*Journal of Applied Statistics & Machine Learning*
Volume 1; Number 2; 2022 : pp. 89-100

# REVISITING SURVIVABILITY PREDICTION OF BREAST CANCER WITH MACHINE LEARNING TOOLS

**Mandana Bozorgi[1*], Kazem Taghva[2] and Ashok Singh[3]**

[1*]*Computer Science Department, Washington State University, Vancouver, USA. E-mail: mandana.bozorgi@wsu.edu*
[2]*Computer Science Department, University of Nevada Las Vegas, USA. E-mail: kazem.taghva@unlv.edu*
[3]*RGGM Department ,University of Nevada Las Vegas, USA. E-mail: ashok..singh@unlv.edu*

ARTICLEINFO

Received: 27 June 2022
Revised: 29 July 2022
Accepted: 12 August 2022
Online: 26 December 2022

*To cite this paper:*
Bozorgi, M., Taghva, K., & Singh, A. (2022). Revisiting Survivability Prediction of Breast Cancer with Machine Learning Tools. *Journal of Applied Statistics & Machine Learning*. 1(2): pp. 89-100.

ABSTRACT

This paper revisits the problem of five-year survivability predictions for breast cancer using machine learning tools. This work is distinguishable from the past experiments based on the size of the training data, the unbalanced distribution of data in minority and majority classes, and modified data cleaning procedures. These experiments are also based on the principles of TIDY data and reproducible research. In order to fine-tune the predictions, a set of experiments were run using naive Bayes, decision trees, and logistic regression. Of particular interest were strategies to improve the recall level for the minority class, as the cost of misclassification is prohibitive. The main contribution of this work is that logistic regression with the proper setting of class weight gives the highest precision / recall level for the minority class. In addition, this work provides precise algorithms and codes for determining class membership and execution of competing methods. These codes can facilitate the reproduction and extension of our work by other researchers.

***Keywords:*** Machine Learning, Big Data, Learning Algorithm, Logistic Regression, Classification, ROC

## 1. INTRODUCTION

According to the National Breast Cancer Organization [Comprehensive Cancer Information, 2016], “Breast cancer is a disease in which malignant (cancer) cells form in the tissues of the breast.” Over 230,000 women are diagnosed with breast cancer in the United States annually [Comprehensive Cancer Information, 2016]. In addition, about one in eight women will develop breast cancer. These alarming statistics have led to tremendous research efforts and studies associated with breast cancer in recent years. In addition, many organizations have compiled statistical data pertaining to individual patients. One such database is Surveillance, Epidemiology, and End Results (SEER) database which is maintained by National Cancer Institute (SEER) [Surveillance, epidemiology, and end results, 2016]. The SEER database is a rich source of information for statistical learning analysis.

For example, Abdelghani and Guven carried out a comparative study of three data mining techniques in order to predict five-year survivability based on SEER data [Abdelghani, B., & Guven, E. (2006)]

Since SEER database is updated on a regular basis with new patients, it is logical to repeat some of the past experiments. As the first step, we wanted to repeat the same experiments to establish a basis for comparison with a new up- dated SEER data. It turns out that we could not repeat experiments reported by Bellaachia and Guven (2006) and Delen, Walker, and Kadam (2005). This is because the reported data preparation, cleaning, and data processing were incomplete and ambiguous. As a result, these cited works were not reproducible (Peng, 2011).

This paper revisits the topic of prediction of five-year survivability for breast cancer with machine learning tools, following the principles of TIDY data and reproducible research as discussed by Peng (2011) and Wickham 2014). Of particular interest in how to set up an environment that other researchers could use to apply the same techniques on other types of cancer.

This paper is organized into six sections, including this introduction. Section 2 summarizes some of the notable works associated with data science, SEER database, and machine learning techniques. Section 3 gives a brief introduction to specific techniques used in this work. Section 4 provides detailed description of the experiments, and Section 5 provides some statistics on the performance of the methods used in this study. Section 6 explores directions for future work.

