**Prediction of Breast Cancer Survival**

Practical Data Science (COSC1295)

May 2018

Team Members:

Janarth Punniyamoorthy

(s3706154@student.rmit.edu.au)

Charles Galea

(s3688570@student.rmit.edu.au)

Lecturer:

Dr. Yongli Ren

Table of contents

Table of Contents

[Dr. Yongli Ren 1](#_Toc513468644)

# Abstract

# Introduction

Breast cancer is a common cancer accounting for 1 in every 4 cancer diagnoses. It is estimated that in 2018 over 18,000 Australians will be diagnosed with breast cancer and more than 3,000 will die from this disease (<https://breast-cancer.canceraustralia.gov.au/statistics>). Early diagnosis can lead to a dramatic increase in long term survival rates from 56% to more than 86%. It has been reported that shown that cancer that has not spread beyond the breast to the lymph nodes has a better prognosis. Metastatic breast cancers which have spread to the lymph nodes and more distant sites account for 90% of all deaths from this disease (Fouad et al., 2015, Peart, 2017). In this report, we have explored whether patient information, including the patient’s age and number of auxillary nodes detected, could be used to predict whether a patient survived longer than 5 years.

# Methodology

## Data

The Haberman’s Survival dataset (Haberman, 1976) was obtained from the University of California, Irvine Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/l>). The dataset contained data from a study conducted between 1958 and 1970 at the University of Chicago’s Billings Hospital on the survival of 306 patients who had undergone surgery for breast cancer. The data comprised: patient age at time of operation, year of operation, the number of auxillary nodes detected and the patient survival status (designated with a value of 1 for patients surviving for 5 years or longer or 2 for patients who died within 5 years) (Table 1).

## Data Analysis

Data manipulation and statistical analyses were performed using the python packages pandas (<https://pandas.pydata.org/>) and NumPy (<http://www.numpy.org/>). The data were initially checked for errors and missing data points. Statistical parameters (mean, standard deviation, minimum, maximum, quantiles) were then determined for the entire patient cohort as well as for each target group (based on survival status).

## Data Visualization

The python plotting package Matplotlib (<https://matplotlib.org/>) was used for data visualization.

## Machine Learning

### *k*-Nearest Neighbors

The *k*NN classifier defines the class of a test instance based on the majority vote of its *k*-nearest neighbors derived from the training data (Friedman et al., 2001). Due to the relatively small size of the dataset *k*NN hyperparameters (*k*, distance metric p and weights) were optimized, to achieve the best balance between the bias and variance of the model, using 5-fold cross validation. The distance metric was determined by the following equation:

Where and .

For 5-fold cross-validation the data was divided into 5 folds and the *k*NN model was applied to make predictions on the 5th segment. Predictions were performed using the equation:

Where is the ith case of the sample and is the prediction of the query point.

The process was then applied successively applied to other 4 segments. The computed errors were then averaged to yield a measure of performance of the model. These steps were then repeated for various parameter values and the value achieving the lowest error (or highest classification accuracy) was then chosen as the optimal value. The *k*NN classifier (*K*NeighborsClassifier) was implemented using the python sklearn package (Pedregosa et al., 2011).

### Data Scaling for *k*NN classifier

The *k*NN classifier was also tested on the dataset following feature scaling to compensate for differences in the measurement scale of the various features. Two scaling methods (standardization and normalization) (Kelleher et al., 2015) were used to rescale each feature prior to applying the *k*NN classifier. The data were standardized to a standard normal distribution with and using the formula:

Where is the standard score (also called the z-score) is the mean (average) and is the standard deviation from the mean. The mean was calculated using the equation:

Where is the total number of observations and is the ith observation. While the standard deviation was derived from the equation:

The data were normalized (also know as Min-Max scaling) by applying the following formula:

Where is the normalized feature value, is the original value, is the minimum value of the feature and is the maximum value of the feature.

