Prediction of Breast Cancer Recurrence using Classifications Methods

Ade Adeoye

Table of Contents

[Abstract 3](#_Toc43473532)

[Introduction 4](#_Toc43473533)

[Literature Review 5](#_Toc43473534)

[Data set 9](#_Toc43473535)

[Approach 13](#_Toc43473536)

[Step 1 13](#_Toc43473537)

[Step 2 17](#_Toc43473538)

[Step 3 19](#_Toc43473539)

[Results 19](#_Toc43473540)

[Conclusion 24](#_Toc43473541)

[References 25](#_Toc43473542)

# Abstract

Breast cancer is one of the leading causes of cancer-related deaths in the world. Its incidence is however quite biased. According to the Centers for Disease Control and Prevention (CDC), most cases of breast cancer are found in women, with estimates putting the likelihood at about a hundred times more likely than in males. For cancer-related diseases, early detection is crucial. If discovered early, a patient is highly likely to survive the disease, even if there may be a few cases 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](https://archive.ics.uci.edu/ml/datasets/breast+cancer) 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, after initially carrying out exploratory data analysis. The methods used include decision tree, naïve-Bayes, logistic regression, and artificial neural networks. As a 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 settle on the best means of carrying out the classification task for this data set.

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

# Introduction

Breast cancer is one of the most common cases of cancer in women. According to the World Health Organization, it 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 of concerning 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. Using the attributes provided in the data set, we attempt to predict breast cancer recurrence with the aid of 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 also obtain a probability value associated with recurrence and some of the attributes which may weigh heavily on this probability. In all, we employ classification techniques such as logistic regression, naive-Bayes and decision tree. We also examined some deep learning classification techniques, and as a final related statistical approach to analysing the available data, compared the metrics of these different classification techniques in a bid to establish which did best.

.

# Literature Review

In recent times, data mining has turned out incredibly useful in revealing latent information otherwise difficult to obtain from large swathes of data. With the increasing generation of data in volume, velocity and variety (Laney, 2001), it is becoming highly likely that the huge streams of data will never stop. Taking advantage of the depth and breadth of available data about an event, transaction or incident, becomes more important regardless of domain or subject. This emerging field of data exploitation is called data mining or data munging. It takes an approach of searching for relevant patterns and regularities (or irregularities) in large masses of data in order to elicit useful information that may inform decision-making. It appears to be domain-agnostic and 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 repository domiciled with most of 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 stores 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.

According to Lisboa, Vellido, Tagliaferri and Napolitano (2010), the last two decades prior to 2010 saw “an increasing throughput of data from microarray screening, spectral imaging and longitudinal studies [that has turned] the understanding of cancer pathology into as much a data-based as a biologically and clinically driven science, with potential to impact more strongly on evidence-based decision support moving towards personalized medicine.” In unravelling the information provided by collected data, machine learning techniques such as clustering appear to be of good significance. It can be used to mine databases and obtain similarities in multi-dimensional space for a collection of patient-associated attributes such as genes, mode of action drugs, patient age, or disease sub-types. Principal component analysis, or any other feature selection method, is also particularly prevalent, since it is in the nature of the medical sciences to churn out a lot of data. Knowing which to prioritize in any analysis or establishing which are the most discriminative attributes is typically an appropriate task.

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. 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. A search conducted on the online database of the mentioned journal, for example, turned out 10354 results for search item “machine learning”, 2664 results for “artificial neural networks”, and 28964 results for ‘principal component analysis”. 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. 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 radiologists (Advisory.com, 2020). 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%”.

As the most frequent incidence of cancer among ages above 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. Its mortality rate is most associated with metastization and recurrence, 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. 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. The volume of research into breast cancer prediction using data mining means that there is a sizeable amount of work that has been conducted using publicly available data sets. While some used proprietary data, most have used the Wisconsin (Diagnostic) Breast Cancer Data 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 the 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 include 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 have 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 linear 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.

|  |  |
| --- | --- |
| 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* |
| *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 1: The data set attributes and their descriptions

The 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 generates random synthetic data to fill this gap. Packages available in R which treat this unusual case of class imbalance include *SMOTE* and *ROSE*.

When *class* is split across *age*, we see how imbalance the data set is. The major part of this proportion that is imbalance lies between ages of 40 and 69.

|  |  |  |
| --- | --- | --- |
| Age | no-recurrence-events | recurrence-events |
| *20-29* | *0.35* | *0.00* |
| *30-39* | *7.34* | *5.24* |
| *40-49* | *22.03* | *9.44* |
| *50-59* | *24.83* | *8.74* |
| *60-69* | *13.99* | *5.94* |
| *70-79* | *1.75* | *0.35* |

Table 2: Distribution of "no-recurrence-events"

and "recurrence-events" by age.

