Prediction of Breast Cancer Recurrence using different Classifications Methods

Ade Adeoye

Ryerson University

Table of Contents

Abstract

Breast cancer is one of the leading causes of cancer-related deaths in the world, and like many cancer-related illness, early detection is crucial. If discovered early, a patient is highly likely to survive the disease, even if there may be a case of recurrence. This has given rise to a suite of data mining and machine learning efforts attempting to gain an advantage against this ailment by predicting the chances of a patient developing the disease, or if diagnosed already, the chances of the cancer remaining benign or turning malignant.

In this study, we examine a breast cancer data set obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia; and, applying a range of machine learning techniques, we predict the chances of recurrence of breast cancer in 286 patients. The essential part of the data mining procedure conducted is classification, where we categorize the sample according to its original binary split using a range of classification methods. The methods used include decision tree, naïve-Bayes, logistic regression, and neural networks. As conclusive analysis, we conduct a comparison of the metrics of these classification models, examining how the different models performed on the sample data. Finally, we identify the best means of carrying out the classification task for this data set.

\keywords{breast cancer, classification, logistic regression, decision tree, naive-Bayes.}

Introduction

With the increasing generation of data in volume, velocity, and variety (Laney, 2001), taking advantage of the depth and breadth of available data about an event becomes more important, regardless of domain or subject. This emerging field of data exploitation is called data mining or data munging, and in recent times, has turned out incredibly useful in revealing latent information otherwise difficult to obtain from large swathes of data. It takes the approach of searching for relevant patterns and regularities (or irregularities) in large masses of data to elicit useful information that may inform decision-making. It is domain-agnostic and is significantly fruitful if conducted appropriately.

Data mining has also enjoyed some applications in the field of medical research in recent times. It is in the tradition of clinics to collect and consistently update patient information. While the field of machine learning and artificial intelligence grew over the past decades, not a lot of attention was paid to this data stores domiciled with most health institutions. This was for a handful of reasons ranging from government regulations to privacy concerns. However, the widespread availability of innovative computational methods and machines, including those able to handle or condense large image files, and digest data in data warehouses, has encouraged clinical research to buy into data munging with interesting results. These efforts have gone into an array of sub-fields in medicine, with most enjoying different applications of machine learning and artificial intelligence. Applications in cancer research have been particularly notable.

Perhaps it is not unusual that among cancer-related research breast cancer features prominently. As the most common cancer in women worldwide, it seems to enjoy fair dedication in terms of research output. Breast Cancer Research, an international, peer-reviewed online journal which publishes original research, reviews and editorials on breast cancer, estimates that it has continually placed in the first quartile of the ‘Oncology’ category of Journal Citation Reports (Breast Cancer Research: Celebrating 20 Years, 2020). Though research into breast cancer moves at several fronts, its data analysis component has grown vigorously over the past years. This field has seen machine learning techniques applied to both the diagnosis of breast cancer and estimation of chances of recurrence in patients. While it is difficult to determine which of these results are directly related to breast cancer diagnosis, it is quite obvious that there has been significant increase over the years in the papers proposing machine learning solutions, leading inevitably to better improved and more viable results. For example, Advisory.com reported in January 2020 on an algorithm developed by Google which showed

quite improved statistics at the prediction of breast cancer when compared to diagnosis conducted by radiologists. According to the report, “the AI system reduced missed cases of breast cancer in the United States by 9.4% and in the United Kingdom by 2.7%” (Advisory Board, 2020).

According to the World Health Organization, breast cancer registers as the highest incidence of cancer among all ages above 34 in 2018 (World Health Organization, 2020). Data released by the American Cancer Society in 2019 showed that the average risk of a woman in the United States developing breast cancer sometime in her life is about 13%, about 1 in 8 (American Cancer Society, 2019). However, in diagnosing breast cancer, primary diagnosis and recurrence diagnosis are nearly equally important, since the chances that a woman who has had breast cancer runs the risk of relapsing. Primary diagnosis refers to the first case of breast cancer diagnosis in a patient, while in recurrence the cancer reappears often in or around the same area as the original cancer. Several medical studies have been conducted to establish what may cause relapse or recurrence. Many found that sometimes the original treatment is unable to kill all the cancer cells. At other times, external factors reignite the growth of cancer cells which have been dormant for some time. While medicine continues to investigate some of the risk factors responsible for local or regional recurrence, it appears statistical or data mining techniques exploring the association of related variables may prove just as useful. This motivates our research question which explores the strongest indicators suggestive of possible recurrence of breast cancer in a patient.

In this study, we examine a data set of breast cancer patients, some of whom suffered relapses. The data set is from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. Using this data set, we attempt to predict breast cancer recurrence with machine learning techniques. This is a standard classification problem; and in this case, a binary classification problem where the samples consist of two main groups: patients who relapsed, and those who did not. We obtain a probability value associated with recurrence and identify the attributes which may weigh heavily on this probability. In achieving this, we employ classification techniques such as logistic regression, naive-Bayes, decision tree, and neural networks. As a final, related statistical approach to analysing the results, we take an additional step of comparing the metrics of these different classification techniques and establish which provided the best output.

.

Literature Review

As the most frequent incidence of cancer among women above the ages of 34 (World Health Organization, 2020), breast cancer has been gaining significant attention at all levels. This makes it ever more important to employ all possible resources in combating this disease. Among other factors, metastization and recurrence appear to contribute mostly to its mortality rate, as highlighted by Moody et al (2005). Metastization is the spread of cancer to other areas of the body typically remote from the breast, while recurrence describes the reappearance of cancerous cells in a patient who has initially been successfully treated. This implies that predicting diagnosis may be just as important as predicting recurrence of breast cancer in a patient. Though not all people diagnosed with breast cancer will have a recurrence, those who have had this cancer are at a high risk of experiencing a recurrence. In an analysis of 4926 women originally diagnosed with primary invasive breast cancer conducted by Lafourcade et al (2018) between June 1990 and June 2008, they found that 1334 cases had a recurrence after a median time of follow-up of 7.2 years, with 469 dying. Some of the patients who experienced a recurrence had cases with high grade, large tumor size, axillary nodal involvement, and negative estrogen and progesterone receptors. Lafourcade et al also found that “for cases with a medium risk profile in terms of tumor characteristics and lifestyle factors, the probability of dying between 5 and 10 years after diagnosis was 6, 20 and 36% for 0, 1, or 2 recurrences within the first 5 years after diagnosis, respectively.” Additionally, a patient with greater number of lymph nodes with cancer at the time of mastectomy stands a higher chance of relapsing (Sarah G.K., 2018). Typically, the higher the number of lymph nodes, the more the cancer has metastasized away from its source. Similarly, premenopausal women also appear to be at a higher risk of breast cancer than are women who are past menopause.

Unsurprisingly, there is significant literature on machine leaning applications using a breast cancer data set. While some used proprietary data, others have used either the Wisconsin (Diagnostic) Breast Cancer Data or the data set that is the subject of this report. Both are available at UCI Machine Learning repository. Abreu et al (2010) reviewed 17 previous works conducted on the subject of the application of machine learning to breast cancer prediction and found that 9 used publicly available data sets, while the rest used private ones. A lot of attention appears to go to the subject of primary diagnosis or survival, not recurrence. Delen, Walker and Kadam (2005) conducted a comparison of three data mining methods in predicting the survival

chances of breast cancer patients using the 200,000-strong SEER incidence database. They found that decision tree (C5) was the best predictor with 93.6% accuracy, followed by artificial neural networks and logistic regression, with 91.2% and 89.2% respectively. Lundin et al (1999) used a neural network in the prediction of breast cancer survivability after 5, 10 and 15 years. Using eight variables: tumor size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis and patient age in a data set of 951 patients, they were able to establish good accuracy for the neural network to quite some extent. The AUC values for the neural network models for 5-, 10- and 15-year breast cancer specific survival were 0.909, 0.886 and 0.883, respectively. In particular, auxiliary lymph node status weighed quite decently on the rate of false predictions. Belciug et al (2010), on the other hand, compared the performances of an array of unsupervised learning tasks using the Wisconsin Breast Cancer data. The solutions they assessed included k-means, Self-Organizing Map, and a cluster network. Prediction accuracy from this suite of clustering methods ranged from 62% to 78%. Ahmad et al (2013) analysed a different data set from the National Cancer Institute of Tehran in predicting the 2-year recurrence rate of breast cancer. They used three data mining techniques which included decision trees, support vector machines, and artificial neural networks to predict the recurrence of breast cancer and to find which methods performed better. In their results, support vector machines outperformed both decision tree and artificial neural networks in the prediction of recurrence.