### Decision Trees

The decision tree was built by partitioning instances into local subsets using recursive splits. ……….

## Model Evaluation

Model performance was evaluated using the following parameters:

Where was the number of false positives (negative target features instances incorrectly predicted as positive), was the number of false negatives (positive target features instances incorrectly predicted as negative), was the number of true positives (correctly predicted positive target feature instances) and was the number of true negatives (correctly predicted negative target feature instances).

The average class accuracy was used to determine classification accuracy to account for the imbalance between patient survival status values in the target feature.

Where is the set of levels that the target feature, , can assume; is the size of the set; and refers to the recall achieved by the model for level (Kelleher et al., 2015).

The Area Under Curve (AUC) metric was utilized to compare the performance of the *k*NN and Decision Tree classifiers. AUC is the area under the ROC (Receiver Operator Characteristic) curve, which is a plot of False Positive Rate (FPR) versus True Positive Rate (TPR). Where the TPR is identical to the Recall term outlined above and the FPR is:

# Results

## Data Pre-processing

### Statistical Analysis

Data collected for the Haberman Survival Studies (Haberman, 1976) for 306 breast cancer patients was used to analysis several classification models for the prediction of caner patient survival. The data consisted of the features: patient age, year of operation, number of positive auxillary nodes detected and survival status (Table 1).

*Table 1.* Haberman Breast Cancer Dataset

|  |  |
| --- | --- |
| Number of patients | 306 |
| Patient ages | 30-83 |
| Year of operation | 1958-1969 |
| Number of nodes | 0-52 |
| Patient survival status |  |
| Less than 5 years | 81 (73.5%) |
| 5 years or longer | 225 (26.5%) |

Table 1. Statistical parameters for the entire breast cancer patient cohort.

|  | **Patient Age** | **Year of Operation** | **Number of Nodes** | **Survival Status** |
| --- | --- | --- | --- | --- |
| **count** | 306 | 306 | 306 | 306 |
| **mean** | 52.46 | 62.85 | 4.03 | 1.26 |
| **std** | 10.80 | 3.25 | 7.19 | 0.44 |
| **min** | 30.0 | 58.0 | 0.0 | 1.0 |
| **25%** | 44.0 | 60.0 | 0.0 | 1.0 |
| **50%** | 52.0 | 63.0 | 1.0 | 1.0 |
| **75%** | 60.8 | 65.8 | 4.0 | 2.0 |
| **max** | 83.0 | 69.0 | 52.0 | 2.0 |

Table 2. Statistical parameters for the entire breast cancer patients surviving for 5 years or longer.

|  | **Patient Age** | **Year of Operation** | **Number of Nodes** | **Survival Status** |
| --- | --- | --- | --- | --- |
| **count** | 225 | 225 | 225 | 225 |
| **mean** | 52.02 | 62.86 | 2.79 | 1.00 |
| **std** | 11.01 | 3.22 | 5.87 | 0.00 |
| **min** | 30.0 | 58.0 | 0.0 | 1.0 |
| **25%** | 43.0 | 60.0 | 0.0 | 1.0 |
| **50%** | 52.0 | 63.0 | 0.0 | 1.0 |
| **75%** | 60.0 | 66.0 | 3.0 | 1.0 |
| **max** | 77.000000 | 69.000000 | 46.000000 | 1.0 |

Table 3. Statistical parameters for the entire breast cancer patients surviving for less than 5 years.

|  | **Patient Age** | **Year of Operation** | **Number of Nodes** | **Survival Status** |
| --- | --- | --- | --- | --- |
| **count** | 81 | 81 | 81 | 81 |
| **mean** | 53.68 | 62.83 | 7.46 | 2.00 |
| **std** | 10.17 | 3.34 | 9.18 | 0.00 |
| **min** | 34.0 | 58.0 | 0.0 | 2.0 |
| **25%** | 46.0 | 59.0 | 1.0 | 2.0 |
| **50%** | 53.0 | 63.0 | 4.0 | 2.0 |
| **75%** | 61.0 | 65.0 | 11.0 | 2.0 |
| **max** | 83.0 | 69.0 | 52.0 | 2.0 |

A screenshot of a cell phone

Description generated with very high confidence

Fig. 1. (A) Ages, (B) year of operation and (C) number of auxiliary nodes for patients in the breast cancer cohort. The data were plotted as a histogram (top) and boxplot (bottom) to highlight the distribution of the data.

