Madhuri Basava
07/25/2024
Applied Data Science
DSC680-T302 (2247-1)

Breast Cancer Survival Prediction

Madhuri Basava

Bellevue University

Applied Data Science DSC680

Professor Amirfarrokh Iranitalab

Madhuri Basava
07/25/2024
Applied Data Science
DSC680-T302 (2247-1)

Milestone 1

Project Proposal and Data Selection

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. 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. Machine learning has the potential to predict breast cancer based on features hidden in data.

Research Questions:

Which features are most relevant for predicting Breast Cancer survival?

Which models are suitable for Breast Cancer survival prediction?

Madhuri Basava
07/25/2024
Applied Data Science
DSC680-T302 (2247-1)

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

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. Support Vector Classification
2. Logistic Regression
3. Random forest Classifier

4. Decision Tree Classifier
5. XGBoost

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

Ethical Considerations:

The model should not discriminate against users based on protected attributes such as race, gender, or age. 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.

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.

Madhuri Basava
07/25/2024
Applied Data Science
DSC680-T302 (2247-1)

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)*