## 2. BACKGROUND

The primary goal of many artificial intelligence (AI), machine learning, and data science is the discovery of new facts from data based on statistical and logical methods. The secondary goal of these disciplines is to communicate the new facts (Aumann *et al.*, 2003; Dhar, 2013). Of course, the discovery should be valid and reproducible. Unfortunately, many reported discoveries are not reproducible due to sloppy data preparation and clean up (De Fraja, Oliveira, & Zanchi, 2010).

Typically, many projects use data sets that were not necessarily collected for those projects. For example, SEER database is built for summarizing cancer data and not survivability prediction. The survivability prediction problem is a binary classification with uneven distribution of data points (Vapnik, 1995; Xiao, et al. 2009). In order to prepare SEER data for binary classification, we must first decide how to assign data points to each class. The most widely used metric involves calculating the percentage of patients alive after five years, using a direct method outlined by Parkin and Hakulinen (1991). Section 4.1 provides a detailed explanation of our approach to data assignments based on the direct method.

One of the earliest and most cited work on survival predictability with machine learning tools are the experiments reported by Delen, Walker & Kadam (2005). These experiments identified decision tree as the best predictor, compared with artificial neural networks (ANN) and logistic regression. A follow-up set of experiments by Bellaachia and Guven (2006) reported similar results that decision tree was superior to naive Bayes and ANN. Neither work was reproducible research, as there is no code book description of recipes on data preparation and algorithms. Both of the above-mentioned studies were conducted using SEER data. Closely related studies on lung cancer using SEER data found that decision tree was the best predictor (Agrawal, Misra, Narayanan, Polepeddi, & Choudhary, 2012). This study further identified the importance of two out of 11 features when predicting survivability.

In another interesting and related study using SEER data, (Zolbanin, Delen, & Zadeh, 2015) the prediction of survivability on comorbidity of cancers, for example, breast and prostate cancer, was investigated.

Salama, Abdelhalim, & Zeid (2012) performed comparison studies on Wisconsin Breast Cancer (WBC) database (Lichman, 2013), and reported that Multi-Layer Perception (MLP) was superior to decision tree for that database. It is important to point out that WBC collects a different set of features for breast cancer than does SEER. It is also worth mentioning that another study (Christobel & Sivaprakasam, 2011) identified the Support Vector Machine (SVM) as the best predictor for the WBC database. Finally, we want to draw attention to binary classification based on missense mutation in genome (Wei & Dunbrack Jr, 2013).

## 3. MACHINE LEARNING TOOLS

This section provides a brief introduction to binary classification with naive Bayes, logistic regression, and decision tree. In general, classification starts with a vector of features $\vec{X} = (x_1, x_2, \ldots, x_n)$ which can serve as a template for each data point in the data set. We wanted to build a binary classifier $Y$ that predicts survivability. Essentially this construction was based on the characteristics of the initial data set, in this case, the SEER database.

The simplest learning algorithm is the naive Bayes (Friedman, Geiger & Goldszmidt, 1997). This classification technique relies on Bayes' rule that the outcome of an event *A* can be predicted from evidence *B*:

$$P(A \mid B) = \frac{P(B \mid A).P(A)}{P(B)} \tag{1}$$

In practice, there are more events (or features) that contribute to this equation. The word naive stems from the fact that features $X_i^{'}s$ are assumed to be independent of each other.

Notice that the numerator is the joint probability *P* (*A*, *B*). For a more general vector of features $\vec{X}$, this joint probability for a new data point to be classified is simply the product of the individual probabilities:

$$P(X_1, X_2, \ldots, X_n) = P(X_1) \;.\; P(X_2). \;\ldots\; . \;P(X_n) \tag{2}$$

With logistic regression [Lin, C. J., Weng, R. C., & Keerthi, S. S. (2008)], the feature vector $\vec{X}$ is used to fit the data point in the equation:

$$P(\vec{X}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + + \beta_n x_n \tag{3}$$

Since this value is not necessarily between 0 and 1, a link function, logit is used:

$$P(\vec{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n}} \tag{4}$$

The Maximum Likelihood Estimate (MLE) is used to find the values of the coefficients β*i*'s from the data. The decision tree [Quinlan, J. R. (1986)] uses a tree structure to classify the data points. The leaves represent classes (survived or not), and branches represent conjunction of features from the feature vector. This is a popular method as it represents a conceptual thought process that one can start at the root and make conclusions at the leaves.

