Madhuri Basava
08/05/2024
Applied Data Science
DSC680-T302 (2247-1)

Breast Cancer Survival Prediction

Madhuri Basava

Bellevue University

Applied Data Science DSC680

Professor Amirfarrokh Iranitalab

Madhuri Basava
08/05/2024
Applied Data Science
DSC680-T302 (2247-1)

Milestone 3 – White Paper

Topic: The project I chose is Breast Cancer Survival Prediction.

This project focuses on predicting Breast Cancer Survival by creating models that can predict the likelihood of survival based on the given features and help them in every possible manner.

Clustering the people who survived breast cancer based on similar features can help to provide the best care to breast cancer patients.

Business Problem:

Breast Cancer is one of the most common cancers in Women worldwide with approximately 30% of the female cancers. Machine learning has the potential to predict breast cancer based on features hidden in data.

Background/History:

Breast cancer is the most common cancer among women in 154 countries and the main cause of cancer-related death in 103 countries. In 2018, there were approximately 2.1 million new cases of breast cancer in women, accounting for 24.2% of the total cases, and the mortality rate was approximately 15.0%. Over the past 30 years, this disease has increased, while the death rate has decreased may be due to mammography screenings and improvements in cancer treatment.

Madhuri Basava
08/05/2024
Applied Data Science
DSC680-T302 (2247-1)

Research Questions:

1) Which features are most relevant for predicting Breast Cancer survival?
2) Which models are suitable for Breast Cancer survival prediction?
3) What criteria should be used to evaluate the performance of the models?
4) How will missing data be handled in the dataset?
5) What steps will be taken to ensure data quality and integrity?
6) How will data privacy be ensured, given the sensitivity of medical data?
7) How will the ethical implications of survival predictions be communicated to patients and healthcare providers?
8) How will the models be updated with new data and advancements in medical research?
9) How will real-time data be incorporated into the prediction models?
10) How will the success of the project be measured in terms of improving patient outcomes and healthcare efficiency?

Madhuri Basava
08/05/2024
Applied Data Science
DSC680-T302 (2247-1)

Dataset:

The Breast Cancer survival dataset contains 16 columns and 334 rows.

This dataset is taken from the Kaggle website.

Survival Prediction Breast Cancer (kaggle.com)

Data Description: There are a total of 16 fields.

- **Patient_ID**: unique identifier ID of a patient.
- **Age**: age at diagnosis (Years)
- **Gender**: Male/Female.
- **Protein1, Protein2, Protein3, Protein4**: Expression levels of these proteins (units undefined). These data help understand the biological activity within the tumor cells and can be used to identify potential therapeutic targets.
- **Tumour_Stage**: Tumor stage, classified as I, II, or III.
- **Histology**: Histological type of the tumor, which can be Infiltrating Ductal Carcinoma, Infiltrating Lobular Carcinoma, or Mucinous Carcinoma. This classification helps define the type of cells that form the tumor and impacts treatment options.
- **ER status**: Estrogen Receptor status, indicated as Positive or Negative. This shows whether the tumor responds to estrogen, influencing the decision to use hormone therapies.

- **PR status**: Progesterone Receptor status, indicated as Positive or Negative. Similar to ER status, this data helps guide therapeutic decisions based on the hormone sensitivity of the tumor.
- **HER2 status**: HER2 status is indicated as Positive or Negative. A positive result may qualify patients for specific treatments targeting HER2, a growth factor that can promote tumor progression.
- **Surgery_type**: Type of surgical procedure performed, which can be Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, or Other.
- **Date_of_Surgery**: Date on which surgery was performed (in DD-MON-YY format).
- **Date_of_Last_Visit**: Date of the last visit (in DD-MON-YY format).
- **Patient_Status**: Status of the patient, specified as Alive or Dead.

Methods:

I plan to do the prediction analysis with 5 different models as below.

1. Random forest Classifier
2. Decision Tree Classifier
3. XGBoost Classifier
4. Support Vector Classification
5. Logistic Regression

I chose these models to improve accuracy and reduce false positives and false negatives (Inaccurate predictions).