The first use of the data set that is the subject of this work was by Michalski et al (1986) who used it to evaluate a multi-purpose incremental learning system called AQ15, one of the earliest forms of supervised learning. Since then, it has enjoyed quite some usage. Abreu et al (2010) discussed some authors who have employed this data set in some of their works. Tomczak (2013) used a classification restricted Boltzmann machine (classRBM) in his analysis of this breast cancer data set. According to his results, using different variations of classRBM, he achieved classification accuracy at least 13% better than human predictions provided by two different oncologists. He also found that the most important attributes for prediction are the histological type of tumor and the level of progesterone receptors in the tumor. A surprising result was that tumor stage over 50mm does not seem to matter in breast cancer recurrence. As a final revelation, he stated that conclusive decisions must still be left to doctors who should examine the results from a clinical viewpoint. Chaurasia and Pal (2014) used a diagnosis system based on RepTree, RBF Network and Simple Logistic, while applying a 10-fold cross-validation method to evaluate the proposed system performance. They were able to obtain a

correct classification rate of 74.5%. RepTree is a decision tree learner which uses reduced error pruning; RBF networks are artificial neural networks that use radial basis functions as activation function, while Simple Logistic is a logistic regression tool. Murti (2012) also analysed the same data set using three rule-based classifiers to predict breast cancer recurrence. After pre-processing to remove missing values, he achieved classification accuracy of 72.27%, 72.72% and 75.17% for the respective classifiers which are RIPPER, Decision Tree and Decision Table with Naïve-Bayes (DTNB). RIPPER is a propositional rule learner proposed by William Cohen. Similarly, Strumbelj et al (2009) used this data set to compare the performance of several well-known classifiers with the evaluation of two oncologists. However, it appears they used a more expanded version of this data set which included a handful more features than are publicly available.

Data set

The data set we used in this work is from 1986 and was originally presented by Zwitter and Soklic (1986), physicians at the Institute of Oncology, University Medical Center, Ljubljana, Yugoslavia. The data is publicly available and downloadable at UCI's Machine Leaning Datasets website. It consists of 286 cases of women who did or did not experience a recurrence of breast cancer. This means it has at least one class attribute which is binary: recurrence-events and non-recurrence-events. There are 201 instances of no recurrent events and 85 instances of recurrent events. The data also has 9 attributes; some ordinal and some nominal, but all consisting of information which may be useful in diagnosing the chances of recurrence of breast cancer in a patient. These attributes include *patient's age*, a parameter identifying the patients' ages at the time of diagnosis; *menopause*, a ternary variable indicating whether the patient is pre-, or post-menopausal; *tumorsize*, an interval describing the greatest diameter in mm of the removed or excised tumor; *invNodes*, the number of auxiliary lymph nodes with visible metastatic breast cancer at the time of diagnosis; *nodeCaps*, a binary variable indicating whether the cancer metastasized into a lymph node or not; *degMalig*, an attribute identifying the histological grade (range 1-3) of the tumor; *breast*, which of the patient's breast the tumor occurred; *breastQuad*, location of the tumor within the breast area (upper left, upper right, central, lower left, or lower right); and *irradiat*, a binary variable indicating whether or note the patient received radiation therapy.

There is not much descriptive statistics we can conduct on this data set composed entirely of categorical variables besides obtaining the mode for each attribute and generating the summary statistics. The mode for each attribute is shown in Table 1, while the description of each attribute is shown in Table 2.

Attribute	Mode
class	no-recurrence-events
age	50-59
menopause	premeno
tumorSize	30-34
invNodes	0-2
nodeCaps	no
degMalig	2
breast	left
breastQuad	left_low
Irradiat	no

Table 1: Mode for each attribute

Name	Description
age	*Patient's age (n years at last birthday) at time of diagnosis*
menopause	*A ternary variable describing whether the patient is pre-, post- or at menopause at the time of diagnosis. This variable is expressed as a function of the age when menopause sets in.*
tumorSize	*The greatest diameter in mm of the removed tumor*
invNodes	*Number of auxiliary lymph nodes with visible metastatic breast cancer at the time of diagnosis*
nodeCaps	*A binary variable indicating whether the cancer metastasized to a lymph node or not*

Name	Description
degMalig	*Histological grade (in range from 1 to 3) of the received tumor*
breast	*A binary variable indicating on which of the patient's breast the original tumor occured (left or right)*
breastQuad	*Location of tumor within breast area (upper left, upper right, central, lower left, or lower right)*
irradiat	*Whether the patient received radiation therapy or not*
Class	*Output class stating recurrence or non-recurrence of breast cancer in the sample patient.*

Table 2: The data set attributes and their descriptions

Attributes		no-recurrence-events	recurrence-events
age	20-29	1	0
	30-39	21	15
	40-49	63	27
	50-59	71	25
	60-69	40	17
	70-79	5	1
menopause	ge40	94	35
	lt40	5	2
	premeno	102	48
tumorSize	0-4	7	1
	5-9	4	0
	10-14	27	1
	15-19	23	7
	20-24	34	16
	25-29	36	18
	30-34	35	25
	35-39	12	7
	40-44	16	6
	45-49	2	1
	50-54	5	3
invNodes	0-2	167	46

Attributes		no-recurrence-events	recurrence-events
	3-5	19	17
	6-8	7	10
	9-11	4	6
	12-14	1	2
	15-17	3	3
	24-26	0	1
nodeCaps	yes	25	31
	no	171	51
degMalig	1	59	12
	2	102	28
	3	40	45
breast	left	103	49
	right	98	36
breastQuad	central	17	4
	left_low	75	35
	left_up	71	26
	right_low	18	6
	right_up	20	13
irradiat	yes	37	31
	no	164	54

Table 3: Frequency table of categorical attributes in data set

```
```{r}
summary(bdata)
```
```

```
                class        age        menopause     tumorSize   invNodes   nodeCaps   degMalig   breast        breastQuad  irradiat
 no-recurrence-events:201   20-29: 1   ge40   :129   30-34  :60   0-2  :213   no  :222   1: 71    left :152   central  : 21   no :218
 recurrence-events    : 85  30-39:36   lt40   :  7   25-29  :54   12-14:  3   yes : 56   2:130    right:134   left_low :110   yes: 68
                            40-49:90   premeno:150   20-24  :50   15-17:  6   NA's:  8   3: 85                left_up  : 97
                            50-59:96                 15-19  :30   24-26:  1                                   right_low: 24
                            60-69:57                 10-14  :28   3-5  : 36                                   right_up : 33
                            70-79: 6                 40-44  :22   6-8  : 17                                   NA's     :  1
                                                     (Other):42   9-11 : 10
```

Table 4: Summary statistics of attributes

The frequency table for each attribute split along the *class* variable is shown in Table 3, and the summary statistics for the data set is shown in Table 4. Besides showing the count of each

variable in the data set, these tables also show the distribution of each variable relative to the *class* variable. For example, when *class* is split across *age*, we notice an imbalance in the data set is. Most patients lie between the ages of 40 and 69. Using the *table* function in R, we further examine the different distributions of each attribute against the *class* attribute in raw numbers and in proportions. For example, after splitting the *class* attribute along *age*, we can tell that for patients between the ages of 40 to 49, there are more than 2 times the number with no recurrent cases than there are with recurrent cases. Though this is about the same for patients between the ages of 60 to 69, the split is worse for patients between the ages of 50 to 59. For all other patients, age distribution relative to recurrence-events seems about even. Splitting along *nodeCaps* or *degMalig* also reveal quite disproportionate values for each class. For patients who experienced metastization into a lymph node, the split seems a little even. However, those who did not were significantly more for patients with no recurrent event. And while for patients with grade 1 and 2 tumors the split between those with “recurrence-events” and those without was heavily biased towards the latter, patients with grade 3 tumors seem about an even split along class lines. Only the *breast* attribute has about the most even split of all the attributes in the data set

This 2.36:1 ratio of the non-recurrence class to the recurrence class shows that the data set is highly imbalanced. This means the samples are heavily biased towards one class. In this case, the samples contain more women who suffered non-recurrent cases than it does women with recurrent cases. Such imbalance could have a detrimental effect on our subsequent machine learning analysis. Usually, it has the effect of biasing the prediction towards the over-represented class. This means we need to take special precaution by adequately treating this imbalance using the range of randomly sampling methods available. Some of what we could do include oversampling the recurrent-events cases, undersampling the no-recurrence-events cases, or both. There are also techniques which generate random synthetic data to fill this gap. Packages available in R which treat this unusual case of class imbalance include *SMOTE* and *ROSE*.

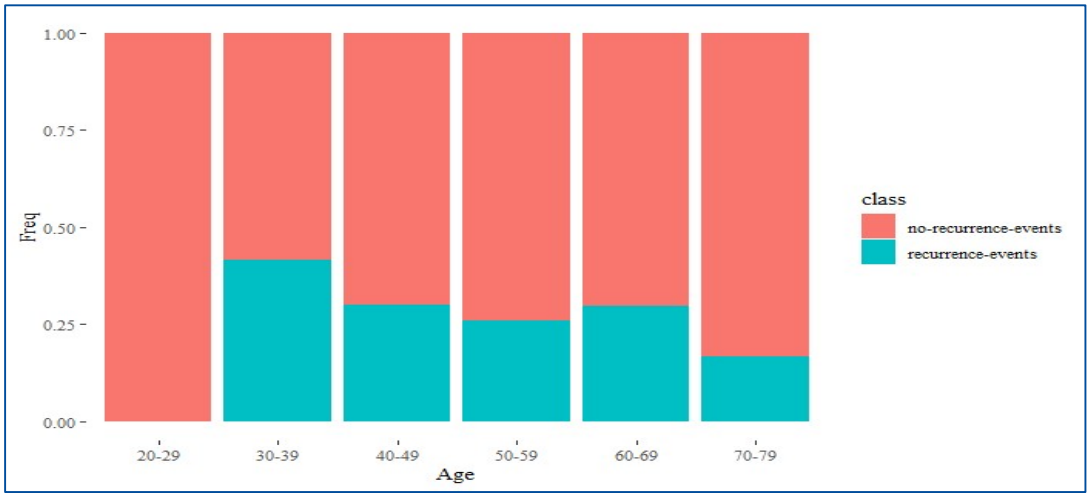

Figure 1: A bar plot of the different age frequencies show that the data set is imbalanced

The bar charts in Fig 2 show a split of a*ge* across both *class* and *tumorSize*. Most of the data is comprised of records with “no-recurrence-events”. We also observe that most of the surveyed patients have tumor sizes between 10 and 39mm. This is where the bulk of the data lies. For some other *tumorSize* facets such as 0-4mm, 5-9mm, 45-49mm and 50-54mm, very few patients are represented.