## 4. METHODOLOGY

Three distinct steps in running experiments with machine learning tools are data preparation, program execution, and interpretation of the results. The first two steps are discussed in this section and a detailed explanation of the third step is given in Section

### 4.1. Data Preparation

As mentioned in Section 1, we first were interested in reproducing the experiments reported by Bellaachia and Guven (2006) in order to extend the work on the more recent SEER data. Unfortunately, neither the data sets nor the results could be reproduced, mainly due to the lack of exact and explicit instructions for data preparation. This is very common in scientific literature and major obstacle in reproducible research (Peng, 2011; Baggerly & Coombes, 2009). Following Wickham, 2014), the data preparation must include four components:

1. The raw data
2. A TIDY data set
3. A code book describing each variable and its value
4. An explicit and exact recipe from which one needs to produce components one and two from component oneThe raw data we used is the data repository as reported in "SEER RESEARCH DATA RECORD DESCRIPTION CASES DIAGNOSED IN 1973-2013" (Breast Cancer Information and Support, 2016). This repository contains 769,261 records with 134 attributes. Since the records cover various kinds of cancer, not all attributes apply to our work on breast cancer. Furthermore, there is a set of attributes that only applies to data collected after 1988. One such set used for this study was EOD Tumor Size, EOD Extension, EOD Lymph Nodes; the data for this set were collected from 1988 to 2003. The same data was collected after 2003 with different labels and positions (columns), namely, CS Tumor Size, CS Extension, and CS Lymph Node Involv, respectively. We used 18 attributes, as described in Table 1.

The attributes patientId, COD, year of Diagnosis, and survival Months were not used as features for classification. However, survival Months, year of Diagnosis, and COD were used to label the two classes for binary classification.

Since a patient may have more than one record reflecting different visits, we decided to only consider the first visit record. Also, some of the attributes such as tumour size was only collected after 1988; therefore, we selected only patients with a first diagnosis made during or after 1988. This produced 545,188 records.

The next step in data preparation and cleaning was to label records based on five year's survivability, according to direct method as outlined by Parkin and Hakulinen [Parkin, D. M., & Hakulinen, T. (1991)]. It worth

**Table 1: The 18 Attributes Used for the Experiments in This Study**

| *Variable* | *Variable Definition* | *Values* |
|---|---|---|
| patient Id Number | uniquely identifies a patient | up to 8 digits |
| race | Two digits code race identifier | 01-99, 01 for white,02 for black |
| marital Status | one digit code for marital status | 1-9, 1 for single, 2 for married |
| behavior Code | code for benign etc. | o-4,0 for benign,1 for malignant potential, etc. |
| grade | cancer grade | 1-9, 1 for Grade I, etc. |
| vital Status Record | alive or not | 1-4, 1 for alive, 4 for dead |
| histologic Type | microscopic composition of cells | 4-digit code |
| cs Extension | extension of tumor | 2-digit code |
| csLymphNode | involvement of lymph nodes | 2-digits code |
| radiation | radiation type code | 0-9, for none, 1 for Beam, etc. |
| SEERHistoricStageA | codes for stages | 0-9, 0 for in situ, 1 for localized |
| age at Diagnosis | First diagnosis age | 00-130, actual age, 999 for unknown |
| csTumorSize | size in millimeters | 000-888, 000 for no tumor |
| regional Nodes Positive | negative vs positive nodes | 00-99, exact number of positive nodes |
| regional Nodes Examined | positive and negative nodes examined | 00-99, exact number |
| survivalMonths | number of months alive | 000-998, exact number of months, 9999 for unknown |
| COD | Cause of Death | 5-digit code, 2600 for breast cancer, 00000 alive |
| year of Diagnosis | This visit year | 4-digit code |

