# Title: Predicting whether a patient would survive longer than 5 years after breast cancer surgery

Authors: Wesley Paul Nderi – s3635870
Nincy Mathew Pookkoden - s3650735

This report has been compiled in the fulfilment of a Master of Analytics at the Royal Melbourne University of Technology (RMIT University)

Contact Details:
s3635870@student.rmit.edu.au
s3650735@student.rmit.edu.au

Date of report: 5th May 2018

## Executive summary

The aim of this report was to investigate the survival rate of a patient within the first five years after undergoing breast cancer surgery based on the Haberman's data set. A study was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital to collect this data. This was sourced from the UCI Machine Learning Repository. Overall, the preliminary results indicate that a patient has a high likelihood of survival after surgery and that breast cancer has a relationship with the age of the patient. It is recommended that early stage detection of breast remains the single most important identification method of cancer.

## Introduction

Today, in Australia breast cancer is the regarded as the leading cause of cancer deaths in young women. An analysis conducted by the Australian Bureau of Statistics(A.B.S 2016) highlights that in 2016, breast cancer was ranked as one of the five leading types of cancers causing death in young Australians. Further, that in the same year the 30-39 age group had the highest recorded number of breast cancer deaths which could be attributed to a difference in characteristics and risk factors. For example, a higher proportion of younger women are found to have larger tumour size cancers associated with lower survival rates (AIHW 2012). In addition, even in cases where the cancers are small, the five year survival rate of younger women is lower than that for women aged over 40 years (ABS 2016). The Australian Institute of Health and Welfare (AIHW 2012) suggests that 37 Australian women are diagnosed with this cancer each day, with about 7 succumbing to the disease daily. Cancer is a global phenomenon with the World Health Organization (WHO) estimating that about 8.8 million people lose their lives to cancer, with the majority being from low-and- middle income countries.

The WHO advocates for the early detection of cancers and suggests that this allows for a more effective treatment plan, which improves the chances of survival, reduces morbidity and also significantly reduces the costs associated with treatment. In addition, screening programmes can be effective in identifying specific cancer types, for example, using mammography screening to identify breast cancer(WHO 2018). In this report we present the use of supervised machine learning techniques to classify the survival status of patients after undergoing breast cancer surgery based on the *Haberman's dataset* available from the UCI Machine Learning repository.

The rest of this report is organized as follows. The next section reviews the methodology used to conduct this classification task. Experimental results are presented in Section 3. Discussion and Conclusion are given in the last sections.

## Methodology

This report used supervised machine learning methods techniques as advocated for by Yongli Chen (2018) during his lectures at RMIT. In this paper, we have embarked on a classification task and investigated two modelling techniques; the *k Nearest Neighbour* and the *Decision Tree*. This classification task involves identifying what category a new observation belongs. Specifically, this report used these modelling techniques to classify whether a patient would survive longer than five years based on the *Haberman's Survival Data Set*.

The *k Nearest Neighbour* technique based on the *k-NN* algorithm. It is described as a conventional non parametric classifier(Cover & Hart 1967) and is usually used as the baseline classifier in many classification problems involving patterns (Jain et al. 2000). It uses a distance measure of similarity between instances- usually the *Euclidean* distance. It uses this proximity to predict a given query. Other distance measures have been mentioned in literature such as the *Minkowski* or *taxi-cab* distance(Batchelor 1978), cosine similarity measure(Manning et al. 2008) and chi square(Michalski et al. 1981). We modify the algorithm and use a *distance weighted k nearest neighbour* approach by weighting points by the inverse of their distance so that closer neighbours of a query point will have a greater influence than neighbors further away. For the purposes of this report, we do not address the imbalance in the target variable and focus instead on the use of these techniques in the classification problem at hand. The parameters used in this report after parameter tuning are that **n_neighbours** are set to 35 where the accuracy seems to plateau, **weights** set to *'distance'*, **metric** set to *'minkowski'* and **p** set to 1.

The *Decision Tree* technique is based on the Iterative Dichotomizer 3(ID3) algorithm that attempts to create the shallowest decision tree that is consistent with the data

given. The algorithm builds a recursive, depth- first manner, beginning at the root node and working down to the leaf nodes. The algorithm begins by choosing the best descriptive feature to the test, based on the information gain. The design of this algorithm assumes that a correct decision tree for a domain will classify instances from that domain in the same proportion as the target level occurs in that domain.In this algorithm parent node represents a single node and splits on that variable.The leaf nodes of the tree contains the output variable which is used to make the predictions . In this report, the split is done based on nodes and age, further it evaluates all possible split nodes till it reaches the terminal node.The decision tree classifier parameters used in this report are **max_depth** set to **5**, **criterion** set to **Gini** and **minimum sample leaf** is set to **2**. The Gini index used provides an indication how pure the nodes are and also indicates how good the split is.

We selected these two classification techniques to find the most suitable one for predicting cancer survivability rate after surgery. The focus shall be on three attributes- the age of participants, year the case was discovered and the number of positive axillary nodes detected. Using a backward elimination strategy to find the most informative feature in order to avoid the curse of dimensionality. We find that the most informative feature in predicting whether a patient survives longer than 5 years is the number of nodes detected as it has the highest accuracy.