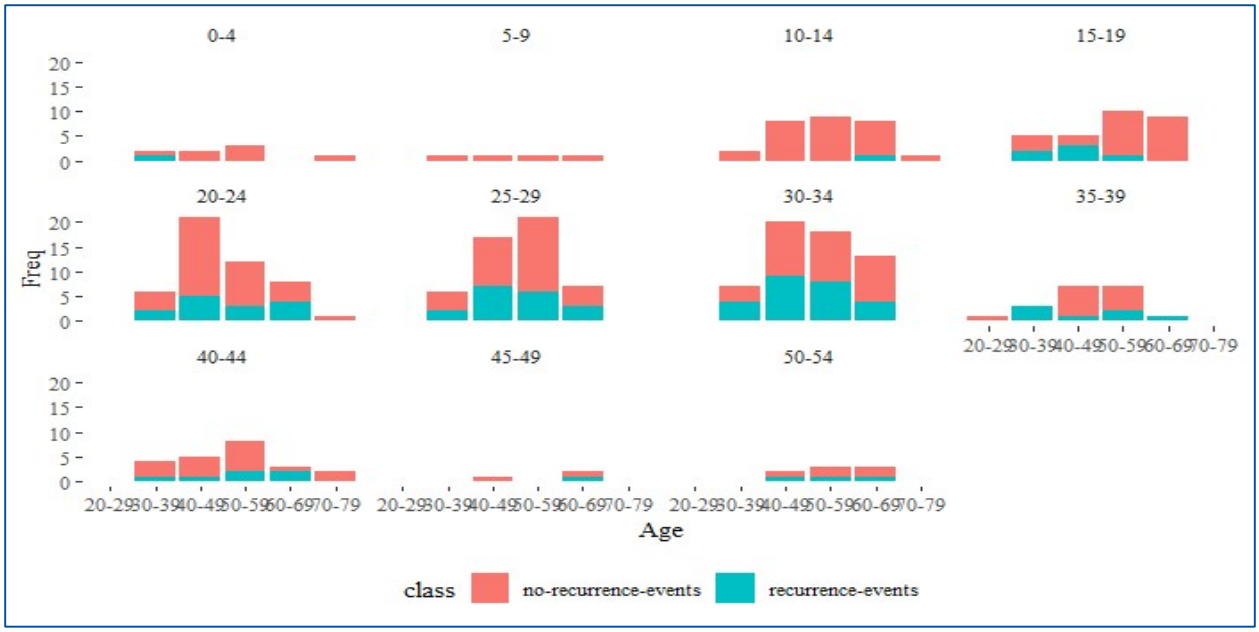

Figure 2: Most patients have tumorSize between 10 and 39mm. Some are also not well represented for this variable

The relationships between other attributes in the data set also reveal the same bias towards "no-recurrence-events". For example, the mosaic plot in Fig. 3 shows the relationship between *tumorSize* and *invNodes*. Here, we see that there are larger blocks for "no-recurrence-events", though the matrix itself is quite sparse. For most cases, there are fewer (0-2) lymph nodes infected, even when the tumour size is large. For other cases of *tumorSize*, there are progressively fewer lymph nodes involved. Though there may be a correlation, this is not immediately visible.

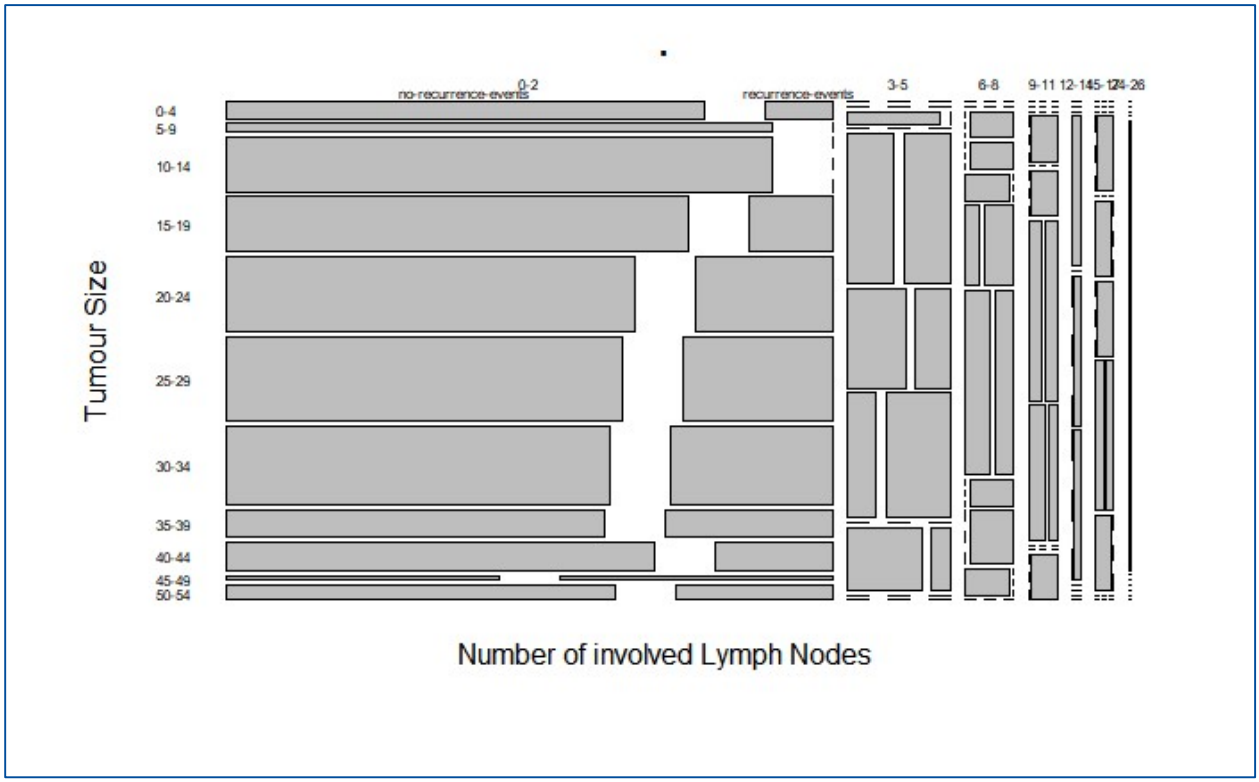

Figure 3: A mosaic plot shows the relationship between tumorSize and invNodes. There are progressively smaller blocks for increasing values of both.

There is a total of 9 missing values in the data set. There are 8 missing values under *nodeCaps* and 1 under *breast*. Missing values may or may not constitute a problem in the analysis of a data set. This depends on whether they are missing at random or intentionally omitted. The attributes containing missing values in this data set suggest that these missing values are not intentionally omitted. Since the data set is already a small one, it is desirable to attempt imputing these missing values. For imputing, we tried three different techniques mainly for comparison purposes: using modal values, k-nearest neighbours and multiple correspondence analysis

(MCA). All three imputing techniques produced nearly the same answers. As a result, we settled on the first approach: using the modal values.

Approach

The development approach is divided into six steps. These include data preparation, exploratory data analysis, data balancing, model fitting, model selection and analysis, and model evaluation and reports. The first five steps require an exhaustive approach and involve learning every detail about the data set. The last step deals with the reporting of the results obtained in the previous stages. The steps are also not conducted in isolation, as there exists a feedback loop which opens model selection back to exploratory data analysis.

