Breast Cancer Survival Prediction

CS 619 Final Project

Name: Sairam Kodimella

ID: U01833658

Introduction:

In this project, I intend to do survival prediction for breast cancer patients. Collected dataset has around 400 samples. The problem that I am attempting to solve is whether can we predict whether a person has breast cancer or not given his present medical conditions. To accomplish this, data analysis and transformations are applied initially. Null values and duplicate rows have been removed from data followed by splitting of dataset in 80:20 ratio for training and testing scenarios. Once data is split, 4 data mining algorithms are applied. They are Naïve Bayes, Decision Table, Random Forest and ZeroR. Out of 4 algorithms, Random Forest has given the best accuracy of 93.8% on the test dataset. Python is used to generate some of the visualizations shown below. Weka tool has been extensively used for transformations and data mining algorithms.

Dataset description:

One of the many forms of cancer that can originate in the breast is called breast cancer. Breast cancer is more common in women, but men are not immune to the disease. It is the second largest cause of mortality in women. Predicting whether a person has breast cancer or not at early stages has immense value in saving his/her life. Given sufficient historic patient data, Data Mining algorithms can be powerful enough to predict whether the patient has cancer or not. This data collection includes of individuals diagnosed with breast cancer who have had surgical removal of their tumours. Dataset has been downloaded from Kaggle. Link to download the dataset is https://www.kaggle.com/datasets/amandam1/breastcancerdataset

There are 16 attributes in the dataset. A brief description of each of the attribute is as follows:

Attribute Name	Description
Patient_ID	ID of the patient
Age	Age of the patient
Gender	Gender of the patient
Protein1	expression levels
Protein2	expression levels
Protein3	expression levels
Protein4	expression levels
Tumour Stage	Breast cancer stage of the patient
Histology	Infiltrating Ductal Carcinoma, Infiltration Lobular Carcinoma, Mucinous Carcinoma
ER Status	Positive/Negative
PR Status	Positive/Negative
HER2 Status	Positive/Negative
Surgery type	Lumpectomy, Simple Mastectomy, Modified Radical Mastectomy, Other
Date of Surgery	The date of Surgery
Date of Last Visit	The date of the last visit of the patient
Patient Status	Alive/Dead

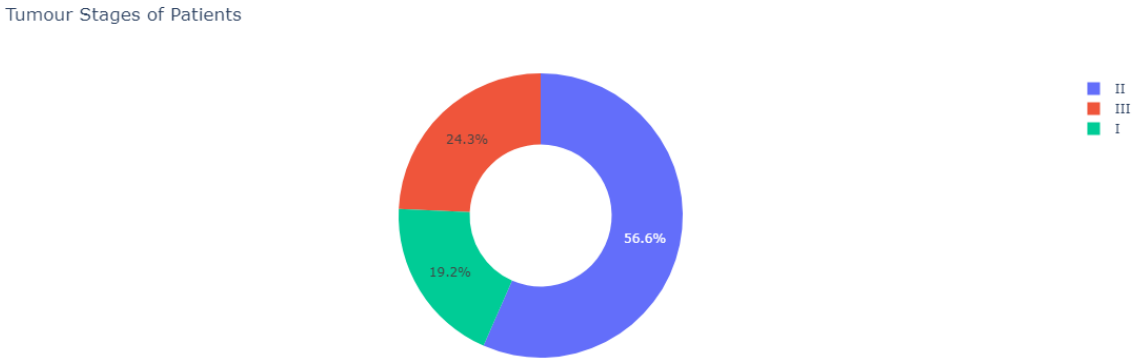
Data preparation:

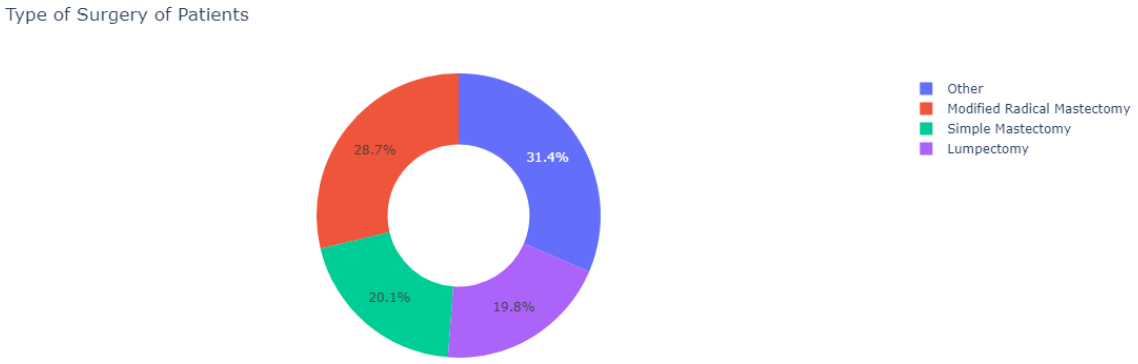
Up on observing all attributes of data, following actions were performed on the data.