A screenshot of a cell phone

Description automatically generated

Figure 1: A bar plot of the different age frequencies

show that the data set is imbalanced

When *age* is split across both *class* and *tumorSize*, one of the defining variables in the data set, we see again the imbalance. 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 30 and 39mm. This is where the bulk of the data lies. For some other *tumorSize* facets such as 45-49mm and 50-54mm, very little patients are represented.

A screenshot of a cell phone

Description automatically generated

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”. With a mosaic plot, we can look at some of these relationships. In this plot between *tumorSize* and *invNodes*, 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.

A screenshot of a cell phone

Description automatically generated

Figure 3: A mosaic plot shows the relationship between tumorSize and invNodes.

There are progressively smaller blocks for increasing values of both.

The data set is not without missing values. There are 9 missing values in all. 8 under *nodeCaps* and 1 under *breast*. Missing values may or may not constitute a problem in the analysis of a data set. This will depend on the nature of the missing values, whether they are missing at random or intentionally omitted. Evidently, the attributes containing missing values in this data set suggest that these missing values were not intentionally omitted, and since the data set is a rather small one, it should be desirable to attempt imputing these missing values. In addition, there are other attributes available for each missing value. This suggests that there is sufficient information available for imputing. In solving this while exploring the data set, we tried different techniques which include MCA and KNN, and subsequently imputed the missing values before proceeding with additional analysis.

A rather unusual property of this data set is that all the attributes are categorical. This presents unique challenges in handling since we are unable to use many of the available numerical approaches to data analysis. It also means we must take different unique approaches, such as one-hot encoding, in the exploitation of these categorical variables. The nature of this data set is unlike the more widely available Wisconsin Breast Cancer data set (Street, 1995), equally available for download at UCI Irvine Machine Learning Repository, but more suited to cases of breast cancer diagnosis than to those of breast cancer recurrence. Though our data set saw a lot of usage in the succeeding years after its release in 1986, this usage has faded somewhat in recent years compared to the more widely available Wisconsin (Diagnostic) data set. In a survey of multiple resources on the web which conducted analysis on breast cancer related information, we found only about 17% representing information from this data set compared to 83% which analysed the Wisconsin alternative. In the paper by Abreu et al mentioned earlier, of their 17 reviewed works, four used the Wisconsin data set, compared to 3 which used this data set.

# Approach

The development approach is divided into three basic steps which include data preparation, exploratory data analysis and data balancing, and model fitting and selection. Each step requires an exhaustive approach and involves learning every detail about the data set. The steps are also not conducted in isolation, as there exists a feedback loop which opens the model selection back to exploratory data analysis.

## Step 1

In analysing this data set, we follow the process outlined in Figure 2. Firstly, we prepare the data, reorganize it, treat missing values and change data types, if necessary. While imputing missing values, we explore three different approaches including using nearest neighbours, clustering using multiple correspondence analysis (MCA), or simply replacing the missing items with the modal values. Since there are many more attributes available for each missing value, using the nearest neighbour appear just as good as using the modal value.

The data set has all categorical values. This means we cannot use the many numerical approaches typical of most machine learning procedures. However, this does not detract from deriving insights from the data. Preparing the data set involves exploring the significance and distribution of each of the attributes. Using the *table* function in R, we can 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.

There is a total of 9 cases with missing values under the *nodeCaps* and *breastQuad* attributes. Prior to conducting our classification tasks, we examine different approaches to imputing these missing values. Imputing involves identifying and replacing missing values with closest alternatives using statistical approaches. In this work, we take three different approaches: imputing modal values, k-nearest neighbour imputation and MCA. The modal values are the most frequently occurring values under each attribute. For *nodeCaps*, this is “no”; and for *breastQuad*, it is “left\_low”. The k-NN computation uses a distance measure developed for categorical or mixed data called Gower’s distance, given in the equation below.



where is a weight for the th variable. takes the value of 1 when both and are known; otherqise, it takes zero. is the square of the distance between the th value of the two observations and . is given as:

The *kNN* function available in the *VIM* package allows imputing using Gower’s distance. After imputing using this function, the results obtained were observed to be identical to the modal values.

The third approach of data imputation is achieved by examining the data set’s principal components. The attributes in this data set are all categorical. This means we cannot use R's base PCA package. 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 this type of data". 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. The first 8 dimensions retrieve about 39% of the data’s total inertia. Admittedly the number of variables is very small, and we do not necessarily achieve any gain by reducing the number of attributes, nevertheless using PCA provides us with a convenient clustering of the different attributes along the principal components. With this clustering, we gain better insight into the distribution inherent in the data set. When we plot the first two principal components, the data distribution in terms of age shows a well-distributed, 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. This favourable distribution gives us additional insight into the data.