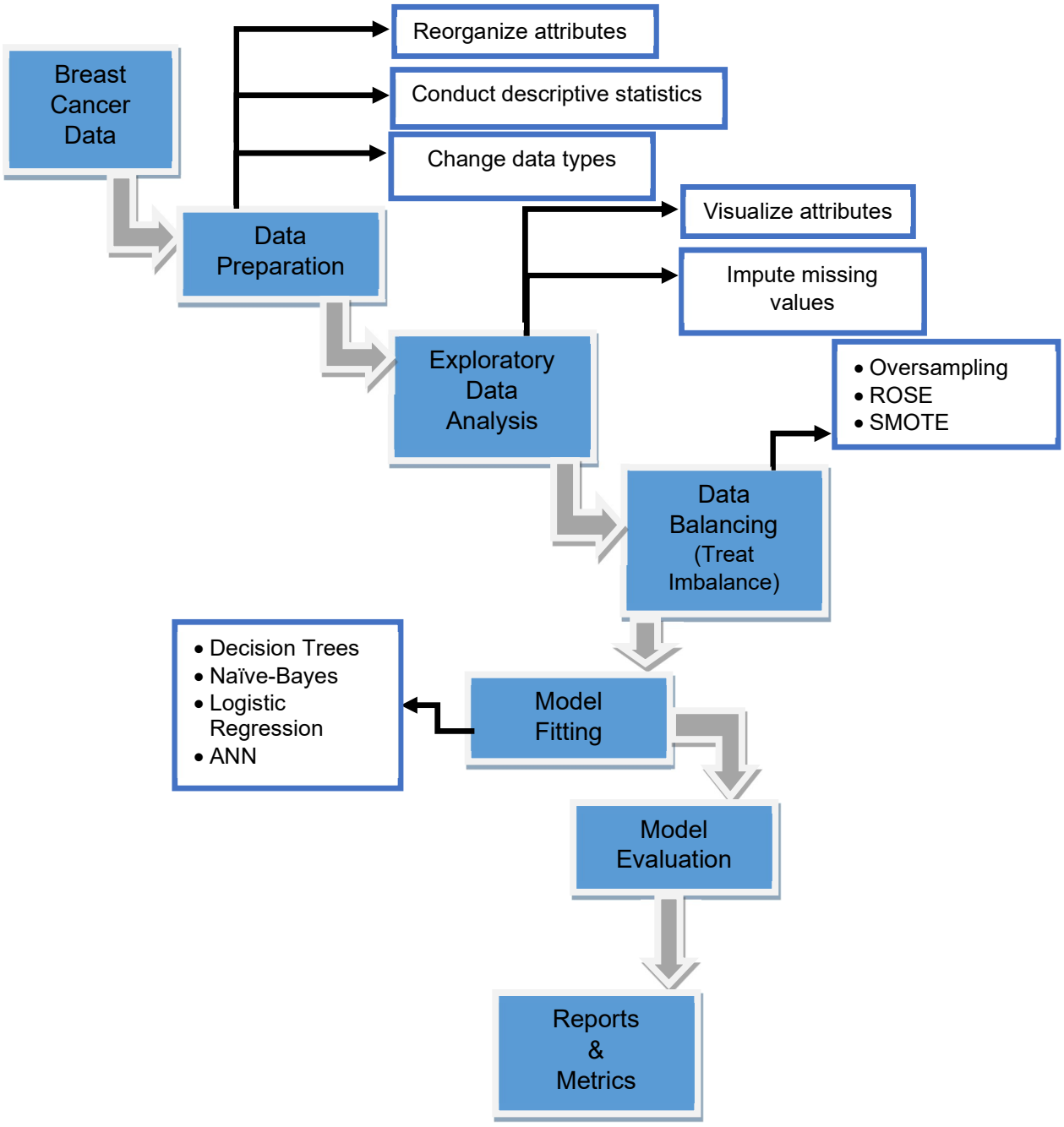

Figure 4: An outline of the workflow for analysing the Breast Cancer data set

Data Preparation

In analysing this data set, we follow the process outlined in Figure 4. Firstly, we load the data into the R environment, and thereafter prepare it by reorganizing and changing data types where necessary. We then conduct some basic descriptive statistics tasks. Imputing is saved for the exploratory stage so that we could gain more insight into the data distribution prior to filling missing values.

We use the *readr* package to load the data set into the R environment. This package is from the *tidyverse* universe of packages and handles input better than R's base package. After loading into R, we prepare the data by changing some of the data types. The data set has all categorical values which exist as string or numeric types in the original data set. However, some are in fact ordinal types, while the others are nominal. For example, *degMalig*, the histological grade (in range from 1 to 3) of the received tumor, exists in the dataframe as a numeric variable. It is in fact best handled as an ordinal variable and carries little meaning when expressed as an integer. This means that a degree of malignancy of 3 is higher than that of 2, which in turn is higher than that of 1. The same applies to *tumorSize*, which represents the greatest diameter in millimeters of the removed tumor. Other variables such as *breastQuad*, *irradiat,* etc., are converted into nominal forms. All attributes are converted into factor variables in R, with the ordinal types becoming ordered factors.

The descriptive statistics task we carry out entails obtaining the mode for each attribute and examining how many missing values are present in the data set. Using the *table* function, we examine the distribution of each attribute along the class attribute or any other attribute, both in proportions and in raw values. The *summary* function, on the other hand, was used to examine the number of missing values in each attribute. A total of 9 missing values were found.

Exploratory Data Analysis

Our exploratory data analysis step includes visualizing the distribution of the different attributes and imputing missing values. We conduct exploratory data analysis by visualizing the data using bar charts and exploring the strengths of its different features using MCA. We impute missing values using the mode, but only after comparing the results generated using three different methods: mode, k-nearest neighbours and MCA. By saving imputing for this stage, we get a global view of the data distribution before imputing.

We use *ggplot* for visualization where necessary. The data set has all categorical variables, so some of the visualizations entailed faceting over another variable. We create bar charts for the important variables and employ faceting when we need to examine the relationship of one variable across another.

We explore three different approaches for imputing missing values. These include replacing with the mode of the attribute, using k-nearest neighbours, and clustering using multiple correspondence analysis (MCA). All produce nearly the same answers. However, we settle for the first approach in our solution: using the mode.

There is a total of 9 cases with missing values under the *nodeCaps* and *breastQuad* attributes. Replacing the missing values with the mode of the attribute is quite straightforward. We use the *sapply* function to perform a column-wise operation on the data set and to obtain the mode for each attribute. The modal values are shown in Table 1.

k-NN examines the nearest neighbour of any instance in the data set. The k-NN computation we use employs a distance measure developed for categorical or mixed data called Gower's distance. This is given in the equation below.

$$d_{p,q} = \frac{\sum_{i=1}^{n} w_i d_i^2}{\sum_{i=1}^{n} w_i}$$

where w_i is a weight for the ith variable. w_i takes the value of 1 when both p and q are known; otherqise, it takes zero. d_i^2 is the square of the distance between the ith value of the two observations p_i and q_i. d_i is given as:

$$d_i = \frac{|p_i - q_i|}{R_i}$$

The *kNN* function available in the *VIM* package allows imputing using Gower's distance. We obtain the nearest neighbour result using this function and compare with our earlier results obtained using the modal values.

MCA is achieved by examining the data set's principal components. Though the attributes are not numerical, we find that there are alternative approaches to handling categorical types when examining "principal components". Chavent et al (2017) discuss how to handle multivariate

analysis of mixed data in their paper **Multivariate Analysis of Mixed Data: The R Package *PCAmixdata*. The package extends standard multivariate analysis methods to incorporate data of mixed types. It offers the function *PCAmix* which makes no distinction between ordinal and nominal variables and can be used for principal component analysis of both using MCA via Generalized Singular Value Decomposition. Using this function help us achieve a truly global view of the data set where we find that the first 8 dimensions retrieve about 39% of the data's total inertia. Admittedly the data set is small, and we do not necessarily achieve any gain by reducing the number of attributes, nevertheless using MCA provides us with a convenient clustering of the different attributes along the principal components and provides a global view of the data set. With this clustering, we gain better insight into the relationships inherent in the data set. For example, when we plot the first two principal components, the data distribution in terms of age shows a well-distributed and insightful plot. Most of the older populations are located at the top of this age distribution, while the younger population are located somewhat at the bottom.