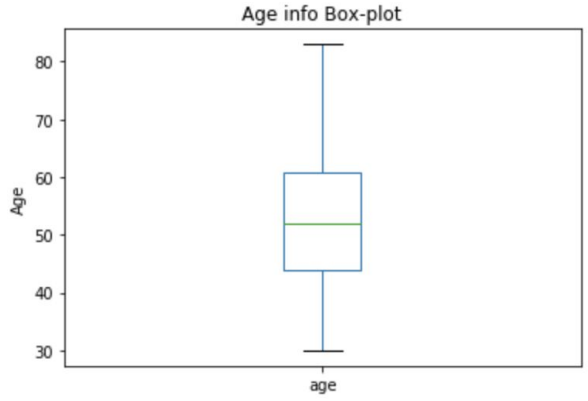
## Results

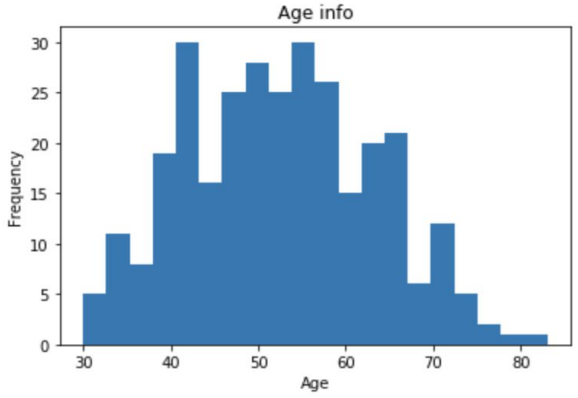
The *Haberman's Survival Data Set* contains 4 features, namely the age of participants, the patient's year of operation, the number of positive axillary nodes detected and the survival status of the patient. The survival status is the target variable while the three other features are the descriptive features. The data has a total of 306 rows.

## Univariate analysis

We shall proceed to explore each of these attributes below in order to get a better understanding of the data.

### Age attribute



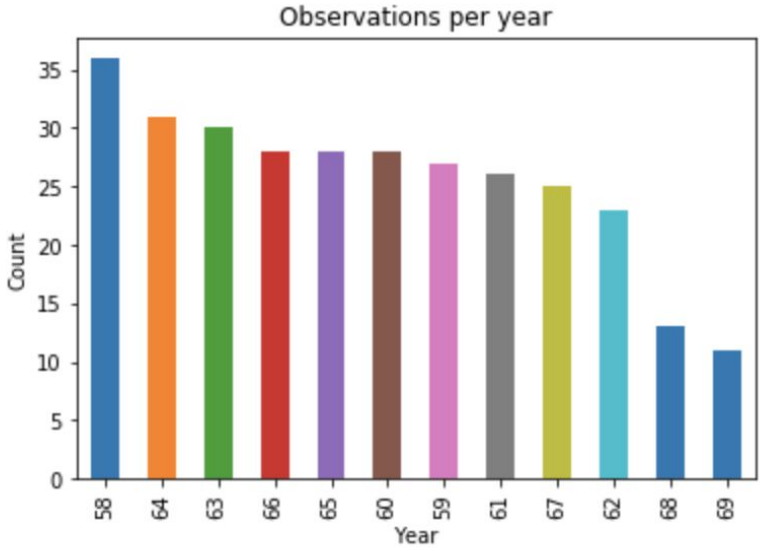




This feature is numerical in nature, with the minimum age being 30 and the maximum age being 83. Using a histogram to visualize the age distribution, we see a relatively symmetrical distribution of the ages. The box plot indicates that there are no outliers present in this attribute and as a result, no impossible values.

For a more meaningful interpretation in the context of the data, we group the ages in ranges of 5 as shown below and create another variable **age_group**. The age group that records the highest count is the 50-55 years group. The 40-45 years and 50-55 years have relatively high counts as well.

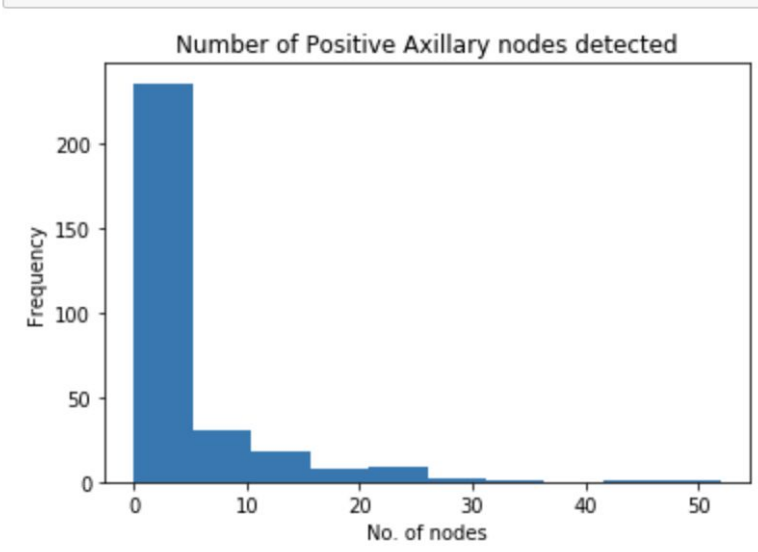
```
30 - 35     14
35 - 40     26
40 - 45     40
45 - 50     44
50 - 55     56
55 - 60     43
60 - 65     35
65 - 70     27
70 - 75     16
75 - 80      4
80 - 85      1
Name: age_group, dtype: int64
```