mentioning that many of the studies on SEER data ignored this step (e.g., Bellaachia and Guven; and Delen et al.) [Abdelghani, B., & Guven, E. (2006), Delen, D., Walker, G., & Kadam, A. (2005)]. Consider the three patient records shown in Table 2. There are four records for patient 1; the first record showed that the patient survived 110 months from the visit in October of 2004. Based on this record, Patient 1 will be labelled as survived. Patients 2 survived 47 months from the date of first visit in January 2010; this patient will be marked as ignore and was not used for training. Patients 3 and 4 are both deceased and the cause of death for both patients is breast cancer. Patient 3 survived beyond five years, so she was labelled as survived. Patient 4 was labelled as not-survived. The record of the first visit for each patient for training purposes. Finally, any record that had had empty or unknown values in regionalNodesPositive, regionalNodesExamined, CSTumorSize, and EODTumersize were removed. Out of 338,596 patients, 300,215 were labelled survived and 38,381 were labelled not-survived.

Note that the number of survived data points were almost eight times the number of not-survived data points.

**Table 2: A Sample of Four Patients Records**

| *PatientId* | *VSR* | *STR* | *month of Diagnosis* | *year of Diagnosis* | *COD* |
|---|---|---|---|---|---|
| 1 | 1 | 110 | 10 | 2004 | 00000 |
| 1 | 1 | 85 | 11 | 2006 | 000000 |
| 1 | 1 | 15 | 9 | 2012 | 00000 |
| 1 | 1 | 14 | 10 | 2012 | 00000 |
| 2 | 1 | 47 | 1 | 2010 | 00000 |
| 2 | 1 | 9 | 3 | 2013 | 00000 |
| 2 | 1 | 8 | 5 | 2013 | 00000 |
| 3 | 4 | 96 | 3 | 2005 | 2600 |
| 3 | 4 | 46 | 5 | 2009 | 2600 |
| 4 | 4 | 23 | 7 | 2006 | 2600 |
| 4 | 4 | 22 | 8 | 2006 | 2600 |

## 4.2. Experiments

Many natural problems can be solved using binary classification techniques. Known examples of binary classifications are the detection of fraudulent credit card fraudulent transactions (Phua, Alahakoon & Lee, 2004), spam identification (Benevenuto, Rodrigues, Almeida, Almeida & Gonçalves, 2009), classified documents (Taghva, 2009), and privacy detection (Taghva, Beckley & Coombs, 2006). Naive Bayes, decision trees, logistic regression, artificial neural network (ANN), and support vector machine (SVM) are among the most popular techniques for binary classification. In this study, the performance of naive Bayes, decision trees, and logistic regression were evaluated for their performance in predicting five-year survivability of breast cancer patients. These three approaches were chosen because they were techniques used in past studies on survivability prediction. The implementations for these three approaches developed by Pedregosa et al. (2011) were used in these experiments.

As mentioned previously, the number of data points in the survived class is eight times the number of not-survived data points. Typically, this imbalance affects the classification accuracy (Wei & Dunbrack Jr, 2013). Many approaches have been developed to overcome the problems associated with the unbalanced training data. The simplest one is to provide the prior weights of the training class to the classifier. The balanced value for class-weight parameter for both decision tree and logistic regression experiments. In addition, the class prior (0.12, 0.88) was used for naive Bayes experiments. Stratified 10-fold cross validation was used for training and testing to make sure that each fold preserved a similar distribution as the original classes. Aside from the default setting, the only other parameter used was newton method for the solver method of the logistic regression.

## 5. RESULTS

Regarding the prediction accuracy when using precision/recall metrics and ROC curve, in the 10-fold cross validation method, the entire data set was split into 10 random sub-samples. Each classifier uses nine folds for training and one-fold for testing. The final confusion matrix is the average of the 10 runs. Let $TP$ be the number of true positives, that is, the number of patients which the classifier predicts survived and the patients actually have survived. Let $FN$ be the number of false negatives, i.e., the number of patients that actually survived but the classifier predicts not-survived.