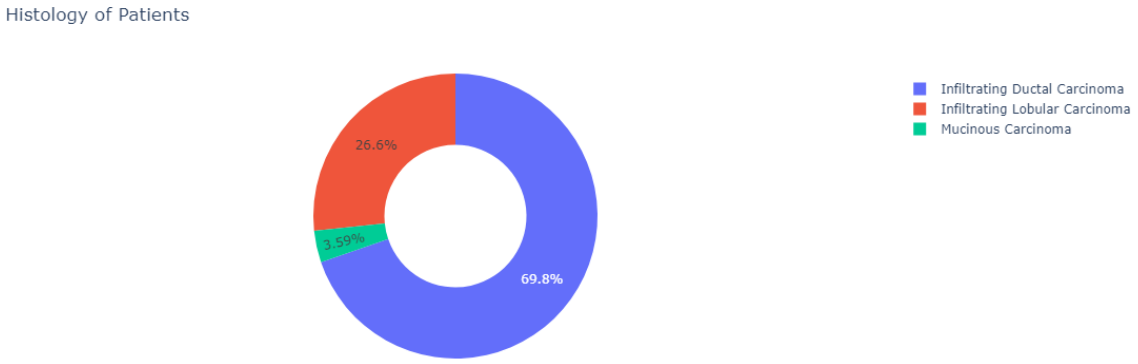
Date will have no effect on whether a patient diagnosed with cancer or not. So, date of surgery and date of last visit are removed from the dataset.

Dataset is examined for null values and null values are only found in Date of last visit and date of surgery attributes. As these are anyhow removed before training and testing.

Pie charts has been drawn for three attributes tumour stage, histology, and surgery type to understand their distribution among data. Plots has been generated in python using plotly. Below are the plots generated.







Once data cleaning is completed, data has been converted into ARFF format using WEKA tool.

Steps to convert csv to ARFF file:

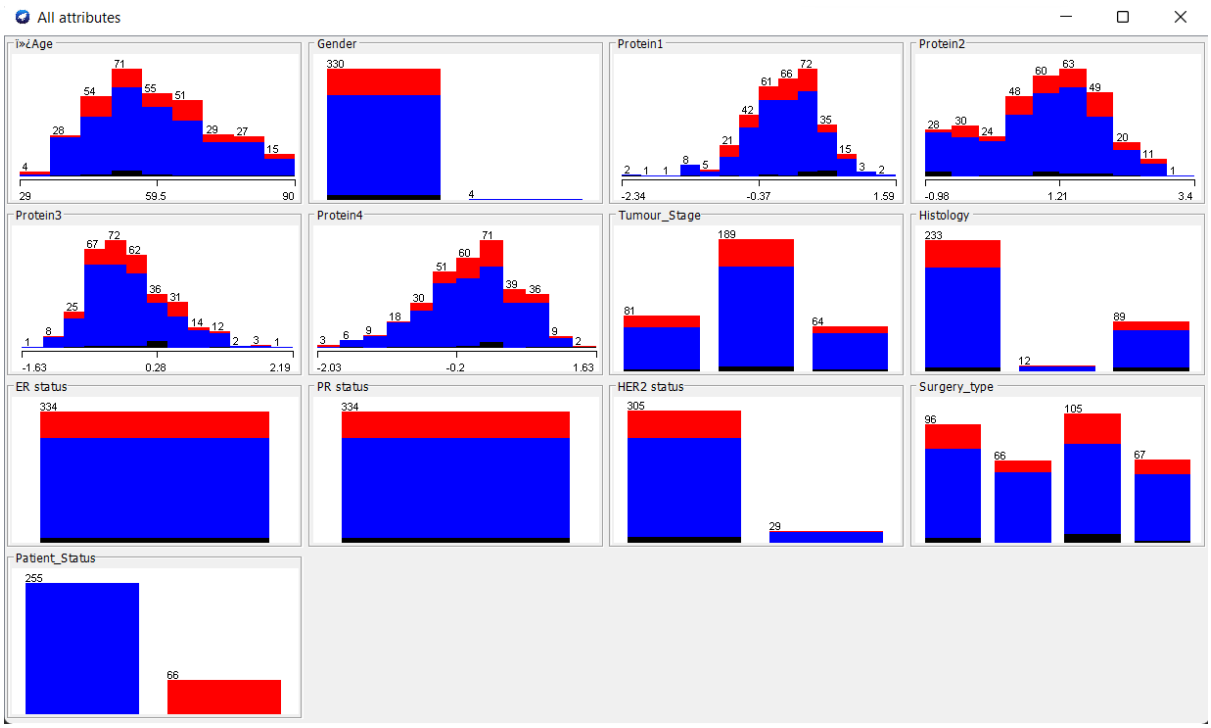
1. Open the Weka GUI Chooser and then click on the tools button in the top menu bar.

