LEVERAGING NAIVE BAYES FOR ENHANCED SURVIVAL ANALYSIS IN BREAST CANCER

Muhammad Garba, Muhammad Abdurrahman Usman and Anas Muhammad Gulumbe

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

ABSTRACT

The study aims to predict breast cancer survival using Naïve Bayes techniques by comparing different machine learning models on a comprehensive dataset of patient records. The main classification groups were survival and non-survival. The objective was to assess the performance of the Naïve Bayes classifier in the field of data mining and to achieve significant results in survival classification, aligning with current academic research.

The Naive Bayes classifier attained an average accuracy of 91.08%, indicating consistent performance, though with some variability across different folds. Conversely, Logistic Regression achieved a higher accuracy of 94.84%, demonstrating proficiency in recognizing instances of class 1, yet encountering challenges with class 0.The Decision Tree model, with an accuracy of 93.42%, exhibited similar performance patterns. With an accuracy of 95.68%, Random Forest surpassed the Decision Tree. Nonetheless, all models encountered challenges in accurately classifying instances of class 0. The Naive Bayes algorithm was juxtaposed with K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). Future research aims to enhance prediction models with novel methods and tackle the challenge of accurately identifying instances of class 0.

KEYWORDS

Data Mining,Naïve Bayes,K-Nearest Neighbors.Support Vector Machines,Logistic Regression

1. INTRODUCTION

Innovative technological advances, excellent information, and advanced methods for analysis have led to major breakthroughs in forecasting breast cancer survivability and providing cost-effective options for therapy for patients[1]. In the 2012 World Health Organization (WHO) classification, the two main categories for breast cancer are sarcomas and carcinomas. An estimated 5,400 Canadian women are expected to lose their lives to breast cancer in 2023, accounting for 13% of all female cancer deaths[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 recurrence after receiving chemotherapy, therapies like targeted,endocrine andothers develop acquired resistance[5].

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.

DOI:10.5121/ijaia.2024.15402

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 [6]
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 patietns[7]
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 [8].
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 prognosis of breast cancer patients hinges significantly on these and additional variables. It is crucial to note the individuality of each patient's circumstances and the multitude of factors impacting prognosis, all of which should be evaluated and addressed by healthcare professionals[8].
Machine learning (ML) represents a critical domain in artificial intelligence, involving algorithms that iteratively improve their performance with experience gained from data. This field primarily centers on predictive modeling derived from established features learned through training datasets. The principal methodologies encompass reinforcement learning, supervised learning, and unsupervised learning. ML finds application across diverse sectors such as bioinformatics, finance, astronomy, medicine, and agriculture. Within supervised learning, classification algorithms play a pivotal role by effectively categorizing new data and observations based on patterns discerned from existing datasets[9].

In this research, the Naïve Bayesian algorithm is employed to categorize breast cancer data with the aim of assessing patient survival probabilities. Various methodologies for constructing classifiers are explored, encompassing Bayesian methodology, decision tree methodology, artificial neural network methodology, support vector machine methodology, genetic algorithm methodology, rough set approach, fuzzy set approach, and others.

Many scholars are attracted to the Bayesian approach due to its unique capability to articulate uncertain information, its adeptness in expressing complex probabilities, and its incremental learning features that incorporate prior knowledge[10]. Several investigations have utilized Bayesian approaches for the prediction of breast cancer, including Bayesian logistic regression[11]. The prediction of breast cancer has also been used with other machine learning techniques, such as ensemble classifiers, naive Bayes, decision trees, support vector machines, and K-nearest neighbors[12]. Positive results for sensitivity, specificity, accuracy, precision, and F-measure were obtained from the study. Furthermore, some research has used Bayesian optimization techniques to improve machine learning algorithms' prediction performance.[10].

2. LITERATURE SURVEY

There is increasing interest in using machine learning techniques to predict the survival rates and important prognostic markers related to breast cancer, as highlighted by both the overall review and individual investigations. These questions shed important light on how machine learning might improve the precision and dependability of models that forecast breast cancer survival. In the very end, this development may lead to better patient outcomes and more knowledgeable medical decision-making. Machine learning algorithms have demonstrated encouraging results in predicting breast cancer survival when compared to traditional methodologies. Thorough research revealed that the 5-year survival rates of patients with breast cancer have been projected using machine learning techniques, namely decision trees[13].

Moreover, a study investigated the efficacy of machine learning algorithms in predicting breast cancer survival in comparison to conventional Cox regression. Among all models evaluated, the study identified that the random survival forest (RSF) model exhibited superior discriminative performance, indicating the potential of machine learning algorithms in such contexts[14].