**Table 3: Confusion Matrix**

| | *Predict No* | *Predict Yes* |
|---|---|---|
| Actual No | True Negative (TN) | False Positive (FP) |
| Actual Yes | False Negative (FN) | True Positive (TP) |

**Table 4: Performance of the Classifiers**

| *Classifier* | *class* | *Precision* | *Recall* | *F1* |
|---|---|---|---|---|
| Naive Bayes | survivednot-survived | 0.361.00 | 0.990.77 | 0.530.87 |
| Logistic Regression | survivednot-survived | 0.411.0 | 0.970.82 | 0.580.90 |
| Decision Tree | survivednot-survived | 0.600.95 | 0.590.95 | 0.600.95 |

The $TN$ is defined as the number of patients that have not-survived and the classifier also predicts not-survived. The $FP$ is the number of patients that have not-survived but the classifier falsely predicts survived. These four metrics are typically summarized in a confusion matrix as shown in Table 3.

Recall then is defined as:

$$recall = \frac{TP}{TP + FN} \tag{5}$$

And the precision is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

The harmonic mean of precision and recall is called the F 1 measure, defined as:

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{7}$$

The Receiver Operating Characteristic (ROC) curve is extensively used for the performance of binary classification. The ROC curve exhibits the trade-off between true positive and false positive error rates (Duda, Hart & Stork, 2012). The Area Under the Curve (AUC) is the accepted measure of the binary classification performance. The performance of the tree classifiers with 10-fold cross validation is summarized in Table 4 and Figure 1.

The *precision* reports the percentage of data points that are classified as positive that are actually positive. The *recall* reports the percentage of

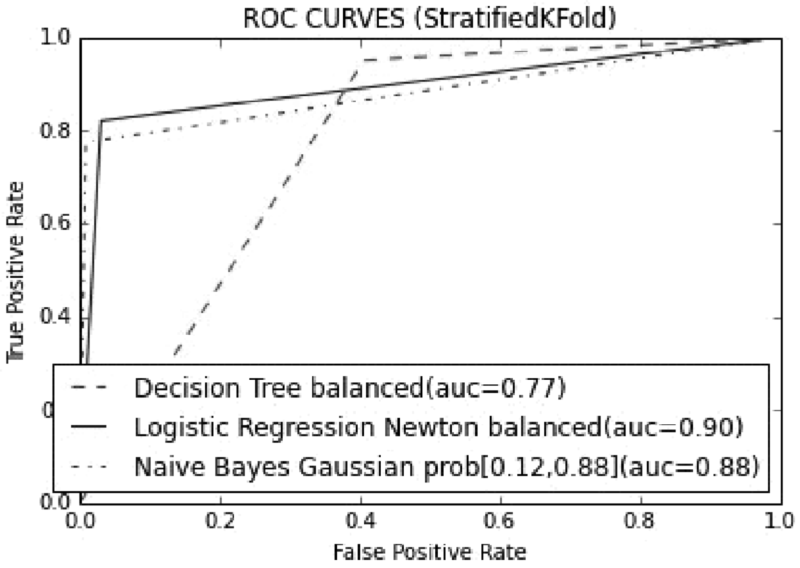

**Figure 1: ROC Curve**

correctly labelled data points. Precision is sensitive to the class distribution. In general, the precision is affected by the class distribution while recall is not. All three methods have low precision for the not-survived class, but both logistic regression and Naive Bayes have very high recall values for this class. This is a crucial point as the cost of misclassification is prohibitive for this class. The idea being that when a patient is put in the not-survived class, then we may require further test to be assured of the patient condition. The ROC curve suggests that logistic regression is also superior based on the AUC value. The difference between AUCs for Naive Bayes and logistic regression may not be statistically significant. A closer look at the coefficients reveals that race and vitalStatusRecord are not significant and can be eliminated.

## 6. CONCLUSION AND FUTURE WORK

This paper reported on application of machine learning tools for predicting cancer survivability. This work was based on reproducible research principle, a larger data set, and unbalanced nature of cancer data set. Results indicate that logistic regression is a good choice for cancer prediction as compared to decision trees and naive Bayes.

There are three possible extensions to this project that we are currently pursing. The first extension is to apply the Synthetic minority over-sampling technique (SMOTE) to balance the training set (Bozorgi, Taghva & Singh, 2017). Second extension is to apply these experiments to other types of cancers using SEER data. The third extension is to build a web-based application that could be used as an advisory tool for survivability prediction.