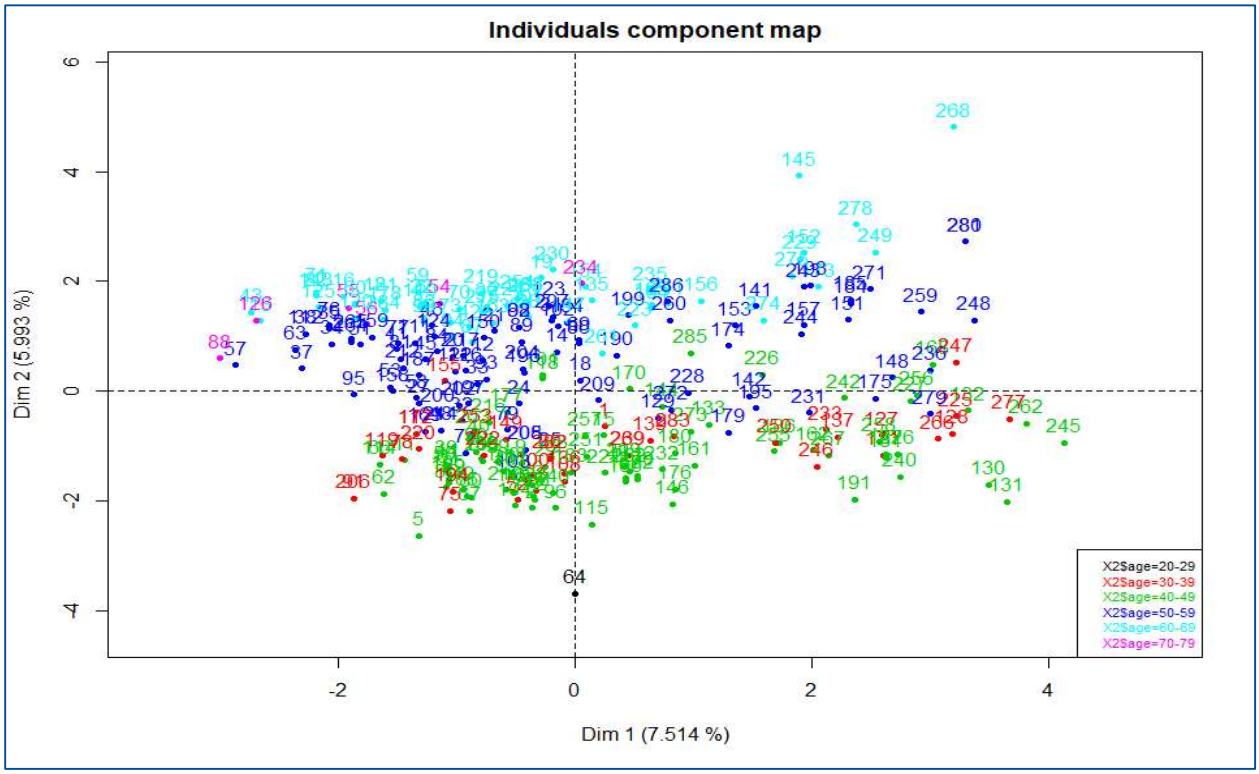

Figure 5: Using PCAmix shows a well-distributed plot of the age variable in terms of the principal components

Similarly, when we examine the distribution of *nodeCaps*, we see that the “no” cases are clustered on the left, while the “yes” cases are clustered on the right. One of our missing values (the blue dot) is located around the middle of this plot. When examined using the PCA coordinates, it appears to be grouped with the “yes” cluster. This is the same result we obtained using both the kNN method and the modal values.

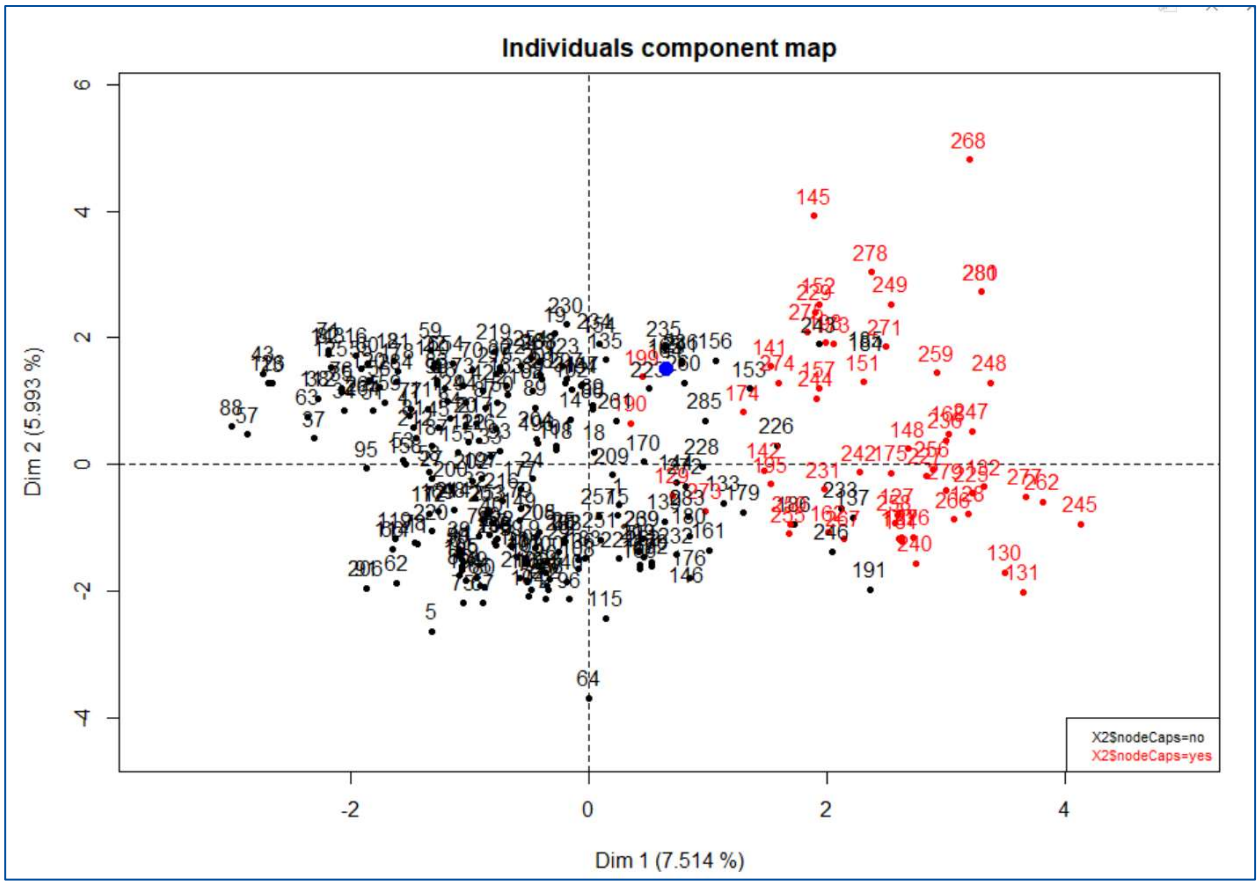

Figure 6: Cluster distribution of nodeCaps. The blue dot is a missing value in the data set and appears to be grouped with the "yes" cluster.

Data Balancing

We observed earlier that the data set is imbalanced because there are more *no-recurrence-cases* than there are *recurrence-cases*. In this step, we apply sampling techniques to balance the data set. At first, we use all three approaches of undersampling, oversampling or synthetically generating random fillers. We find that all different approaches produce near similar results, though the results seemed best when we sample up. As a result, we have sampled up in many cases where we need to test the model.

By treating imbalance, we avoid an unproductive scenario where the machine learning model is heavily biased towards *non-recurrence-events* which is in the majority. We also examine different techniques of random sampling and observe which is much better for the data set.

The *ROSE* and *SMOTE* packages in R offer functions with which randomly sampled data can be used to balance out imbalanced data sets. The *caret* package also makes these functions available to its *train* function when an imbalanced training set is used. All these options are used both before and while building the models in R. The *train* function in the *caret* package makes it easy to apply sampling while simultaneously conducting cross-validation using the training set. We follow this procedure in the creation of our models.

Model Fitting

In fitting the model, we build firstly a decision tree. Next, we fit a naive-Bayes model and then a logistic regression model. Finally, we build a neural network for the classification task. All models are created using R.

We use the *ctree* function from the *partykit* package to create the decision tree. For all other models, we use the *caret* package. This package offers an array of functions for preprocessing, model building and model evaluation. The neural network was created using the *nnet* and *neuralnet* packages.

Cross-validation is done at each model fitting step to test the robustness of the derived model, and to obtain average metrics which reflect the overall health of the model. A 10-fold cross-validation was conducted while creating each model. This involves dividing the model into ten different chunks and using, in repeated cycles, all 9 chunks for training and the remainder for testing.

For most of the models, we try different sampling techniques while fitting to get an idea which sampling technique could be best.

Model Evaluation

Finally, we evaluate each model and put together a summary report of the metrics obtained while fitting.

The step of model evaluation remains open to revision as we explore the metrics and establish if there are adjustments that could improve the model. For example, the models were first fitted

without using *degMalig*, *tumorSize*, and *invNodes* as ordered factors before the decision was made to convert these into factors and then refit the models. We observed that this did not have a lot of impact on the results of the models. Similarly, some of the features were eliminated and the model refitted without these features to see if there were any improvements. Our evaluation stage show, for instance, that the logistic regression model achieves better output with only a few features than with all the features. Finally, we explore some feature engineering steps and check how the results affect model output. We observe that we can create a neural network only after dummifying the categorical features in the data set.

The metrics we used to evaluate each model include sensitivity (recall), specificity, precision, the number of false negatives, among others. We explore different designs which produce the best metrics and try to isolate the most useful features where necessary.

Reports

All the models are evaluated using their respective metrics. These include ROC, AUC, precision, recall and accuracy values. For example, one of the key objectives is to minimize the number of false negatives, i.e. the number of patients who could relapse, but for whom the model suggests the opposite.

We also examine, for the naïve-Bayes model, which sampling technique delivered the best results.

As a final step, we compare the accuracy values for each model to check if they are statistically different from each other.

Results

We built four different models using decision tree, naïve-Bayes, logistic regression and artificial neural networks. The lowest accuracy on the test data obtained from the models was 66% on the logistic regression model, while the highest was 73% using the naïve-Bayes model. Though the models produced a high average specificity value of 79%, the sensitivity average was quite low at a value of 50%. This is because the false negatives were quite high compared to the true positives. The naïve-Bayes model attained a precision score of 56%, the highest among the different methods, while the logistic regression model recorded the lowest value at 48%. Of the classification models examined, the naïve-Bayes model seemed to perform best. However, statistical tests showed that the accuracy mean of the respective models were not statistically different from each other.

In qualifying a model, we not only look at the accuracy, specificity, precision and sensitivity values, we also look at the one which minimizes the false negatives. False negatives represent the number of patients who could suffer a relapse but who are wrongly classified as not likely to suffer a relapse. Keeping this number low ensures that we do not misinform patients who are likely to experience a recurrence.