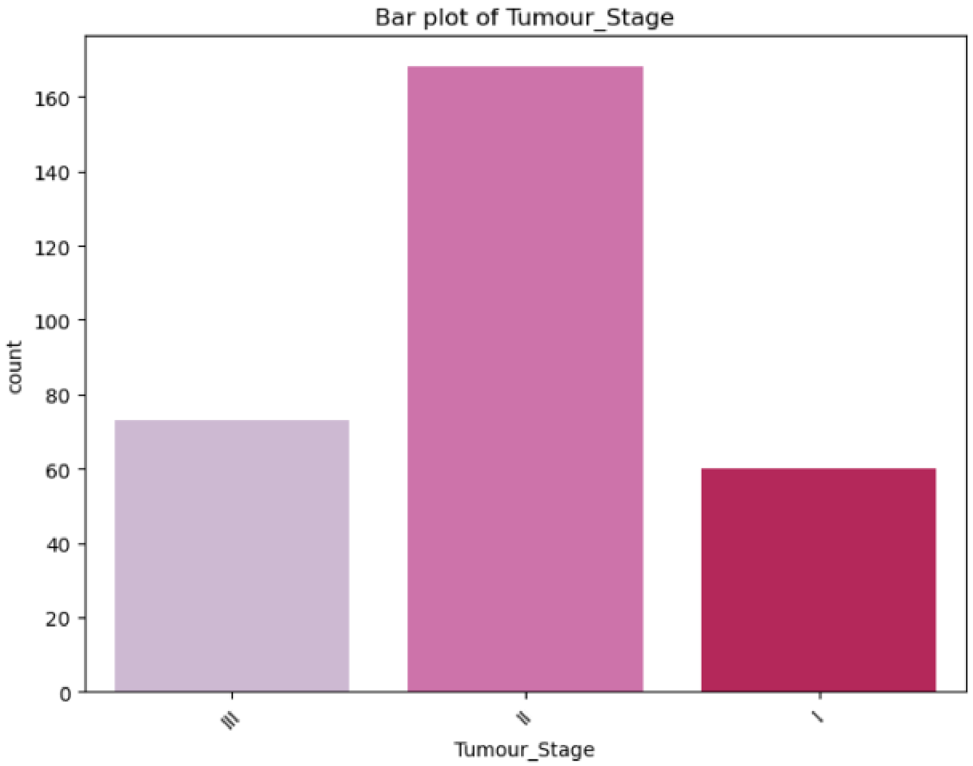
I evaluated the results by calculating the Accuracy of each model.

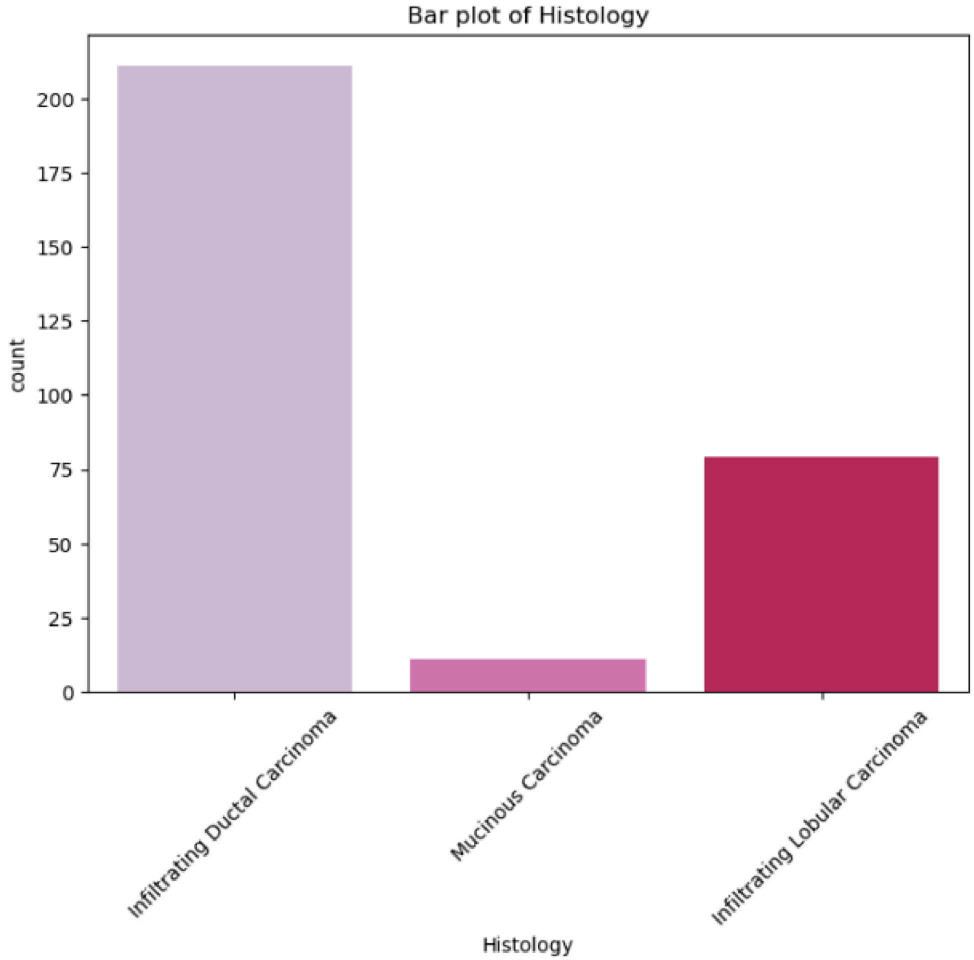
It measures the overall performance of the binary classification model. As both TPR (True Positive Rate) and FPR (False Positive Rate) range between 0 to 1, the area will always lie between 0 and 1, and a greater value of AUC denotes better model performance.