**Year attribute**



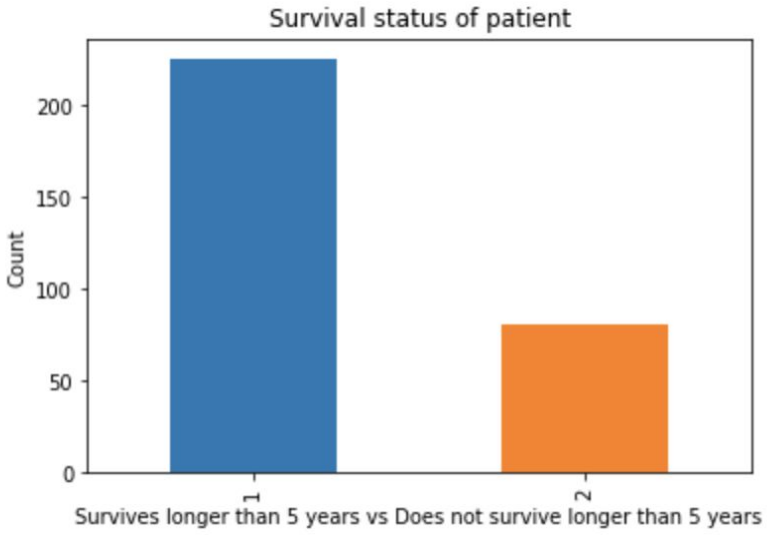

This feature is numerical in nature and contains the number of observations between the year 1958 and 1970, although it does not seem to be inclusive of 1970. There is a clear decline in the counts from the beginning of the study to the end of the study and more so an abrupt decline in 1968.

**Nodes attribute**



This feature is numerical in nature and it is clear that there is a skew in the distribution of the number of positive axillary nodes detected.  The most common number of nodes detected is zero.

**Status attribute**

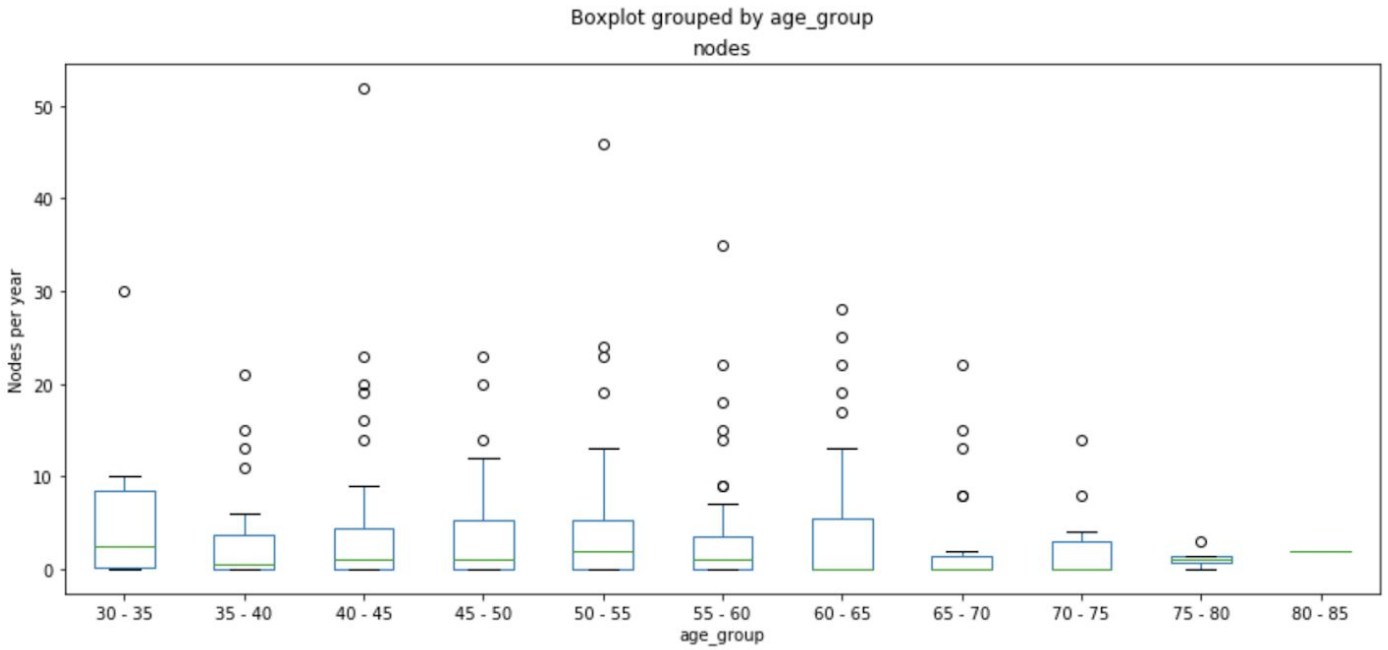



This feature is the target variable and is binary in nature. It contains coded values. The patient's status can either be **1** to mean that they survived longer than five years or **2** to mean that the patient died within the five year window. It is clear that there is an imbalance in the data as majority of the patients survive longer than five years with only a small minority not surviving longer than the five years.