Our initial assessment of the data set showed that it was highly imbalanced. We needed to treat this imbalance appropriately and used the ROSE package for this operation. We created an oversampled data set, and another data set which combined both oversampling and undersampling. Going into the analyses, we observed the oversampled data set appeared to provide better accuracy for this data set. Using the naïve-Bayes model, we tested this hypothesis and eventually confirmed that the oversampled data set was most suitable. In using oversampling and undersampling techniques in modeling building, we leveraged the availability of a sampling option while training which is part of the *caret* package

Initial runs of the decision tree model were conducted using the doubly-sampled data set, and the synthetically generated data set. The doubly-sampled data set combines both oversampling and undersampling. The doubly-sampled data set produced a sensitivity of 42% and a specificity of 82%, while delivering an accuracy of 70%. There were only 15 false negatives in all. The synthetically generated set, on the other hand, produced 11 false negatives. It delivered a specificity of 72% and a sensitivity of 58% with an overall accuracy of 68% on the test set. These initial runs were conducted to get a taste of model building. Cross-validation was conducted on the training set using the *caret* package and the output seemed okay. Our final

prediction on the test data which we achieved by sampling up and using decision tree produced 18 false negatives, a rather poor of sensitivity of 31%, a precision value of 53%, and a good specificity value of 88%. The overall accuracy on the test set was 71%, which seems just as good as predicting *no-recurrence-events* for every instance in the test data. While we could say that the model has not done excellently, we were able to achieve one of the objectives set out for this study: identifying the variables that contribute to determining the recurrence of breast cancer in women. The model identified these as *nodeCaps*, the number of infected lymph nodes and *degMalig*, the histological grade of the tumor. These were the two top features identified by the tree model. This conclusion is consistent with results obtainable from cancer research as highlighted under the literature review. It is also consistent with the result obtained by Tomczak (2013) as earlier highlighted in the literature review.

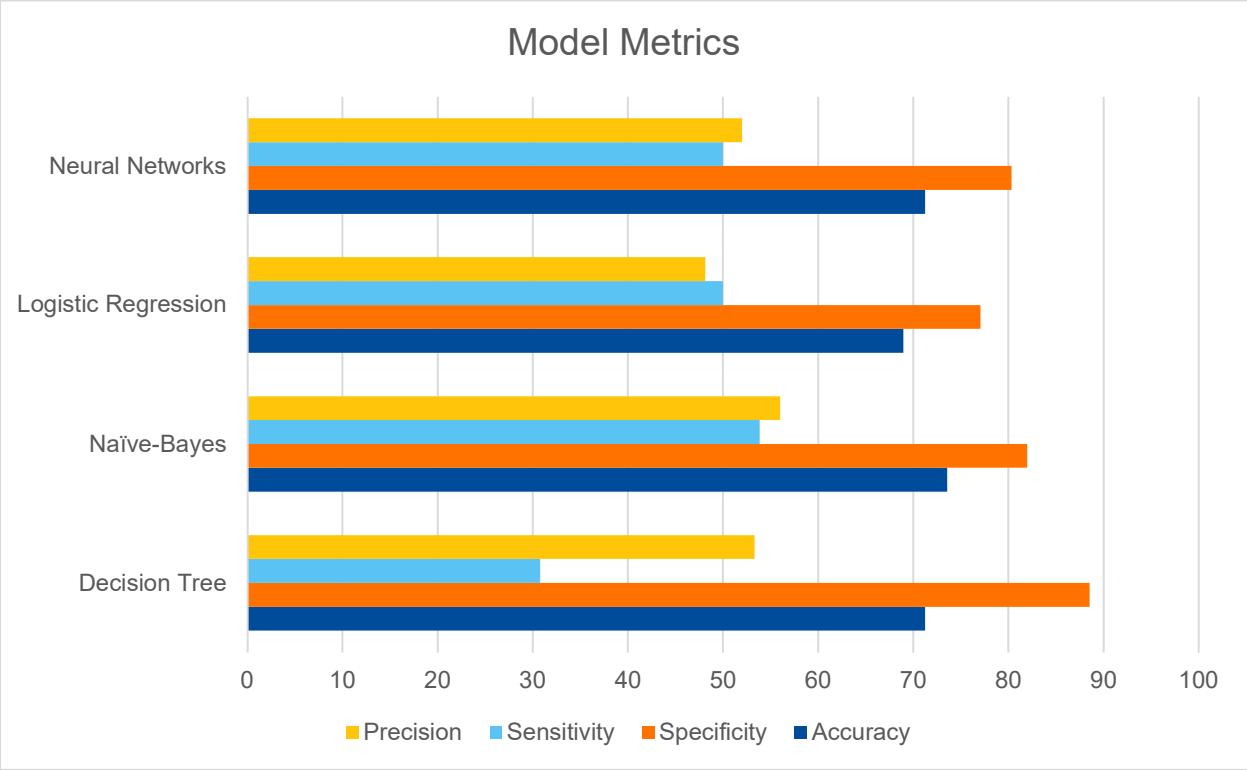

Figure 7: Metrics showing the accuracy, sensitivity, specificity, and precision values for each model

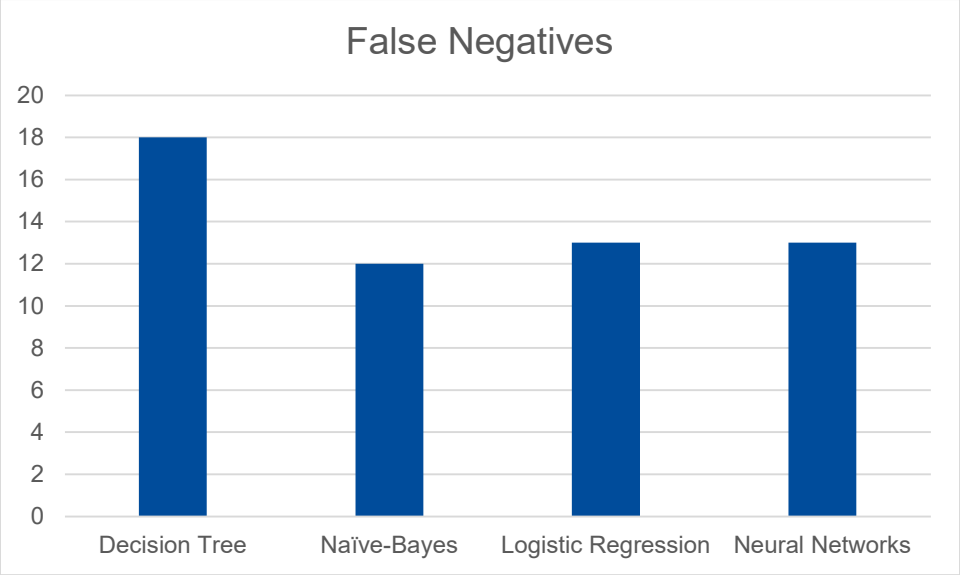

Figure 8: Plot showing false negatives for each model

The naive-Bayes classifier produced the best accuracy of all the classification methods. It delivered a specificity of 82% on the test set, precision of 56%, and a sensitivity of 54%, while maintaining only 12 false negatives. Its overall accuracy was 74% which is much better than random selection. Again, we sampled up using the functionality provided by the *caret* package. The output of this model is only a little better than the previous model. It delivered better specificity and sensitivity, and better overall accuracy even though it has one more false negative than the decision tree model. Like the previous model, it also identified the variables of most importance in model building.

To obtain richer insight and more as an exploratory procedure, we combined every two possible features of the data set so that we can check which delivered the best metrics. We were able to identify *age* and *tumorSize* as two features which delivered well on certain metrics as it relates to the data set. When these two were used to develop a naive-Bayes model, we achieved an accuracy of 69% while maintaining only 10 false negatives.

Additional checks confirmed that the models were quite sensitive to the sampling techniques used. As a result, further examination was conducted to compare results from all the different sampling methods, including oversampling, undersampling, ROSE and SMOTE. We found that, for this data set, sampling up had a greater tendency to produce better accuracy values. Sampling down, on the other hand, was greater disposed to producing lower accuracy models.

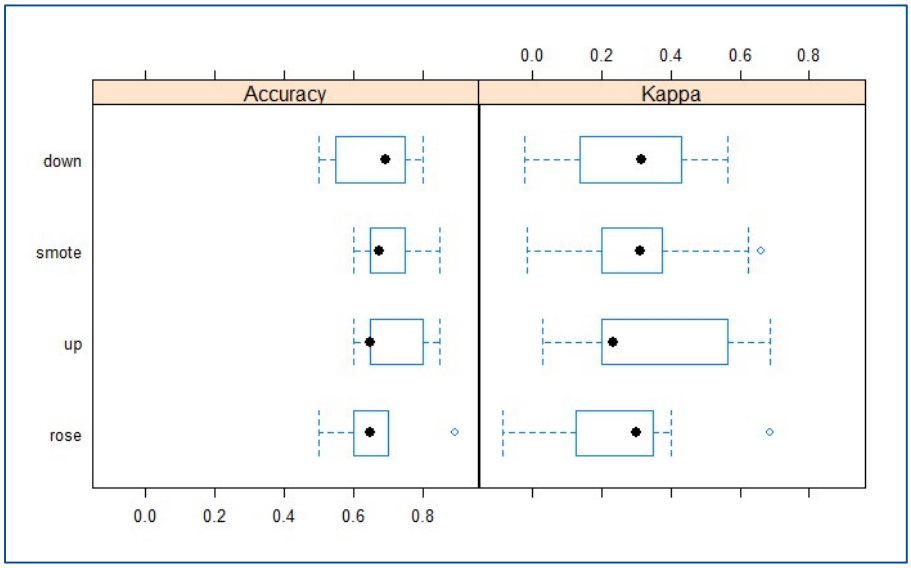

Figure 9: Output results from the naïve-Bayes model using different balancing techniques. "Up" sampling seemed to produce better accuracy values, while "down" sampling produced lesser accuracy values for this data set.