I plan to follow the below steps to evaluate these results.

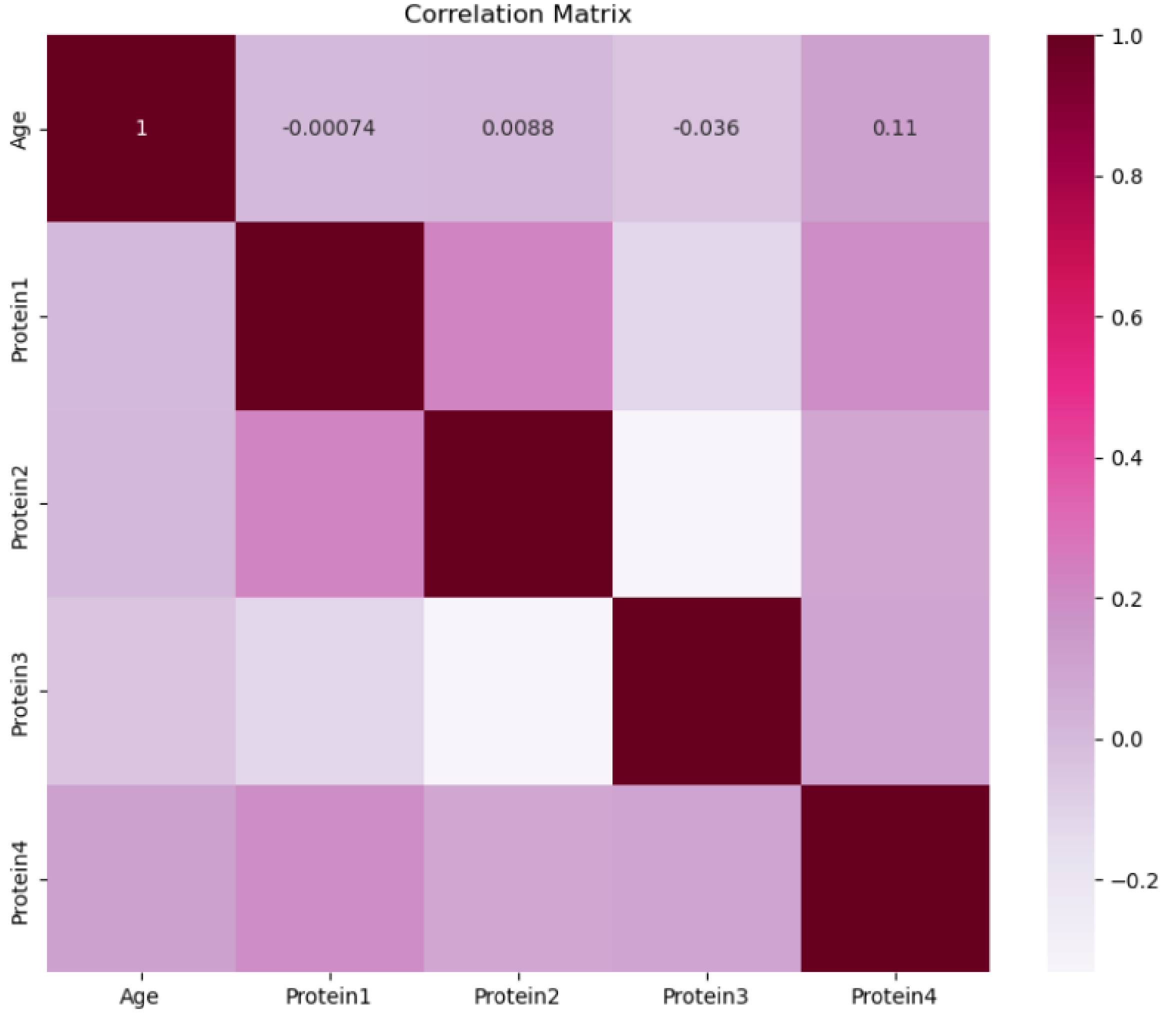
1. Load the data set into a data frame.
2. Perform the Exploratory Data Analysis to understand the characteristics of the data set.
3. Clean the data set. Remove the unnecessary features.
4. Evaluate the correlation between the variables in the dataset.
5. Divide the data set into a train and test data set and apply various models.
6. Create a confusion matrix to show the performance of each model to evaluate the predicted values from the model vs. the actual values from the test dataset.
7. Calculate the Accuracy for each model and choose the best model.