## Multivariate analysis

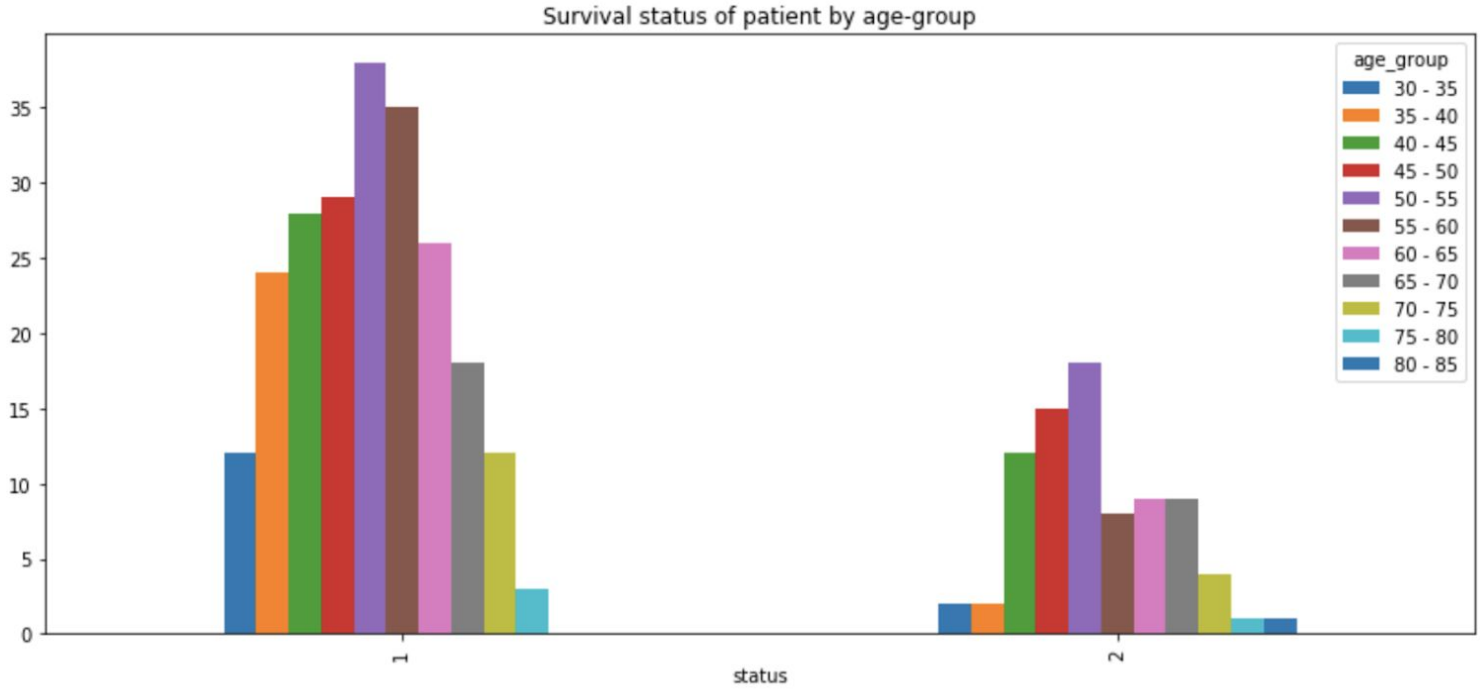
We shall now explore the existence of relationships between the attributes using multivariate visualization techniques as shown below.

### Age groups vs no. of nodes detected




Exploring the relationship between age group and number of nodes, we observe that the median for all the categories tends to be close to zero. This complements our observations in the *Nodes* attribute. However, there are significant outliers that were observed in the course of the study, for example, over 50 nodes were detected in a patient within the 40-45 age group. The outliers seem to be more common as the age of the patient increases. The majority of patients seem to be aged between 30 and 65, as the observations seem to decline slowly at that point. Irrespective of outliers the outliers observed, the age group 30-35 has the highest number of nodes observed.

**Survival status by age- group**

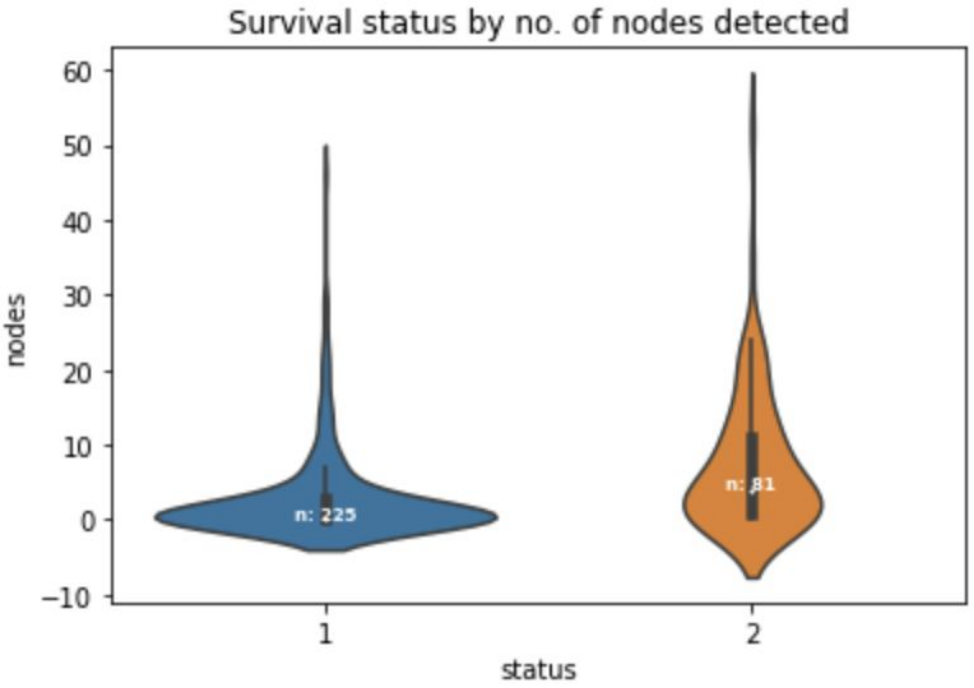



| age_group | 30 - 35 | 35 - 40 | 40 - 45 | 45 - 50 | 50 - 55 | 55 - 60 | 60 - 65 | 65 - 70 | 70 - 75 | 75 - 80 | 80 - 85 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| status | | | | | | | | | | | |
| 1 | 12 | 24 | 28 | 29 | 38 | 35 | 26 | 18 | 12 | 3 | 0 |
| 2 | 2 | 2 | 12 | 15 | 18 | 8 | 9 | 9 | 4 | 1 | 1 |