Researchers were able to forecast breast cancer survival time within a two-year window with up to 72% accuracy using SEER data and a Random Forest classifier, highlighting the potential of machine learning approaches in predicting survival time[15].Several machine learning classifiers' efficacies in forecasting breast cancer outcomes was investigated in a different study. Logistic Regression, Random Forest, XGBoost, AdaBoost, k-Nearest Neighbors, Support Vector Machine, and Naive Bayes were these classifiers. Using machine learning algorithms to predict treatment outcomes and make therapy decisions for breast cancer was demonstrated in this work.[16].

3. METHODOLOGY

The Surveillance, Epidemiology, and End Results (SEER) dataset pertaining to breast cancer was retrieved. A comprehensive collection of population-based data on cancer incidence and survival in the United States of America was developed by the National Cancer Institute (NCI) and is called SEER. Data science approaches have the potential to significantly advance several scientific fields by providing new perspectives on widely asked questions. It is extremely difficult to diagnose patients since very few physicians are able to correctly foresee illnesses. Data mining is a field of study that uses a variety of approaches to extract information and knowledge relevant to making decisions from databases.

Predictions, forecasting, estimate, and decision assistance are some of the practical uses for this extracted knowledge. When it comes to the process of finding patterns in databases through intelligent approaches, data mining is a crucial step in the process. While the incidence of breast cancer increases with affluence for all age groups, women in the world's poorest nations have a disproportionately high breast cancer death rate, particularly for those under 50[17].

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 DataFrame, 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.

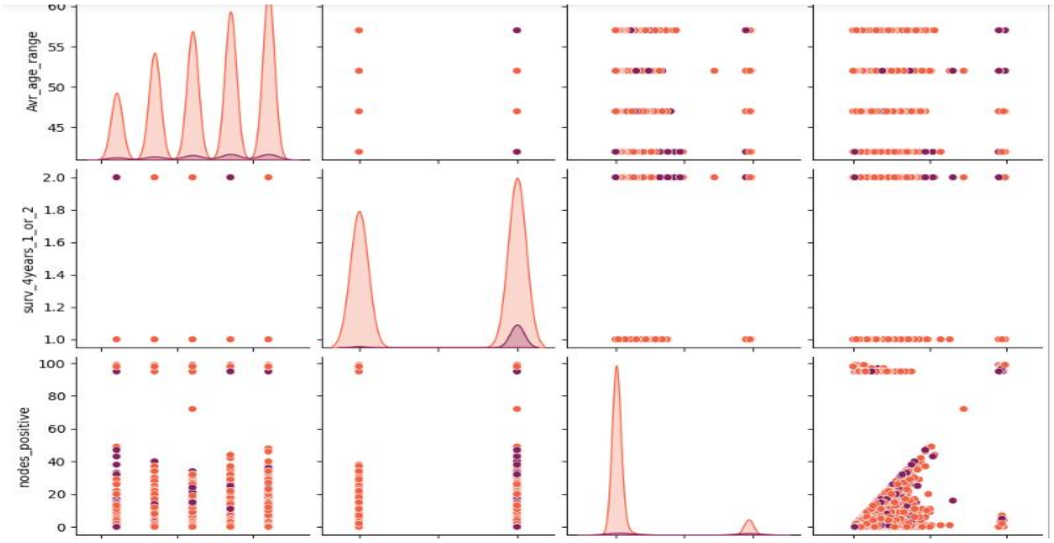
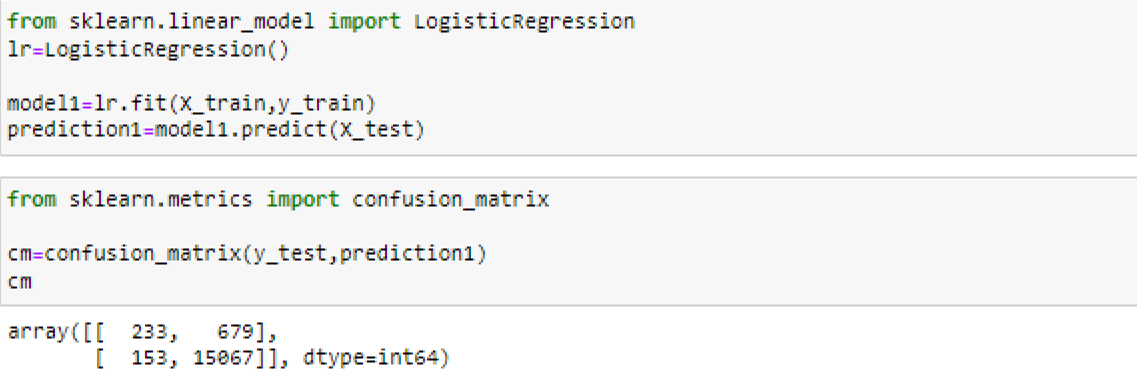