Correlation between the variables in the data set:


Bar plot of Tumour_Stage
count
0
20
40
60
80
100
120
140
160
III
II
I
Tumour_Stage

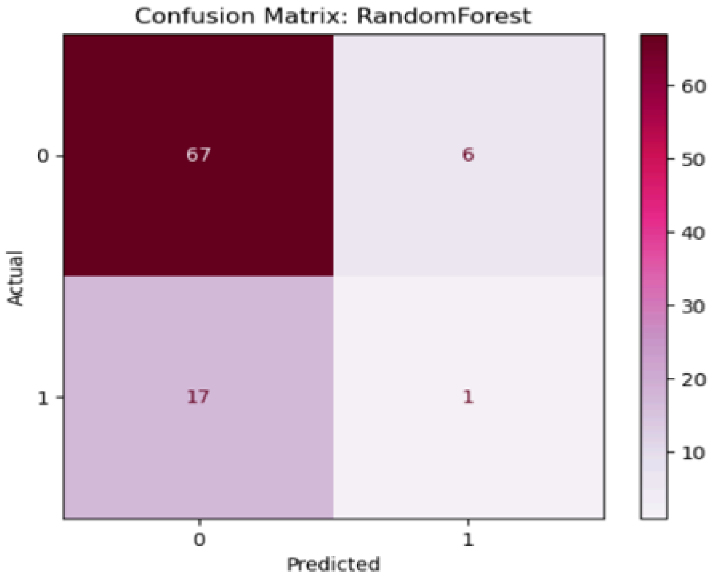


Correlation Matrix:

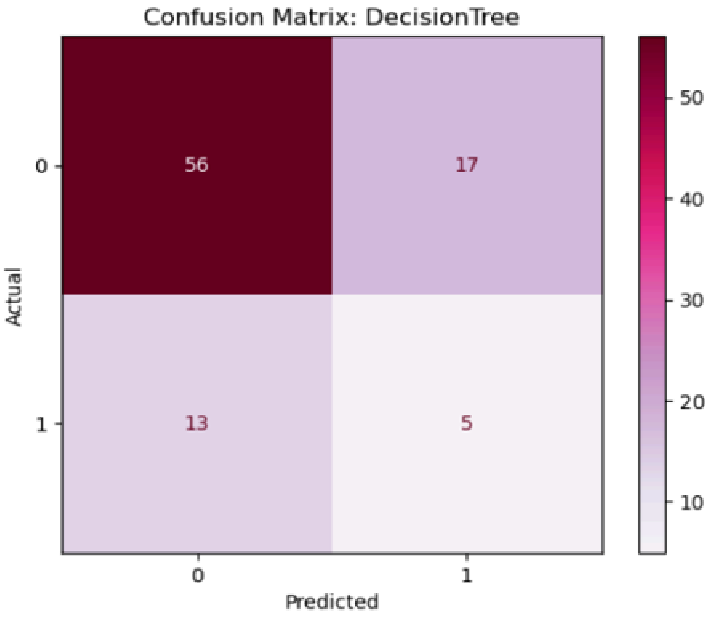


Below is the Confusion Matrix for each of the 5 models.