A screenshot of a map

Description automatically generated

Figure 4: 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 at 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 that was achieved with both the kNN method and the modal values.

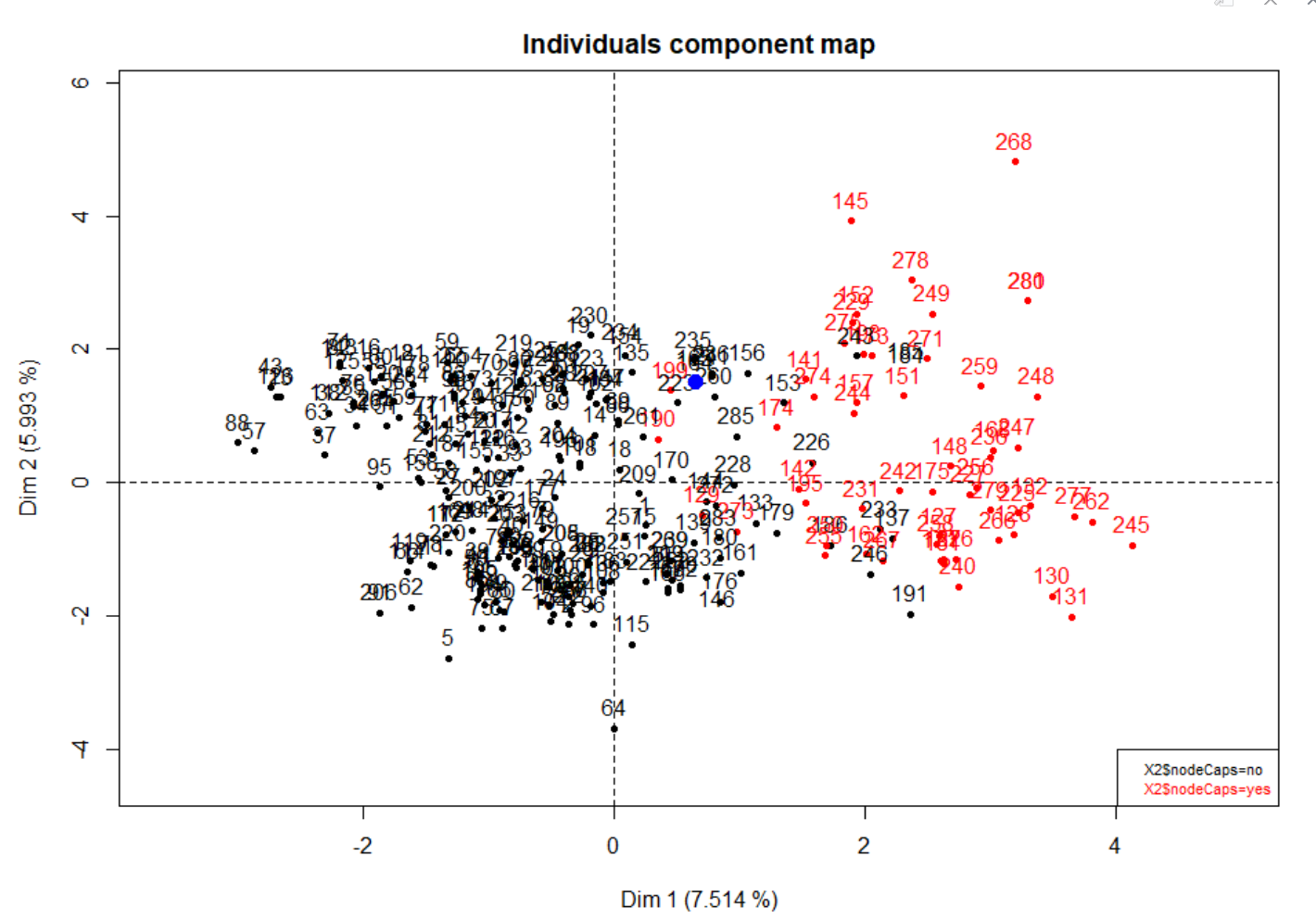


Figure 5: Cluster distribution of nodeCaps. The blue dot is a missing value

in the data set and appears to be grouped with the "yes" cluster.

The data types for this data set came as strings. By converting into factors, we can use the attributes to delineate relationships with the major *class* attribute.

## Step 2

We conduct exploratory data analysis by visualizing the data and exploring the strengths of its different features using MCA. The data is imbalanced as there are more cases of non-recurrence in the data sets than there are cases of recurrence of breast cancer. Balancing the training data set is an important precursor to creating a 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 examine different techniques of random sampling and observe which is much better for the data set. Some of approaches we take include undersampling, oversampling or synthetically generating random fillers. We find that many of these different approaches produce similar results.