## *References*

1. Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L., & Choudhary, A. (2012). Lung cancer survival prediction using ensemble data mining on SEER data. *Scientific Programming*, *20*, 29-42.

2. Abdelghani, B., & Guven, E. (2006). Predicting breast cancer survivability using data mining techniques. 2010 2nd International Conference on Software Technology and Engineering, 2010, pp. V2-227-V2-231, doi: 10.1109/ICSTE.2010.5608818.

3. Aumann, H. H., Chahine, M. T., Gautier, C., Goldberg, M. D., Kalnay, E., McMillin, L. M., ... & Susskind, J. (2003). AIRS/AMSU/HSB on the Aqua mission: Design, science objectives, data products, and processing systems. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(2), 253-264.

4. Baggerly, K. A., & Coombes, K. R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *The Annals of Applied Statistics*, 1309-1334.

5. Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), 113-127.
6. Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-73.
7. Bozorgi, M., Taghva, K., & Singh, A. (2017). Cancer survivability with logistic regression. 2017 Computing Conference, pp. 416-420.
8. Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. Decision Support Systems, 74, 150-161.
9. Xiao, Y., Liu, B., Cao, L., Wu, X., Zhang, C., Hao, Z., ... & Cao, J. (2009). Multi-sphere support vector data description for outlier detection on multi-distribution data. IEEE international conference on data mining workshops (pp. 82-87). IEEE.
10. Wickham, H. (2014). Tidy data. Journal of statistical software, 59(10), 1-23.
11. Wei, Q., & Dunbrack Jr, R. L. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. PloS one, 8(7), e67863.
12. Vapnik, V.N. (1995). The Nature of Statistical Learning Theory.Springer-Verlag, Berlin, Heidelberg.
13. Taghva, K., Beckley, R., & Coombs, J. (2006, February). The effects of OCR error on the extraction of private information. In *International Workshop on Document Analysis Systems* (pp. 348-357). Springer, Berlin, Heidelberg.
14. Taghva, K. (2009). Identification of sensitive unclassified information. In *Computational Methods for Counterterrorism* (pp. 89-108). Springer, Berlin, Heidelberg.
15. Salama, G. I., Abdelhalim, M., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, *32*(569), 2.
16. Lichman, M. (2013). UCI machine learning repository.Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.
17. Peng, R. D. (2011). Reproducible research in computational science. Science Dec 2011: Vol. 334, Issue 6060, pp. 1226-1227
18. Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter*, *6*(1), 50-59.
19. Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern Classification. Wiley-Interscience.
20. Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., & Gonçalves, M. (2009, July). Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 620-627).
21. Christobel, A., & Sivaprakasam, Y. (2011). An empirical comparison of data mining classification methods. International Journal of Computer Information Systems, 3(2), 24-28.

22. Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. Machine learning, 29(2), 131-163.

23. Lin, C. J., Weng, R. C., & Keerthi, S. S. (2008). Trust region Newton method for large-scale logistic regression. *Journal of Machine Learning Research,* 9(4).

24. Parkin, D. M., & Hakulinen, T. (1991). Analysis of survival. Cancer Registration, Principles and Methods. IARC Scientific Publications, (95), 159-176.

25. Quinlan, J. R. (1986). Induction of decision trees. *Machine learning,* 1(1), 81-106.

26. Comprehensive Cancer Information. National Cancer Institute. (n.d.). *https://www.cancer.gov/.*

27. Breast Cancer Information and Support. (2016). *https://www.breastcancer.org/*

28. Surveillance, epidemiology, and end results (2016). *https://seer.cancer.gov*

29. The Economist. How science goes wrong, (2013). http://www.economist.com/news/leaders/21588069- scientific-research-has-changed-world-now-it-needs-change-itself-how-scien.

30. De Fraja, G., Oliveira, T., & Zanchi, L. (2010). Must try harder: Evaluating the role of effort in educational attainment. *The Review of Economics and Statistics,* 92(3), 577-597.