Figure1. Graphs of Scatter Matrix

From the Figure 1, above, the association between several variables and survival time is shown via a scatter matrix diagram. A grid of scatter plots is used to show the pairwise correlations between all the variables in a dataset, including the survival time. To show the distribution of each variable, a density plot, also known as a histogram, is usually shown on the diagonal of the matrix. Researchers can spot patterns or trends in the data and ascertain which variables are most closely linked to survival time by looking at the scatter plot matrix[2]. 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)
```

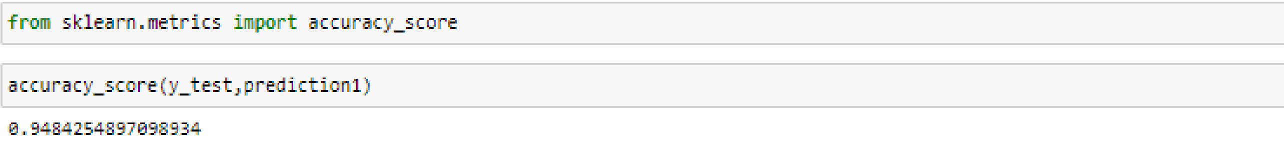
Figure 2. LogisticRegression Selection

From the Figure2, 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 LogisticRegressionclass.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 3. Testing Accuracy

The above function on Figure 3 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.

RandomForestClassifier 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 betweenRandom Forest, Logistic Regression and Decision Three 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.

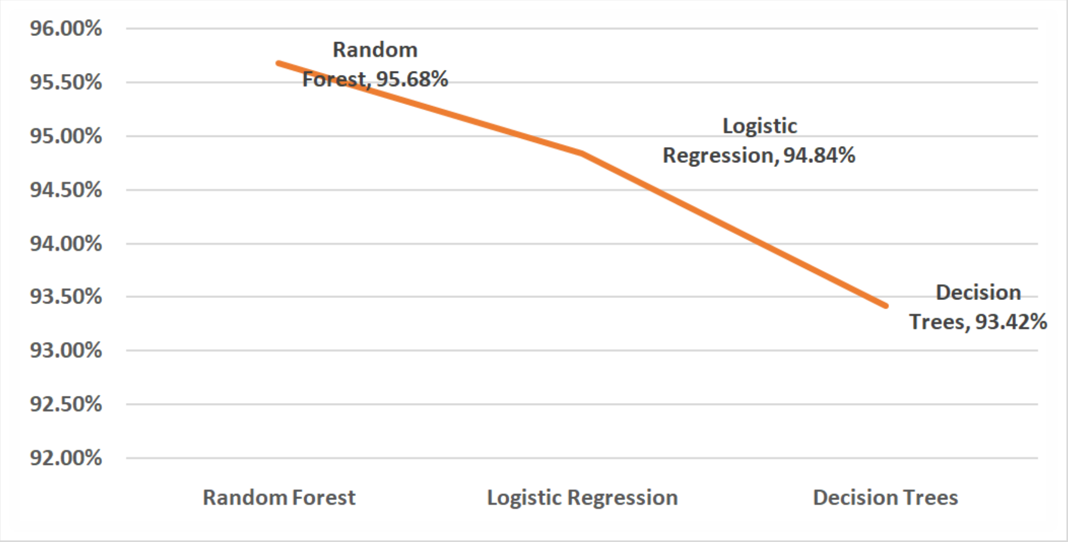


Figure 4. Machine Learning Models Comparison

Popular machine learning techniques used for categorization tasks are K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes. Which approach works best for you will depend on the specifics of the dataset. The KNN algorithm technique allows objects to be categorized according to their properties. An unclassified point is assigned to a class based on a majority vote of its k-nearest neighbors, where k is a positive integer. The algorithm employs Euclidean distance metrics to determine who the closest neighbors are. Using a hyperplane to create a division in the input space, SVM classifies observations based on their location on the hyperplane. The classification approach takes less memory because it is defined by a minimal number of training points, or support vectors.

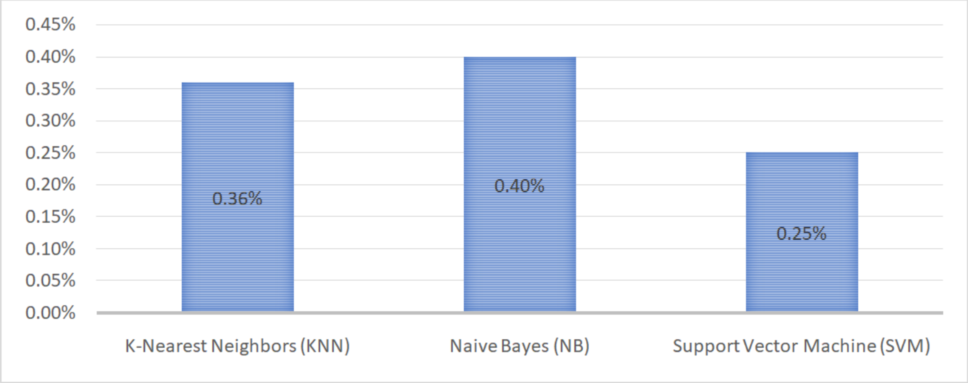


