagi S, Sengar K, Hasan MNS, Jayasudha M. Prognostic Diagnosis for Breast Cancer Patients Using Probabilistic Bayesian Classification. Biomed Res Int. 2022 Jul 25;2022:1–10.

[11] Chang M, Dalpatadu RJ, Phanord D, Singh AK. Breast Cancer Prediction Using Bayesian Logistic Regression. Ann Community Med Pract. 2018;4(3):1039.

[12] Ceylan Z. Diagnosis of Breast Cancer Using Improved Machine Learning Algorithms Based on Bayesian Optimization. International Journal of Intelligent Systems and Applications in Engineering. 2020 Sep 28;8(3):121–30.

[13] Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, et al. Predicting breast cancer 5-year survival using machine learning: A systematic review. Vol. 16, PLoS ONE. Public Library of Science; 2021.

[14] Xiao J, Mo M, Wang Z, Zhou C, Shen J, Yuan J, et al. A Comparison Study of Cox Models and Machine Learning Methods for Developing Breast Cancer Prognostic Prediction Models. Available from: https://doi.org/10.2196/preprints.33440

[15] Naser MYM, Chambers D, Bhattacharya S. Prediction Model of Breast Cancer Survival Months: A Machine Learning Approach. In: SoutheastCon 2023. IEEE; 2023. p. 851–5.

[16] Deep V, Sharma H. SVM Classifier on K-means Clustering Algorithm with Normalization in Data Mining for Prediction. International Journal on Recent and Innovation Trends in Computing and Communication. 2019 Jun 22;7(6):29–34.

[17] Bellanger M, Zeinomar N, Tehranifar P, Terry MB. Are Global Breast Cancer Incidence and Mortality Patterns Related to Country-Specific Economic Development and Prevention Strategies? J Glob Oncol. 2018 Dec;(4):1–16.

AUTHORS

Muhammad Garba, is a senior Member in the Department of Computer Science at renowned Kebbi State University of Science and Technology, Aliero (KSUSTA). He has a PhD in Computing and Technology and over 17 years of teaching and research experience at both post-graduate and undergraduate levels. His main research interests are in the areas of software engineering, with special focus on software product line engineering (SPLE) (its methodologies, techniques and tools). Other areas he is interested in are; software measurement and metrics, empirical software, evidence-based research (such as systematic literature review and mapping studies), distributed computing systems, database management technologies. More recently, he also involved in the investigation of data mining in the healthcare system via MSc supervision. He gained some good experience by heading of a department for a good two tenures (i.e., 4 years) and currently serving as the University Director of Information and Communication Technology (ICT). In addition, I have organized and coordinated workshops, seminars, and conferences in a variety of positions. I have also chaired and participated in a number of committees.

Muhammad Abdurrahman Usman a postgraduate student at Kebbi State University of Science and Technology, Aleiro, specializing in Computer Sciences (Masters) with a focus on data science. Born and raised in Katsina State, Nigeria, He demonstrated early proficiency in mathematics and technology, leading to his distinguished undergraduate studies in computer science. At the postgraduate level, He immersed himself in big data analytics, machine learning, and AI. His research aims to enhance decision-making processes across industries through data science. Apart from his academic pursuits, He regularly engages in data science competitions and promotes the use of data to address socio-economic challenges.

Muhammad Anas Gulumbe, is a Lecturer in the department of Computer Science Faculty of Physical Sciences, Kebbi State University of Science and Technology, Aliero. (KSUSTA). Currently pursuing his Phd in Computer Science at the same Institution and has over 10 years' of experience in teaching and research. His current area of research is Edge Computing, while interested in other areas which include: Cloud Computing, AI, E- Learning, and Information Systems. He gained some experience in the department as Level Coordinator for more than 4 years, Departmental Seminar Coordinator for more than 5 years, Departmental Examination Officer for more than 3 years, and also member of some committees at Faculty and Departmental levels.