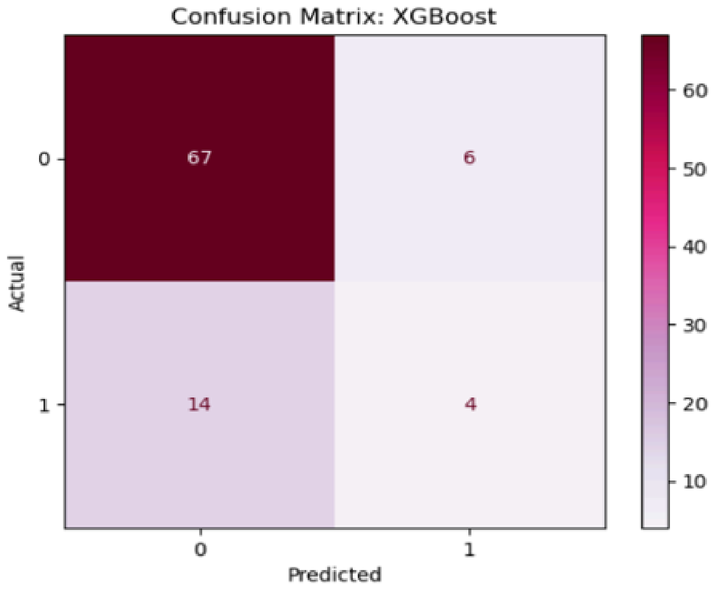
1. Random Forest Classifier



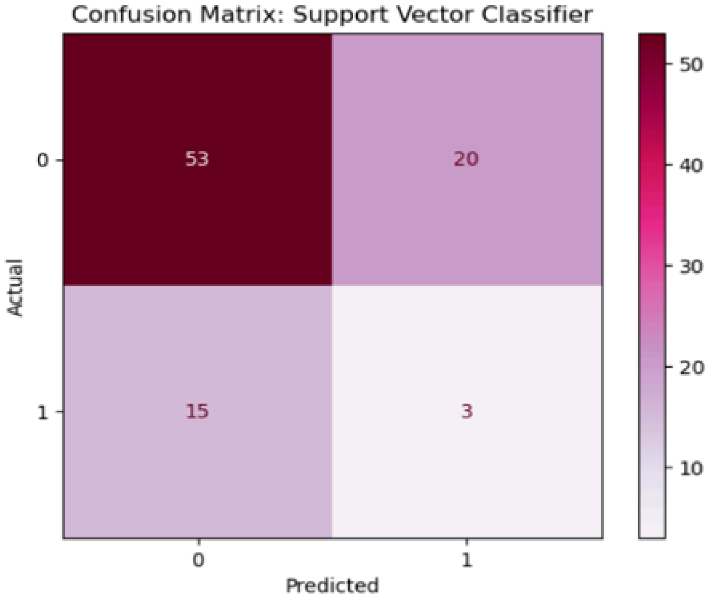
2. Decision Tree Classifier



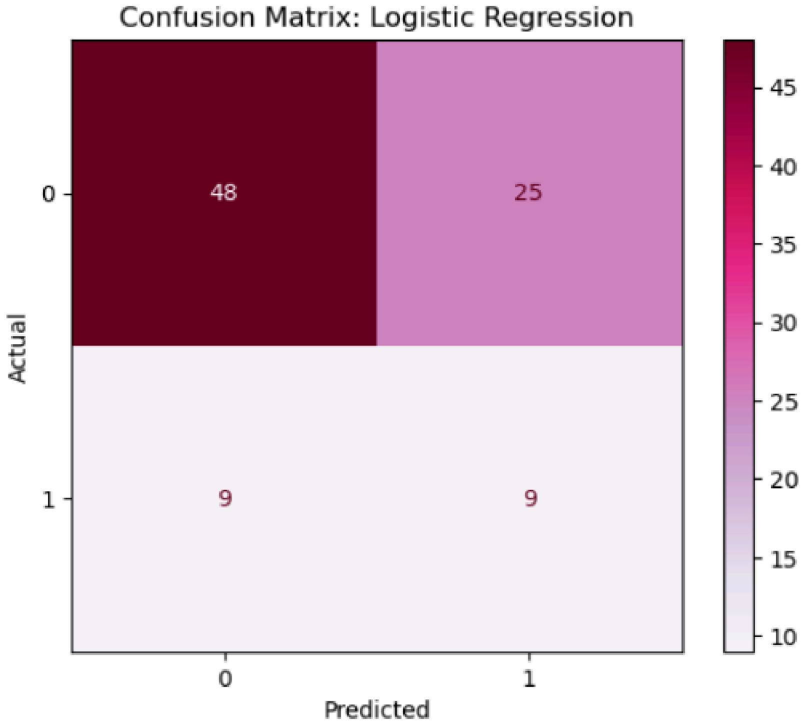
3. XGBoost Classifier



4. Support Vector Machines



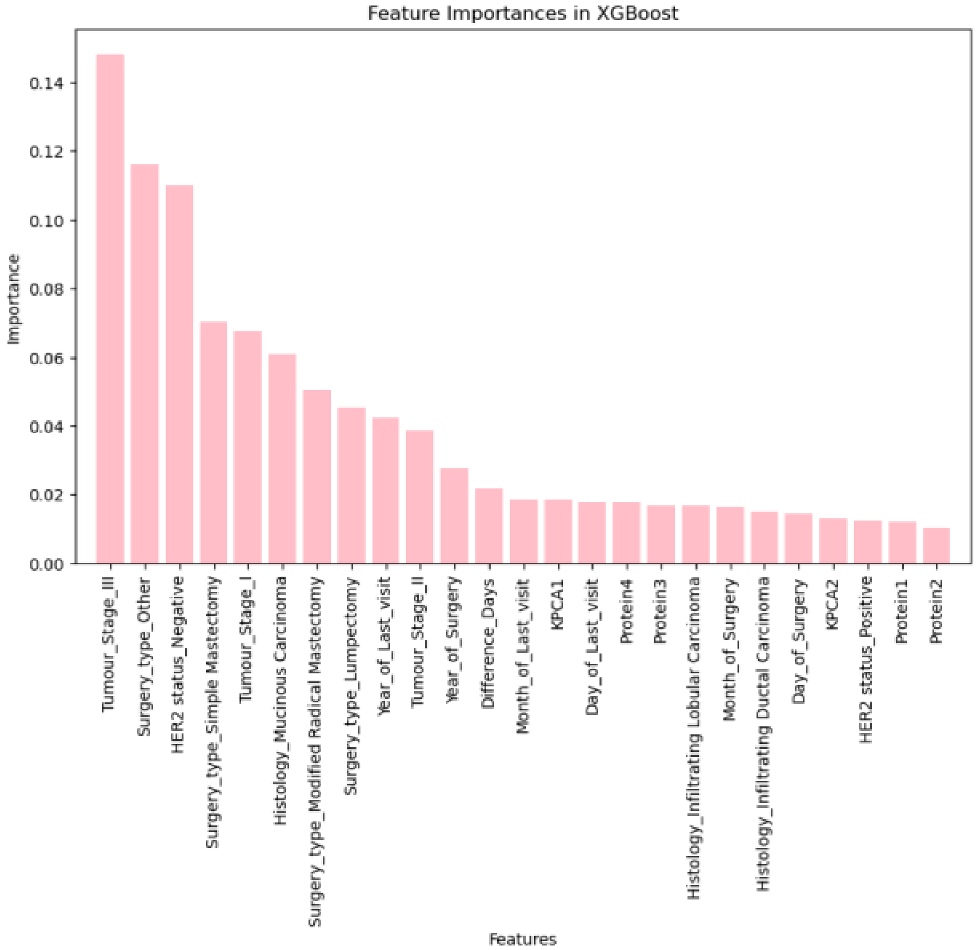
5. Logistic Regression



Answers to the Research Questions:

1) Which features are most relevant for predicting Breast Cancer survival?

   The below chart from the XG Boost Classifier shows that ‘Tumour Stage III’, Surgery type Other, ER2 Status_Negative, Surgery_type_Mastectomy, and ‘Tumour Stage I’ features are important.



2) Which models are suitable for Breast Cancer survival prediction?

Random Forest, Decision Tree Classifier, XG Boost Classifier, Support Vector Classification, and Logistic Regression algorithms are suitable for modeling. Ensemble methods like Random Forest and Gradient Boosting perform well due to their ability to handle complex interactions between various features.

3) What criteria should be used to evaluate the performance of the models?

Metrics like accuracy, precision, recall, F1-score, and confusion Matrix are used to evaluate the performance of the models.

Madhuri Basava
08/05/2024
Applied Data Science
DSC680-T302 (2247-1)

4) How will missing data be handled in the dataset?

Missing data is handled using the imputation technique by removing the rows with null values.