Another classification method considered was logistic regression. Before fitting the model, we conducted stepwise regression to identify which could be the most interesting features for analysis, or the features which should produce the best metric. We identified this as *nodeCaps*, *degMalig* and *irradiat*. We built the model first using these three features, and then a second model using all available features. The first model produced an accuracy of 69% on the test set, a sensitivity of 50%, precision score of 48%, and a specificity of 77%. The false negatives were only 13. The cross-validation results also seemed good. This model appeared to perform better than the model with all the features included. The full model did slightly less than the former model yielding 14 false negatives, a sensitivity of 46%, a specificity of 75% and an overall accuracy of 67% on the test set. Both models, however, did not do as well as the naive-Bayes model, for example.

The final model developed trained a neural network for classification purposes. We used the *nnet* package available through the *caret* package. Neural networks are highly dependent on initial variables, so we set the seed to 80 to conform with our earlier steps. We created a grid of decay and size as hyperparameters with which to tune the model. In the end, the model used size of 1 and decay of 0.5. We found that using multiple hidden layers did not improve model output. Unexpectedly, the neural network did considerably well on the training set producing an

accuracy of 76%. However, it did only a little better than some of the other models on the test set. We recorded 13 false negatives, an accuracy of 71%, a precision score of 52%, a specificity of 80% and a sensitivity of 50%. Prior to passing the features into the neural network, the features were preprocessed into dummy variables through one-hot encoding. The variable importance plot showed some of the dummy features that were important in the training of the model. This feature importance plot suggested additional ideas of feature engineering that may be used to improve any model. For example, *invNodes9-11* and *degMalig2* are features of very high importance in the neural network model. By isolating these two variables and a handful more, we supposed that feature engineering could provide additional opportunities of building a better model.

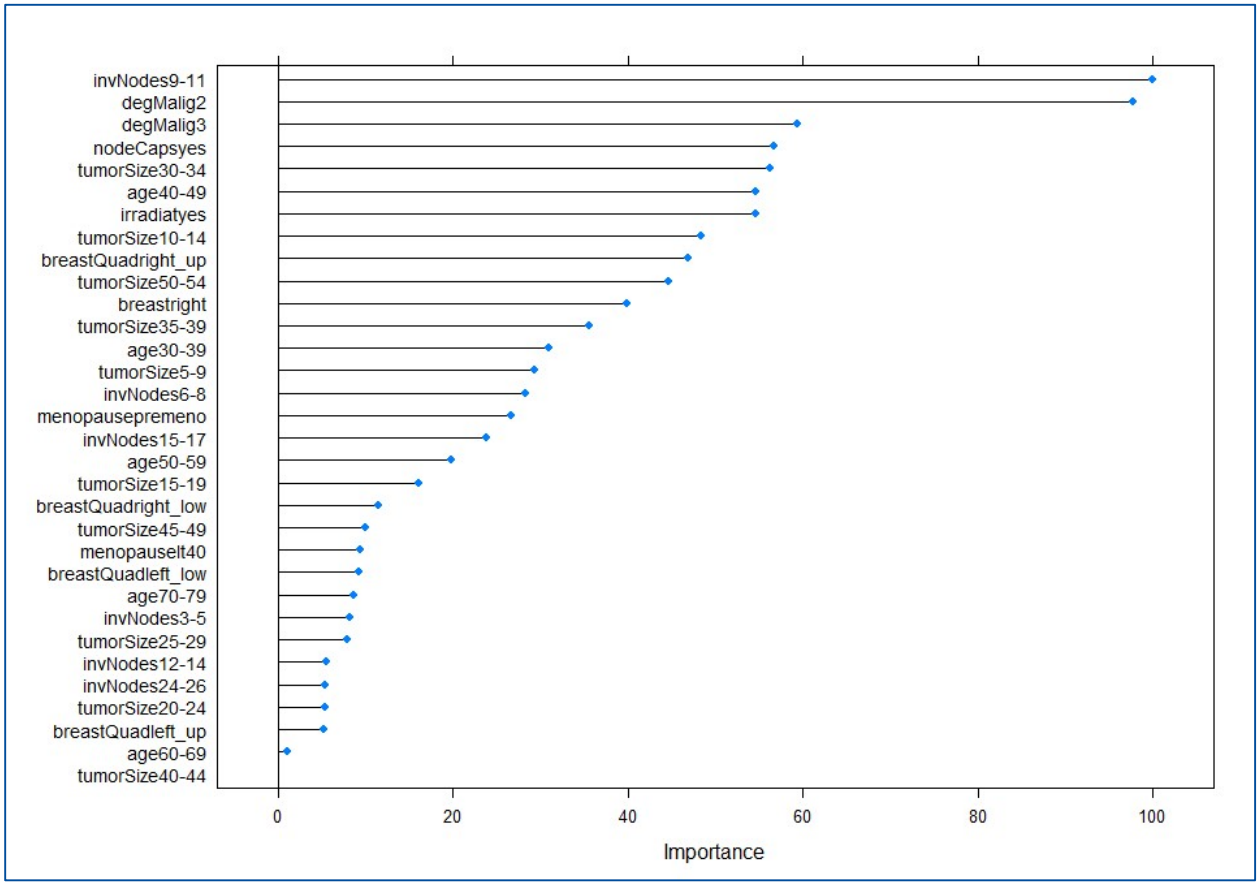

Figure 10: A chart of variable importance for the variables used by the neural network.

We used one-hot encoding to explore these features suggested as important by the neural network results. We focused on the ones which topped the list of variable importance in the plot

of the dummy variables fed into the neural network (Fig 10.). Using these modified features with a logistic regression model, we obtained an output which was no better than the earlier logistic regression results obtained. we were able to obtain another model slightly better than that earlier derived from the purely categorical features. However, a t-test showed that this result was not statistically significant than the earlier result. Additional exploration of dummy variables in the data set derived from one-hot encoding, including attempts at feature crossing, did not produce further interesting results.

As a final check, the accuracy results obtained from each model were compared against each other using *ANOVA*. The aim was to check if any of them is statistically different from the other. The result led us to fail to reject the null hypothesis that the model results are any different from each other. As a result, even though the naïve-Bayes model seemed to produce the best output, its results did not seem entirely different from those produced by the other models.

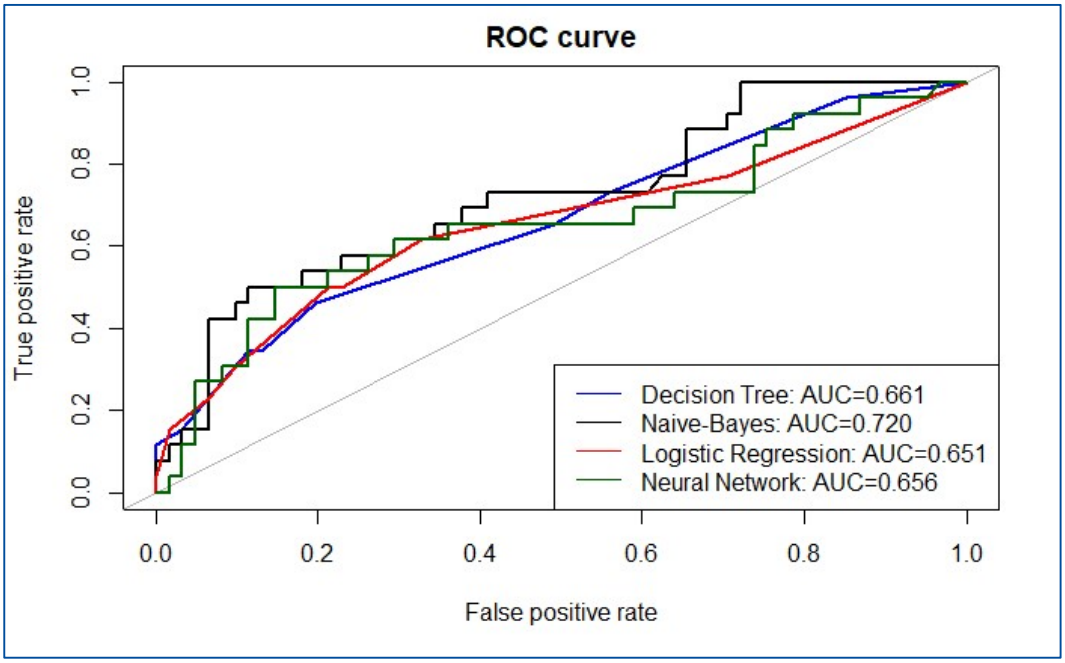

Figure 11: The ROC curves for each fitted model. Naive-Bayes recorded the highest AUC.