| | | |
|---|---|---|
| 1 | 30 - 35 | 12 |
| | 35 - 40 | 24 |
| | 40 - 45 | 28 |
| | 45 - 50 | 29 |
| | 50 - 55 | 38 |
| | 55 - 60 | 35 |
| | 60 - 65 | 26 |
| | 65 - 70 | 18 |
| | 70 - 75 | 12 |
| | 75 - 80 | 3 |
| 2 | 30 - 35 | 2 |
| | 35 - 40 | 2 |
| | 40 - 45 | 12 |
| | 45 - 50 | 15 |
| | 50 - 55 | 18 |
| | 55 - 60 | 8 |
| | 60 - 65 | 9 |
| | 65 - 70 | 9 |
| | 70 - 75 | 4 |
| | 75 - 80 | 1 |
| | 80 - 85 | 1 |

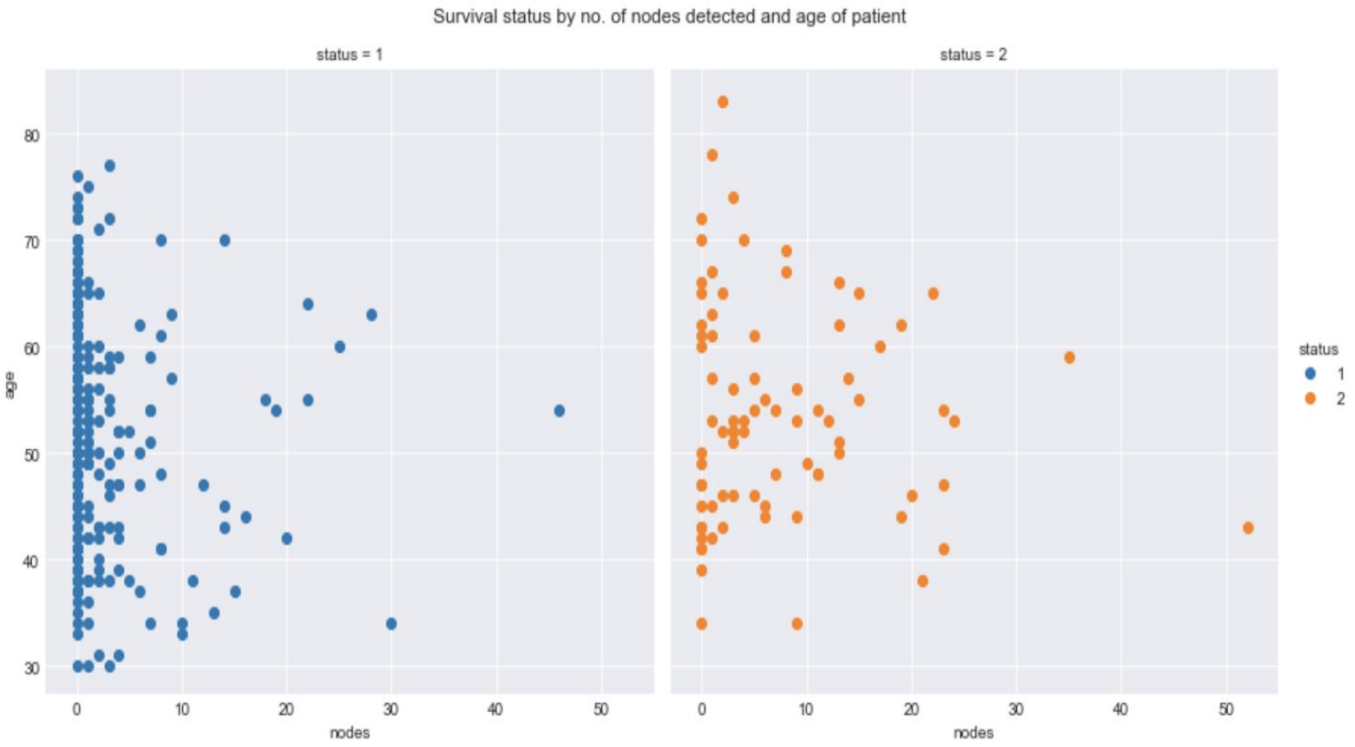
The distributions of age groups based on whether they survived longer than 5 years (**status=1**) or did not survive longer (**status=2**) are completely different. Status 1 appears to have a symmetric normal distribution while Status 2 appears to have a skewed right distribution(which is essentially non normal). The imbalance in the target variable(*status)* is also clear here as depicted by the difference in the height of the bars.

**Survival status by no. of nodes detected**




From the above violin plot, patients less than 5 axillary nodes have a greater survival rate which is clearly visible as the density is more between 0 to 8 nodes. The higher the number of nodes, the greater the chance of not surviving after surgery.

**Survival status by no. of nodes detected and age of patient**




The above scatter plot is faceted based on the survival status and illustrates a relationship between age and nodes in that respect. Both status 1 and 2 seem to have similar behaviour although status 2 exhibits patients with more nodes detected. Although the mode is zero axillary nodes, we can see that it is not as dense in status 2 as it is in status 1. It is possible to infer that the higher the number of nodes observed, the less the chance of survival after surgery.

## Model Outcomes

In this section we shall highlight the steps involved in tuning the parameters as well as the arising interpretations.