2. Click on the Arffviwer. Choose file types to be loaded like, *.csv, *.data.

3. Open *.csv file to view the data and values. Name the file with the. arff extension.

4. Save the file.

Now the ARFF file is opened in the WEKA Explorer. Before starting data analysis, it is important to randomize the dataset and then split the dataset into training and testing. The dataset has been split into 80:20 ratio respectively which means 80% of the overall training samples are taken for training and 20% of the data for testing mode.



Using the RemovePercentage filter in WEKA, dataset has been split in required percentages and saved in separate files. After splitting there are 267 samples in train. Now reopen the train.arff file and start performing data mining operations.

Data analysis and Results:

Following data mining algorithms are implemented to the transformed dataset.

ZeroR:

ZeroR is the most basic classification approach, relying just on the target and ignoring any predictors. The ZeroR classifier predicts just the majority category (class). Although ZeroR has no prediction power, it can be used to establish a baseline performance as a standard for other classification systems.

Results:

Accuracy: 80%

Incorrectly predicted instances: 13

Mean absolute error: 0.3177

Root Mean Square Error: 0.4

Confusion matrix on test data:

	Alive	**Dead**
Alive	52	0
Dead	13	0

Model output is presented in the appendix section of the report.

Verdict:

The model failed to generalise on test data. From above confusion matrix, it can be observed that model predicted alive for all instances.

Naïve Bayes:

The Naive Bayesian classifier relies on the independence assumptions between predictors and Bayes' theorem as its foundation. A Naive Bayesian model is simple to construct and does not require time-consuming iterative parameter estimation, making it especially beneficial for very large datasets. The Naive Bayesian classifier is popular because, despite its simplicity, it frequently outperforms more complex classification techniques.

Results:

Accuracy: 80%

Incorrectly predicted instances: 13

Mean absolute error: 0.3158

Root Mean Square Error: 0.4001

Confusion matrix on test data:

	Alive	**Dead**
Alive	52	0
Dead	13	0

Model output is presented in the appendix section of the report.

Verdict:

Like above algorithm, model failed to generalise on test data. Though accuracy is 80%, it is same as classifying all samples as alive which is not expected.

Decision table:

A decision table is an ordered set of If-Then rules that is more compact and understandable than a decision tree. Decision tables are less complex and need less computing power than decision trees. The classifier rules decision table is described in Building and Using a Simple Decision Table Majority Classifier. The Decision Table classification algorithm summarizes a dataset by utilizing a decision table that has the same number of features as the original dataset. A new data item is assigned a category by searching the decision table for its non-class. Learning decision tables involves selecting the appropriate attributes.

Results:

Accuracy: 78.4%

Incorrectly predicted instances: 14

Mean absolute error: 0.3217

Root Mean Square Error: 0.4048

Confusion matrix on test data:

	Alive	**Dead**
Alive	51	1
Dead	13	0

Model output is presented in the appendix section of the report.

Verdict:

Precision and Recall are zero for dead class. All the predictions are wrong which is not ideal. The model was not able to learn anything.

Random Forest:

Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. On various samples, decision trees are constructed, and their majority vote is used to classify data. Random Forest utilizes ensemble learning.

Results:

Accuracy: 93.8%

Incorrectly predicted instances: 4

Mean absolute error: 0.1565

Root Mean Square Error: 0.2344

Confusion matrix on test data:

	Alive	**Dead**
Alive	52	0
Dead	4	9

Model output is presented in the appendix section of the report.

Verdict:

Model can generalize on the test data. It can predict almost all instances on the test data as well. Apart from accuracy, precision and recall metrics were also good.

Conclusions:

Below table represents each data mining algorithm applied on the dataset along with obtained accuracy. Of all the algorithms, Random Forest gave the best accuracy. In addition to accuracy, it did able to generalize the data very well. Precision and recall metrics were also better than other models.