A close up of a map

Description generated with high confidence

Fig. 2. Comparison of the distribution of (A) ages, (B) year of operation and (C) number of auxiliary nodes for breast cancer patients surviving for 5 years or longer (blue) or less than 5 years (orange). The data were plotted as a histogram overlaid with the corresponding density plot (solid line).

### Feature transformation

A close up of a map

Description generated with high confidence

Fig. 3. Cubed root transformation of the data for the number of auxiliary feature results in an improved distribution of the data. The highly right skewed data for the number of auxiliary nodes (A) was processed with a cubed root function to obtain a better distribution of the data (B). The histogram is show at top while the corresponding boxplot is on the bottom of the figure. (C) and (D) show the histograms and density plots (solids lines) for breast cancer patients surviving for 5 years or longer (blue) or less than 5 years (orange).



Fig. 4. Pie chart illustrating the proportion of breast cancer patients surviving for 5 years or longer (blue) or less than 5 years (orange).

A picture containing text

Description generated with high confidence

Fig. 5. Comparison of survival status (blue - < 5 years and orange - > 5 years) for (A) patient age, (B) year of operation and (C) number of auxiliary nodes.

A screenshot of a cell phone

Description generated with high confidence

Fig. 6. Scatter plot for patient age, year of operation and cube root transformed data for the number of auxiliary nodes.

A picture containing screenshot

Description generated with very high confidence

Fig. 7. 2D contour plots examining the distribution of breast cancer patient age, year of operation and umber of auxiliary nodes.

### Feature Scaling

A close up of a map

Description generated with high confidence

Fig. 8. Comparison of the distribution of breast cancer data before (green) and after standardization (red) or normalization (blue) of the data to compensate for differences in the measurement range for each feature. Plots of patient age versus (A) year of operation and (B) number of auxiliary nodes and year of operation versus number of auxiliary nodes.

A close up of a map

Description generated with high confidence

Fig. 9. Comparison of (A) standardized and (B) normalized data for the breast cancer dataset.

A close up of a map

Description generated with very high confidence

Fig. 11. Scatter plots for (A) standardized and (B) normalized breast cancer data.

## Prediction Modelling

### *k*-Nearest Neighbor (*k*NN) Classifier

# Discussion

# Conclusions

# References

FOUAD, T. M., KOGAWA, T., LIU, D. D., SHEN, Y., MASUDA, H., EL-ZEIN, R., WOODWARD, W. A., CHAVEZ-MACGREGOR, M., ALVAREZ, R. H., ARUN, B., LUCCI, A., KRISHNAMURTHY, S., BABIERA, G., BUCHHOLZ, T. A., VALERO, V. & UENO, N. T. 2015. Overall survival differences between patients with inflammatory and noninflammatory breast cancer presenting with distant metastasis at diagnosis. *Breast cancer research and treatment,* 152**,** 407-416.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. 2001. *Elements of Statistical Learning: Data mining, inference, and prediction.,* New York, NY, Springer.

HABERMAN, S. J. 1976. Generalized residuals for log-linear models. *Proceedings of the 9th International Biometrics Conference.* Boston.

KELLEHER, J., MAC NAMEE, B. & D'ARCY, A. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*.

PEART, O. 2017. Metastatic Breast Cancer. *Radiologic technology,* 88.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M. & DUCHESNAY, É. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research,* 12**,** 2825-2830.