**a)** ***K Nearest neighbour***

We shall focus on finding values for the following parameters in this model: **n_neighbors(k)** , **weights, metric** and **p.**

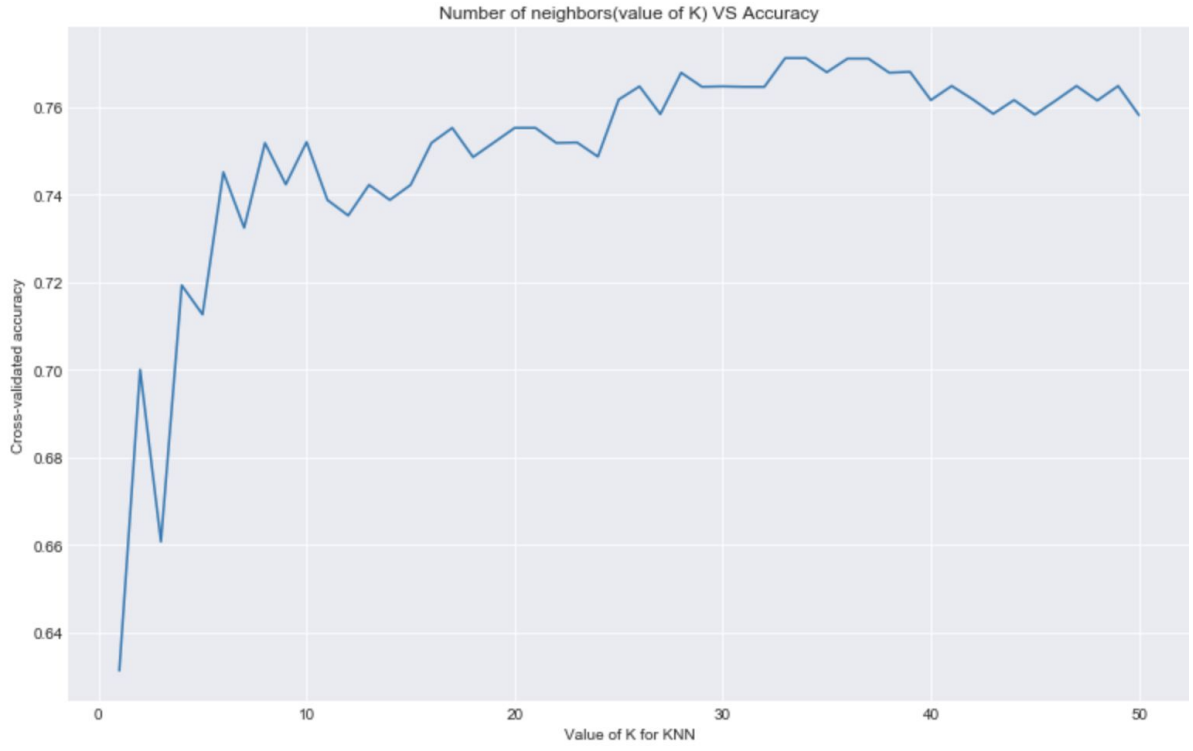
Starting with arbitrary values as shown below and all of the features(age, year and nodes) we build the classifier and use a 10 fold cross validation technique to validate the model. The accuracy measure is about **0.69** with a variance of (+/- 0.20). The confusion matrix is shown below.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.78 | 0.80 | 0.79 | 225 |
| 2 | 0.40 | 0.36 | 0.38 | 81 |
| avg / total | 0.68 | 0.69 | 0.68 | 306 |

From this we can infer that the average precision is 0.68(68%), average recall is 0.69(69%) and the F1 score is 0.68. We observe the effects of the class imbalance as the classifier captures the positive majority(status 1 = patient survives longer than 5 years after surgery) but severely fails to capture the minority class(status 2).

**Optimum value of k**

We define a range of k to be between 1-50 and observe the highest value of accuracy that can be obtained as shown in the graph below. The highest value of k is about 32 with an observed accuracy of **0.77**.




**Optimum value of p**

Using this value of k, we embark on finding the optimum value of p. For the purposes of this report, we shall restrict ourselves to choosing between the Euclidean distance(p=2) and the Manhattan distance(p=1).

With p = 2 and the value of k as 32, we observe the confusion matrix as shown below.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.77 | 0.92 | 0.84 | 225 |
| 2 | 0.51 | 0.23 | 0.32 | 81 |
| avg / total | 0.70 | 0.74 | 0.70 | 306 |

Value of p = 2

The accuracy measure is **0.75** with a variance of (+/- 0.08). From the confusion matrix we can infer that the average precision is 0.70(70%), average recall is 0.74(74%) and the F1 score is 0.70.

We shall now compare these results with when p = 1 and the value of k remains as 32. The results are shown below.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.77 | 0.92 | 0.84 | 225 |
| 2 | 0.53 | 0.25 | 0.34 | 81 |
| avg / total | 0.71 | 0.74 | 0.71 | 306 |

---

The accuracy measure is **0.73** with a variance of (+/- 0.07). The precision goes up by about 1% and while the average recall remains the same, we can see that with value of p as 1, there is a higher score of 0.25 captured for status 2. Similarly, the average F1 score also goes up from 0.70 to 0.71 and there is more of status 2 captured.