Algorithm	Accuracy
ZeroR	80
Naïve Bayes	80
Decision Table	78
Random Forest	93.8

Appendix:

Text output produced by Weka

1. **ZeroR**

   === Classifier model (full training set) ===

   InputMappedClassifier:

   ZeroR predicts class value: Alive
   Attribute mappings:

   Model attributes Incoming attributes
   ---------- ----------
   (numeric) ï»¿Age --> 1 (numeric) ï»¿Age
   (nominal) Gender --> 2 (nominal) Gender
   (numeric) Protein1 --> 3 (numeric) Protein1
   (numeric) Protein2 --> 4 (numeric) Protein2
   (numeric) Protein3 --> 5 (numeric) Protein3
   (numeric) Protein4 --> 6 (numeric) Protein4
   (nominal) Tumour_Stage --> 7 (nominal) Tumour_Stage
   (nominal) Histology --> 8 (nominal) Histology
   (nominal) ER status --> 9 (nominal) ER status
   (nominal) PR status --> 10 (nominal) PR status

(nominal) HER2 status --> 11 (nominal) HER2 status
(nominal) Surgery_type --> 12 (nominal) Surgery_type
(nominal) Patient_Status --> 15 (nominal) Patient_Status

Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances	52	80 %
Incorrectly Classified Instances	13	20 %
Kappa statistic	0	
Mean absolute error	0.3177	
Root mean squared error	0.4	
Relative absolute error	100 %	
Root relative squared error	100 %	
Total Number of Instances	65	
Ignored Class Unknown Instances	2	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.800	1.000	0.889	?	0.500	0.776	Alive
	0.000	0.000	?	0.000	?	?	0.500	0.194	Dead
Weighted Avg.	0.800	0.800	?	0.800	?	?	0.500	0.660	

=== Confusion Matrix ===

a	b	<-- classified as

52 0 | a = Alive
13 0 | b = Dead

2. Naïve Bayes
   InputMappedClassifier:

   Naive Bayes Classifier

Attribute	Class Alive (0.8)	 Dead (0.2)
ï»¿Age		
mean	59.1764	58.2273
std. dev.	12.95	13.4127
weight sum	208	50
precision	1.1091	1.1091
Gender		
FEMALE	206.0	50.0
MALE	4.0	2.0
[total]	210.0	52.0
Protein1		
mean	-0.0327	-0.0567
std. dev.	0.5777	0.5021
weight sum	208	50
precision	0.0145	0.0145
Protein2		
mean	0.9208	1.1467
std. dev.	0.9133	0.8974
weight sum	208	50
precision	0.0146	0.0146
Protein3		
mean	-0.1175	-0.0446

std. dev.	0.601	0.5714
weight sum	208	50
precision	0.0149	0.0149

Protein4

mean	-0.0247	0.0941
std. dev.	0.6379	0.5631
weight sum	208	50
precision	0.0126	0.0126

Tumour_Stage

III	49.0	14.0
II	116.0	30.0
I	46.0	9.0
[total]	211.0	53.0

Histology

Infiltrating Ductal Carcinoma	149.0	38.0
Mucinous Carcinoma	9.0	3.0
Infiltrating Lobular Carcinoma	53.0	12.0
[total]	211.0	53.0

ER status

Positive	209.0	51.0
[total]	209.0	51.0

PR status

Positive	209.0	51.0
[total]	209.0	51.0

HER2 status

Negative	189.0	48.0
Positive	21.0	4.0
[total]	210.0	52.0

Surgery_type

Modified Radical Mastectomy	56.0	16.0

Lumpectomy	44.0	7.0
Other	65.0	21.0
Simple Mastectomy	47.0	10.0
[total]	212.0	54.0

Attribute mappings:

Model attributes		Incoming attributes
(numeric) ï»¿Age	-->	1 (numeric) ï»¿Age
(nominal) Gender	-->	2 (nominal) Gender
(numeric) Protein1	-->	3 (numeric) Protein1
(numeric) Protein2	-->	4 (numeric) Protein2
(numeric) Protein3	-->	5 (numeric) Protein3
(numeric) Protein4	-->	6 (numeric) Protein4
(nominal) Tumour_Stage	-->	7 (nominal) Tumour_Stage
(nominal) Histology	-->	8 (nominal) Histology
(nominal) ER status	-->	9 (nominal) ER status
(nominal) PR status	-->	10 (nominal) PR status
(nominal) HER2 status	-->	11 (nominal) HER2 status
(nominal) Surgery_type	-->	12 (nominal) Surgery_type
(nominal) Patient_Status	-->	15 (nominal) Patient_Status

Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.02 seconds

=== Summary ===

Correctly Classified Instances	51	78.4615 %
Incorrectly Classified Instances	14	21.5385 %
Kappa statistic	-0.0294	
Mean absolute error	0.3217	

Root mean squared error 0.4048
Relative absolute error 101.2495 %
Root relative squared error 101.2024 %
Total Number of Instances 65
Ignored Class Unknown Instances 2

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.981	1.000	0.797	0.981	0.879	-0.063	0.524	0.828	Alive
0.000	0.019	0.000	0.000	0.000	-0.063	0.550	0.230	Dead
Weighted Avg. 0.785	0.804	0.638	0.785	0.703	-0.063	0.529	0.708	

=== Confusion Matrix ===

a	b	<-- classified as
51	1	a = Alive
13	0	b = Dead

3. Decision table

InputMappedClassifier:

Decision Table:

Number of training instances: 258
Number of Rules : 1
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 56

Merit of best subset found: 80.62
Evaluation (for feature selection): CV (leave one out)
Feature set: 13
Attribute mappings:

Model attributes	Incoming attributes
(numeric) ï»¿Age	--> 1 (numeric) ï»¿Age
(nominal) Gender	--> 2 (nominal) Gender
(numeric) Protein1	--> 3 (numeric) Protein1
(numeric) Protein2	--> 4 (numeric) Protein2
(numeric) Protein3	--> 5 (numeric) Protein3
(numeric) Protein4	--> 6 (numeric) Protein4
(nominal) Tumour_Stage	--> 7 (nominal) Tumour_Stage
(nominal) Histology	--> 8 (nominal) Histology
(nominal) ER status	--> 9 (nominal) ER status
(nominal) PR status	--> 10 (nominal) PR status
(nominal) HER2 status	--> 11 (nominal) HER2 status
(nominal) Surgery_type	--> 12 (nominal) Surgery_type
(nominal) Patient_Status	--> 15 (nominal) Patient_Status

Time taken to build model: 0.04 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances	52	80 %
Incorrectly Classified Instances	13	20 %
Kappa statistic	0	
Mean absolute error	0.3158	
Root mean squared error	0.4001	
Relative absolute error	99.4138 %	
Root relative squared error	100.0105 %	

Total Number of Instances 65
Ignored Class Unknown Instances 2

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1.000	1.000	0.800	1.000	0.889	?	0.500	0.776	Alive
	0.000	0.000	?	0.000	?	?	0.500	0.194	Dead
Weighted Avg.	0.800	0.800	?	0.800	?	?	0.500	0.660	

=== Confusion Matrix ===

a	b	<-- classified as
52	0	a = Alive
13	0	b = Dead

4. Random Forest
InputMappedClassifier:

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities
Attribute mappings:

Model attributes	Incoming attributes
(numeric) ï»¿Age	--> 1 (numeric) ï»¿Age
(nominal) Gender	--> 2 (nominal) Gender
(numeric) Protein1	--> 3 (numeric) Protein1
(numeric) Protein2	--> 4 (numeric) Protein2

(numeric) Protein3 --> 5 (numeric) Protein3
(numeric) Protein4 --> 6 (numeric) Protein4
(nominal) Tumour_Stage --> 7 (nominal) Tumour_Stage
(nominal) Histology --> 8 (nominal) Histology
(nominal) ER status --> 9 (nominal) ER status
(nominal) PR status --> 10 (nominal) PR status
(nominal) HER2 status --> 11 (nominal) HER2 status
(nominal) Surgery_type --> 12 (nominal) Surgery_type
(nominal) Patient_Status --> 15 (nominal) Patient_Status

Time taken to build model: 0.09 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correctly Classified Instances	61	93.8462 %
Incorrectly Classified Instances	4	6.1538 %
Kappa statistic	0.7826	
Mean absolute error	0.1565	
Root mean squared error	0.2344	
Relative absolute error	49.2658 %	
Root relative squared error	58.5908 %	
Total Number of Instances	65	
Ignored Class Unknown Instances	2	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	0.308	0.929	1.000	0.963	0.802	0.977	0.993	Alive
0.692	0.000	1.000	0.692	0.818	0.802	0.987	0.952	Dead

Weighted Avg. 0.938 0.246 0.943 0.938 0.934 0.802 0.979 0.985

=== Confusion Matrix ===

```
 a  b   <-- classified as
52  0 |  a = Alive
 4  9 |  b = Dead
```

References:

1. https://www.kaggle.com/datasets/amandam1/breastcancerdataset
2. https://waikato.github.io/weka-wiki/faqs/how_do_i_divide_a_dataset_into_training_and_test_set/