We see from the ROC curves for each model that the naïve-Bayes seemed to have done best, while the other models ranged in accuracy between 65-66%. Given the imbalance in the data

set, these model outputs can be considered good enough. For most, their respective accuracies are better than results obtainable assigning no-recurrence-events to all cases.

The F1 score combines the precision and recall values. By using a harmonic mean, it reduces focus on higher values of precision or recall, penalizing the score if any of these two is significantly lower. Among our models, the naïve-Bayes approach recorded the highest F1-score of close to 55%, while the decision tree model produced the lowest F1 score despite having the highest sensitivity value.

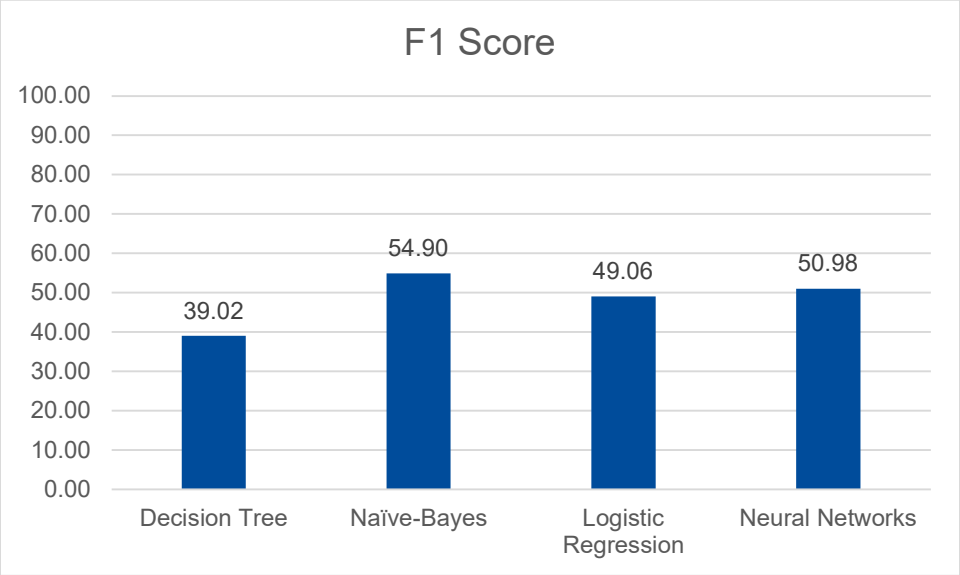

Figure 12: F1 score for the models examined.

Conclusion

We examined the breast cancer data set provided at UCI's machine learning repository. This data set consists of 286 records of patients who either suffered or did not suffer a recurrence of breast cancer. It contained 10 records, 9 of which are the predictors and 1 class or label variable. The predictors include features such as *age*, *menopause*, *tumorSize*, *invNodes*, *nodeCaps*, *degMalig*, *breast*, *breastQuad* and *irradiat*, all of which characterize breast cancer patients.

After conducting an exploratory analysis on the data set, identifying and imputing missing values, we proceeded to create models to predict the class label, i.e. establish using the features patients who are likely to suffer or not suffer a recurrence of breast cancer. Four different classification models were used in the analysis: decision tree, naive-Bayes, logistic regression, and neural network. Though the results of the decision tree model were not particularly excellent, we were able establish some of the highest risk factors leading to recurrence: the number of infected lymph nodes and the histological grade of the tumor. This conclusion is consistent with findings from cancer research. The naive-Bayes model probably provided the best output, with low false negatives, high accuracy, and high specificity. The output of the logistic regression was also good. The stepwise regression conducted prior to fitting the model identified the same variables earlier spotted by the decision tree model. Fitting the logistic regression model with only these three variables produced better output than fitting with all the features. Finally, the neural network we trained probably produced the second-best output. The accuracy was high on both the training and the test sets. Though the models all produced varying values of accuracy, checking if they differ using *ANOVA* revealed that their outputs were not too different from each other. Hence, we failed to reject the null.

We revealed earlier under literature review some of those who had worked on this data set. Tomczak (2013) achieved an accuracy of 73.5% using classRBM, and 71% using naive-Bayes. Chaurasia and Pal (2014) achieved a maximum accuracy of 74.5% using an ensemble method based on RepTree, RBF Network and Simple Logistic, while Murti (2012) obtained a classification accuracy of 75.17% using a Decision Table with naive-Bayes. The metrics obtained from our analysis appear quite consistent with most of these results. We achieved the highest accuracy of 73.5% using a naïve-Bayes model. Like Tomczak (2013), we also found that *degMalig*, the histological grade of the tumor, is one of the major contributors to breast cancer recurrence in women.

As an additional step, we compared the output results from the different models and found that their accuracy values were not statistically different from each other. This means, even though the naive-Bayes model may have provided the highest accuracy, it did not perform significantly better than the other models. We observed that these authors did not take this evaluation step in their analyses, and propose that had they done this, they may have found that the difference in their individual model results are not statistically significant.

Though there are some significantly good results achieved with the different models, there is little doubt a much better result could be achieved if we had more data. With more data we could reduce the heavy dependence of the test results on the randomness inherent in model fitting while also ensuring more robust output. In our analysis, we attempted to minimize the effect of this randomness using cross-validation, and the outputs of the different cross-validation results were consistent with the results obtained with the test data.

Another recommendation would be to perhaps have more features available for this kind of data set. More features would help with understanding the different records better and could in fact lead to better model outputs.

References

[1] Laney, D. (2011, February 6). 3D data management: Controlling data volume, variety and velocity. Retrieved May 28, 2020, from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[2] Lisboa, P., Vellido, A., Tagliaferri R., & Napolitano F. (2010, Februrary 14-18). Data Mining in Cancer Research. Retrieved May 28, 2020, from https://www.researchgate.net/publication/220438411_Data_Mining_in_Cancer_Research

[3] Breast Cancer Research Journal (2020). BCR's 20th Anniversary. Retrieved May 28, 2020, from https://breast-cancer-research.biomedcentral.com/about/breast-cancer-research--celebrating-20-years

[4] World Health Organization. (2012). Estimated incidence, mortality and prevalance of cancer worldwide in 2020 (dashboard). Retrieved from https://globocan.iarc.fr

[5] Moody, S. E., Perez D. ,Pan T. C. ,Sarkisian C. J.,Portocarrero C. P.,Sterner C. J., Notorfrancesco K. L.,Cardiff R. D., and Chodosh L. A.(2005). The transcriptional repressor snail promotes mammary tumor recurrence. Cancer Cell 8, 3 (2005), 197–209.

[6] Lafourcade, A., His, M., Baglietto, L., Boutron-Ruault, M. C., Dossus, L., & Rondeau, V. (2018). s associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort. BMC cancer, 18(1), 171.

[7] Susan G. Komen For the Cure. (2018). Breast Cancer Recurrence. Retrieved from https://ww5.komen.org/BreastCancer/ReturnofCancerafterTreatment.html

[8] Zwitter M. and Soklic M. (1986). Breast cancer data. Retrieved May 25, 2020, from https://archive.ics.uci.edu/ml/datasets/breast+cancer

[9] Unknown. (2018). Predictors for Breast Cancer Recurrence. Retrieved May 28, 2020, from https://www.causeweb.org/usproc/sites/default/files/usclap/2018-1/Predictors_for_Breast_Cancer_Recurrence.pdf

[10] American Cancer Society (2020). How Common is Breast Cancer?. Retrieved from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[11] Delen D., Walker G., and Kadam A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine 34: 113-127.

[12] Lundin M., Lundin J., Burke H. B., Toikkanen S., Pylkkanen L., and Joensuu H. (1999). Artificial neural networks applied to survival prediction in breast cancer. Oncology 57: 281-286

[13] Ahmad L. G., Eshlaghy A. T., Poorebrahimi A., Ebrahimi M., and Razavi A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. Journal of Health and Medical Informatics 4: 2.

[14] Advisory Board (2020). Google's AI beats doctors at detecting breast cancer. (Except when it doesn't.). Retrieved June 06, 2020, from https://www.advisory.com/daily-briefing/2020/01/06/google-ai

[15] Abreu P. H., Santos M. S., Abreu M. H., Andrade B., and Silva D.C. (2016). Predicting Breast Cancer Recurrence Using Machine Learning Techniques: A Systematic Review. ACM Computing Surveys, Vol 49, No. 3 52:1-39

[16] Murti M. S. (2012). Using rule-based classifiers for the predictive analysis of breast cancer recurrence. Journal of Information Engineering and Applications 2, 2 (2012), 12-19.

[17] Belciug S., Gorunescu F., Salem A. B., and Gorunescu M. (2010). Clustering-based approach for detecting breast cancer recurrence. In Proceedings of the International Conference on Intelligent Systems Design and Applications (ISDA). 533–538

[18] Chaurasia V., and Pal S. (2014). Data mining techniques: To predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing 3, 1 (2014), 10–22.