As discussed before, we use a *distance weighted k nearest neighbour* approach by setting **weights** to *distance.*

**Finding the most informative attribute**

Using an elimination technique and comparing the different accuracy measures, we aim to find what feature or combination thereof gives us the most information about the target.

| **Attribute** | **Age, year, nodes** | **Age, nodes** | **Year, nodes** | **Age** | **Year** | **Nodes** |
|---|---|---|---|---|---|---|
| Accuracy score | 0.70 | 0.68 | 0.72 | 0.55 | 0.74 | 0.74 |
| Variance | +/- 0.11 | +/- 0.22 | +/- 0.16 | +/- 0.39 | +/- 0.01 | +/- 0.08 |

From the above, we see that the individual attributes, with the exception of age, have the higher scores. Inferring from domain knowledge in the medical profession, we shall pick nodes as our descriptive feature.
Therefore, the confusion matrix we shall use to evaluate our model is as shown below.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.77 | 0.92 | 0.84 | 225 |
| 2 | 0.51 | 0.23 | 0.32 | 81 |
| avg / total | 0.70 | 0.74 | 0.70 | 306 |

We observe that the average precision of the *kNN* model is 0.70 with the recall being 0.74 and the F1-score being 0.70. This means that about 70 % of the time, this model correctly identifies patients who would live or not live longer than 5 years after their breast cancer strategy. However, only 74% of the time is the model able to classify these patients. The F1 score is the weighted average or harmonic mean of precision and recall and provides that the model is 0.70. Specifically, we can observe that the model has a high recall of 92% for status 1(survived longer than five years) but only 23% for status 2(did not survive longer than five years).

***b) Decision tree***

The goal of this model is to create a tree that predicts the target variable based on the descriptive features.

The first step was to split the data into the feature set and target set. The "X" set consists of the attributes, age and nodes. The features selected were influenced by the earlier discussion in the *kNN* model as well as domain knowledge. The year the patient had surgery does not seem to be very informative. The "y" set consists of the target value which is the survival status.

The data was then split into training and test data .The parameter **test size** used was **0.2** which means the test set is 20% of the data and the remaining 80% will be the training data set.

The next step was to fit a decision tree algorithm into training data set and this was implemented using function **DecisionTreeClassifier().**

**Parameters used**

**Criterion** set to **Gini-** This measures the quality of the split.

**Minimum_samples_split** set to 2(by default). This means if only 1 sample is left then it cannot split further.

We shall check the confusion matrix and accuracy with the parameter used:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.65 | 0.89 | 0.75 | 37 |
| 2 | 0.64 | 0.28 | 0.39 | 25 |
| avg / total | 0.64 | 0.65 | 0.60 | 62 |

The accuracy measure seen is **0.65.**

As the minimum_sample_fit is set to 2, when it goes deep it will overfit the data.Hence the value of  min_sample_fit is changed so as to decrease the depth of the tree and there is less chance to go out of samples to split. This reduces overfitting.

Hence changing the parameter and checking for accuracy of the model

In order to find the best parameters for the decision tree model, a function known as **GridSearchCV** has been implemented as mentioned below.

A sample_split_range is defined that specifies the range values that will be searched. A parameter grid was created so that it can determine the optimum value of minimum sample split based for each value of assigned range.

A grid is initiated with 10 crossfold validation on a decision tree model with accuracy as the evaluation object along with the parameter grid for each of its value.

From this implementation, the best score,best parameter and best estimators are derived for this model.

```
('Best score is:', 0.7459016393442623)
{'min_samples_split': 2}
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=5,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best')
```

From the above,
**The best score is 0.74** which is the single best score achieved among all parameter.
**The best parameter is min_sample_split=2** , its a dictionary which uses this parameter to generate the best score.
**The best estimator** is the actual model objects that fits the parameter and shows all the parameter that should be used to generate the model

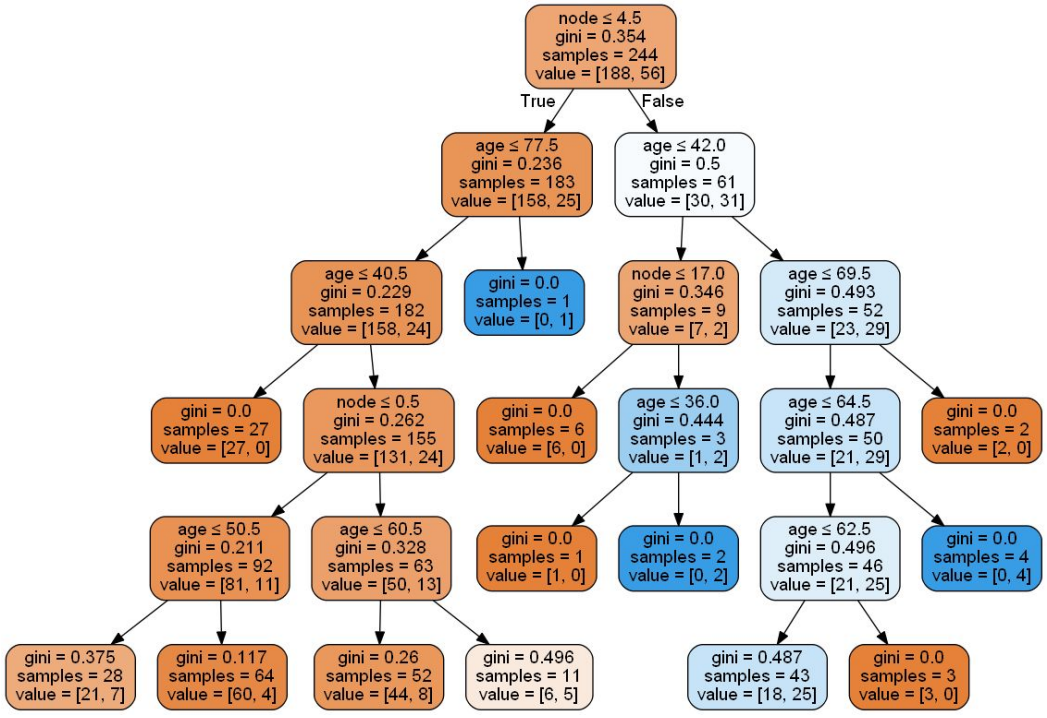
After using the above parameters the following results are seen:

**Accuracy is 0.68.**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.67 | 0.92 | 0.77 | 37 |
| 2 | 0.73 | 0.32 | 0.44 | 25 |
| avg / total | 0.69 | 0.68 | 0.64 | 62 |

We observe that the average precision of the decision tree model is 0.69 with the recall being 0.68 and the F1-score being 0.64. This means that about 64 % of the time, this model correctly identifies patients who would live or not live longer than 5 years after their breast cancer strategy. However, only 68% of the time is the model able to classify these patients. The F1 score is the harmonic mean of precision and recall and provides that the model is 0.68 accurate. Specifically, we can observe that the model has a high recall of 92% for status 1(survived longer than five years) but only 32% for status 2(did not survive longer than five years).

Below is the decision tree plotted with the above parameters and its splitted on node as it is the crucial and most important feature for this modelling.



## Discussion

In this section we shall critique and give a comparison of the outcomes of the two models.

A key observation from the univariate and multivariate analysis section was that the mode of the number of axillary nodes detected was zero. In the medical profession this is referred to as a patient being in stage 0 which means that there is less spread of cancer.

The *k-NN* model is selected primarily because of its non-parametric nature which means that the training phase is the training phase is not extensive and the model also keeps all the training data. In this report, we established that the best value of k or nearest neighbours is 32. This is a relatively large value of k and although the optimal choice of this value is data dependent and although a larger value of k suppresses the noise, it also makes the boundaries less distinct.

The decision tree identifies the variable that is most significant and the value of the variable that would give the best homogenous sets of population. It does this iteratively and splits on all variables to come up with a flow like structure. We sue constraints on the tree size in order to limit the possibility of the tree overfitting.The benefit is that the decision tree is easy to understand and visualize however, in the case of an imbalanced dataset, it can be easily biased.

We observe that the accuracy measure in the *kNN* model is 70 % while that in the decision tree is 69%. In the comparison of these two models, we go a step further and use the confusion matrix specifically, the precision, recall and support. However, it is very clear that the effects of imbalance in the target class affect our interpretation of the observed results.Nonetheless, addressing the imbalance is not the focus of this report although it is taken into account.

To compare these models, we recommend the use of recall in the confusion matrix as we are interested in focussing on weighing the results of the precision and the recall from the confusion matrix. However, the positive class(status 1) is larger and the results reflect mostly the ability of prediction of the positive class and not the negative class(status 2). We see that in the confusion matrix of the *kNN* model, the model recalls 92% of status 1(the majority) but only 23% of status 2(the minority). Similarly, the confusion matrix of the decision tree recalls 92% of status 1 and only 32% of status 2.

## Conclusion

This report highlights the use of two supervised learning machine learning techniques, the *kNN* model and the decision tree model in order to classify the survival status of a patient after breast cancer surgery. It uses a backward elimination technique to find the most informative features, a 10 fold cross validation technique to ensure the validity of the classifiers, and the results are then observed.

The decision tree would be a better model as it is able to recall more of the minority class in light of the class imbalance present in this dataset. This report would recommend the use of techniques such as oversampling of the minority class in the target variable (status 2) or undersampling of the majority class in the target variable(status 1) in order to address the imbalance.

It is recommended that when a patient is detected with Stage 0 cancer, immediate surgery should be done this will increase the survival rate of patient because as per rule lower the number of nodes detected, the less the cancer has spread.Studies have shown that the 5 year relative survival rate for women with stage 0 and 1 (in the context of this report, zero axillary nodes detected) is close to 100% thanks to the advancements in modern medicine.

## References


American Cancer Society(ACS) 2017, *Breast Cancer Survival Rates*, viewed 25 May 2018, <https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html>.

Australian Bureau of Statistics 2018, *Cancer deaths in younger Australians – changes over 20 years*, viewed 13 May 2018, <http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by Subject/3303.0~2016~Main Features~Cancer deaths in younger Australians - changes over 20 years~10000>.

Australian Institute of Health and Welfare (AIHW) 2012, *AIHW Media Releases*, viewed 16 May 2018, <https://www.aihw.gov.au/news-media/media-releases/2012/2012-oct/breast-cancer-survival-improving-but-37-women-sti>.

Batchelor, BG 1978, *Pattern recognition: ideas in practice*, Plenum Press, Berlin, Heidelberg

Cover TM & Hart PE 1967, 'Nearest neighbour pattern classification', *IEEE Trans Inf Theory*, vol. 13, no. 1, pp. 21–27.

Jain, AK, Duin, RPW, Mao J 2000, Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach, vol. 22, no. 1, pp. 4–37.

Kelleher, J, Namee, B & D'Arcy, A 2015, F*undamentals of machine learning for predictive data analytics*, Massachusetts Institute of Technology, United States.

Manning, CD, Raghavan, P & Schutze, H 2008, *An introduction to information retrieval*, Cambridge University Press, Cambridge.

Michalski, RS, Stepp, RE & Diday, E 1981, 'A recent advance in data analysis: clustering objects into classes characterized by conjunctive concepts', *Progress in pattern recognition*, pp 33–56.

World Health Organization 2018, *Cancer*, viewed 19 May 2018, <http://www.who.int/news-room/fact-sheets/detail/cancer>.