Figure 5. Mean accuracy of KNN, NB, and SVM

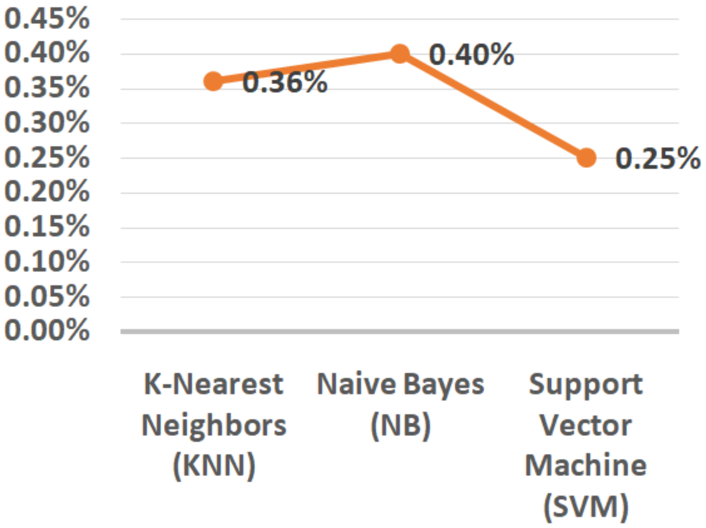


Figure 6. Standard Deviation of KNN, NB, and SVM

From the Figure 5 and 6 above,Support vector machines (SVM), Naive Bayes (NB), and K-Nearest Neighbors (KNN). The findings are displayed throughout the ten folds in terms of mean accuracy and standard deviation. This is how the findings are interpreted: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%.

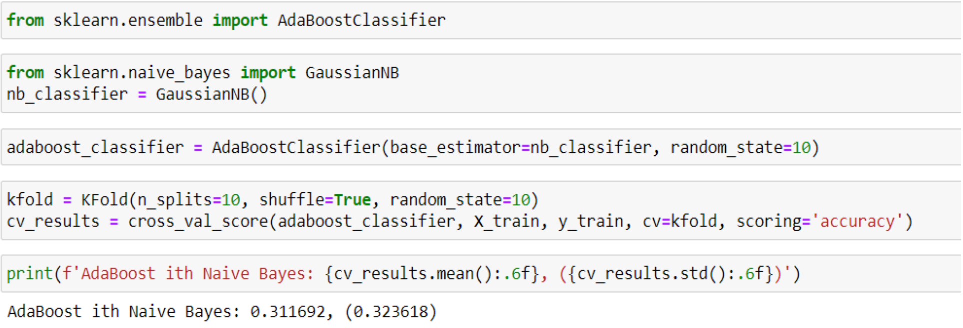
ii. Naive Bayes (NB):
   a. Mean Accuracy: 91.08%
   b. Standard Deviation: 0.40%
   c. Interpretation: The Naive Bayes model demonstrated an average accuracy of around 91.08%, with a standard deviation of 0.40%. This suggests a moderate level of variability in performance across different folds.

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 AdaBoostClassifier 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)
```

Figure7. 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 present work explored and resolved the issues, techniques, and tactics related to the problem of breast cancer survivability prediction in the SEER database. Various data mining techniques and methodologies were used to solve the problem of breast cancer survival. Our study suggests using support vector machine algorithms to enhance breast cancer survival analysis. These algorithms are the most appropriate for this type of analysis and demonstrate excellent and encouraging outcomes.

In addition to expanding the research into other dimensions, future work will concentrate on integrating novel techniques into the current forecast survival model. Particularly when it comes to correctly detecting instances of class 0, there is room for development.

REFERENCE

[1] M.Garba, M.Abdurrahman, A. Gulumbe. " Predictive analytics in breast cancer: A naive bayes perspective," in Conf. Security, Privacy and Trust Management (SPTM 2024), Sydney, AU, 2024, pp.185-196.

[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 scie nce. 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 pu rsuing 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.