Breast

Cancer

Data

Data

Preparation

Reorganization

Cancer

Data

Missing values

Data types

Cancer

Data

Exploratory

Data

Analysis

Data

Balancing

(Treat Imbalance)

Data Viz.

Cancer

Data

PCA

Cancer

Data

* Oversampling
* ROSE
* SMOTE

Model

Fitting

* Decision Trees
* Naïve-Bayes
* Logistic Regression
* ANN

Model

Selection

Cross-validation

Reports

Figure 6: An outline of the workflow for analysing

the Breast Cancer 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. We find that, for this data set, the differences in metrics did not vary much when different sampling techniques are used to balance the data set. However, we did find that using fewer attributes has more impact on the resultant metrics.

With balancing, we have more test data to validate the model, using both basic model validation with test set and cross-validation.

## Step 3

In fitting the model, we use firstly a tree-based model by constructing a decision tree. Next, we fit a naive-Bayes model and then a logistic regression model. Finally, we use a more modern model fitting technique by constructing an artificial neural network. 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. This involves dividing the model into ten different chunks and using, in repeated cycles, all 9 chunks for training and the remainder for testing. Finally, we select the desired model and put together a summary report of the metrics obtained while fitting each model. The step of model selection remains open to revision which takes us back to the data exploration stages.

# Results

Our initial assessment of the data set showed that it was highly imbalanced. We needed to treat this imbalance appropriately and used `ROSE` 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 that we would use oversampling more, since a part of the class, the `recurrence-events` class, was already small enough. In using oversampling and undersampling techniques in modeling building, we also leveraged the availability of a sampling option while training which is part of the *caret* package.

We proceeded to build four separate models so that we can compare their performances on the breast cancer data set. In qualifying a model, we look at the one which minimizes the false negatives. The decision tree model initially used the oversampled data set and then the doubly-sampled data set. We used a seed value of 80 throughout this analysis. The oversampled data set produced a sensitivity of 73% and a specificity of 44%, while delivery an accuracy of 52%. There were only 7 false negatives in all. The doubly-sampled data set, on the other hand, produced 15 false negatives. It delivered a specificity of 82% and a sensitivity of 42% with an overall accuracy of 70% on the test set. This seems okay, but without conducting a cross-validation check it is hard to accept the output produced by this model.

Cross-validation was conducted using the *caret* package and the output seemed promising. With only 11 false negatives and a specificity of 72%, the model seemed quite robust. Its overall accuracy was 67%, which is no better than random selection. Random selection on the imbalanced test set will produce an accuracy of 70%. Its sensitivity was also rather low at a value of 46%. While we can say that the model has not done excellently, we can comfortably settle on the fact that the model definitively indicated that the number of infected lymph nodes and the histological grade of the tumor are rather strong determinants in the recurrence of breast cancer in women. These are the two top features identified by the tree model. This conclusion is consistent with results obtainable from cancer research.

The naive-Bayes classifier constructed using the training data set also appeared quite good. It delivered a specificity of 82% on the test set 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 was 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.

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 these sampling methods including oversampling, undersampling, ROSE and SMOTE. We found that, for this data set, sampling up seemed to have a greater tendency to produce better accuracy values. Sampling down, on the other hand, was greater disposed to producing lower accuracy models.

A screenshot of a video game

Description automatically generated

Figure 7: Output results from using different sampling techniques to balance the data set. "Up" sampling seemed to produce better accuracy values, while "down" sampling produced lesser accuracy values for this data set.

Another classification considered was logistic regression. Prior to fitting a model, we conducted step-wise 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% 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. Unexpectedly, the neural network did considerably well on the training set producing an accuracy of 76%. However, it did quite better than some of the other models on the test set. We recorded 13 false negatives, an accuracy of 71%, 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.

A picture containing screenshot

Description automatically generated

Figure 8: A chart of variable importance for the variables used by the neural network.

Featuring engineering often provides additional opportunity to explore redundant variables. Sometimes, the effort yields positive results. In our case, we conducted used one-hot encoding to explore additional features suggested as important by the neural network results. These features topped the list of variable importance in the plot of the dummy variables fed into the neural network. Using these features with a logistic regression model, 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.

Asa final check, the accuracy results obtained from each model were compared against each other in an *ANOVA* test 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.

A close up of a map

Description automatically generated

Figure 9: 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 from random sampling.

# 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. Since we are seeking to minimize false negatives, this is an encouraging result. The output of the logistic regression was also good. The step-wise 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.

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 to fit the models we could remove their heavy susceptibility to the randomness inherent in model fitting while also ensuring more robust output. To minimize the effects of this randomness, we applied cross-validation at each stage. The outputs of the different cross-validation results were consistent with the results obtained with the test data. This showed that the models were resilient enough. 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 would 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/

[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.com (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.