5) What steps will be taken to ensure data quality and integrity?

Data quality will be ensured through rigorous data cleaning, validation checks, consistency checks, and by setting up a data governance framework to monitor and maintain data integrity.

6) How will data privacy be ensured, given the sensitivity of medical data?

The anonymization of patient data ensures data privacy. Here, in the first column, Patient_Id's values in the data set are anonymized.

7) How will the ethical implications of survival predictions be communicated to patients and healthcare providers?

Ethical implications will be communicated through clear, understandable reports and consultations with healthcare providers, ensuring patients are fully informed about their prognosis and treatment options.

8) How will the models be updated with new data and advancements in medical research?

Models will be periodically retrained with new data, and a continuous learning framework will be established to incorporate the latest medical research findings.

9) How will real-time data be incorporated into the prediction models?

Real-time data will be incorporated through integration with IoT platforms and wearable devices, enabling continuous monitoring and dynamic updates to predictions.

10) How will the success of the project be measured in terms of improving patient outcomes and healthcare efficiency?

Success will be measured through metrics such as improved survival rates, patient satisfaction, reduced treatment costs, and enhanced healthcare provider efficiency.

Conclusion:

Breast cancer survival prediction is a crucial business problem that can significantly impact patient care, healthcare economics, and medical research. Implementing advanced predictive models can lead to better decision-making, improved patient outcomes, and more efficient resource utilization. Here the above 5 models are used to predict Breast Cancer Survival rates and the best model (XGBoost) is chosen.

Assumptions:

It is assumed that the data is representative of the population in the data set and is correctly recorded. Handling of missing data through imputation or other techniques is assumed to be correct. Also, assumed that the selected features have good predictive power. The selected models do not introduce biases and are enough to achieve accurate results.

Limitations:

The dataset might not be representative of the entire population due to sampling bias.

The dataset may contain errors. The models may perform well on training data but may be poorly on unseen data. Handling sensitive data involves strict privacy requirements. Health Care Institutions should understand the model and use it effectively.

Challenges/Issues:

- Ensuring the data quality is a challenge.
- Protecting privacy and security of patient data to avoid breaches is a big challenge.
- Predictions need to be fair and unbiased.
- Integration with the health care system requires a great deal of effort.

Future Uses/Additional Applications:

Future uses include personalized treatment plans, combining predictive models with data from wearable devices to continuously monitor patient health and provide real-time risk assessments and recommendations, and enhancing telemedicine to support remote monitoring and management of breast cancer patients, particularly in underserved areas.

Madhuri Basava
08/05/2024
Applied Data Science
DSC680-T302 (2247-1)

Recommendations:

Recommendations are to ensure the models meet clinical needs and ethical standards, involve healthcare providers, patients, and data scientists in the development process, and provide training and resources to clinicians to use and interpret the prediction models effectively. Also, regularly monitor the performance of the predictive models in real-world settings.

Implementation Plan:

To implement the Breast Cancer Survival Detection project, start by collecting and cleaning high-quality, representative data, ensuring it is anonymized and compliant with privacy regulations. Perform feature selection and engineering based on domain knowledge and statistical techniques, followed by experimenting with various machine learning models and ensemble methods, using cross-validation to ensure performance and avoid overfitting. Develop an interpretable model and deploy it using scalable cloud services and receive predictions with explanations. Implement continuous monitoring and a retraining pipeline to maintain model accuracy, and regularly audit the model for biases to ensure fairness. Document the process thoroughly and provide transparency regarding the model's limitations and ethical considerations to build trust and support.

Ethical Considerations:

The model should not discriminate against users based on protected attributes such as race, gender, or age. It should ensure patient privacy and data security to protect sensitive medical information from breaches. Biases in the data or model predictions may lead to unfair treatment of certain user groups. Individuals need to be notified about the limitations, implications, and potential consequences of the prediction. Additionally, the models should be transparent and explainable, allowing healthcare providers to understand and trust the predictions, and ensuring that decisions are made in the best interest of the patients.

References

- *The Breast Cancer Survival Prediction dataset is retrieved from the Kaggle website:* *Survival Prediction Breast Cancer (kaggle.com)*
- *BRCA_Prediction (kaggle.com)*
- *Prediction of Breast Cancer using Machine Learning Approaches - PMC (nih.gov)*
- *Predicting breast cancer 5-year survival using machine learning: A systematic review - PMC (nih.gov)*