Predicting Breast Cancer Survivability A Comparison of Three Data Mining Methods

Omead I. Hussain

Department of Banking and Financial Science, Cihan University-Erbil,
Kurdistan Region, Iraq

***Abstract*—This study concentrates on predicting breast cancer survivability using data mining, and comparing between three main predictive modeling tools. Precisely, we used three popular data mining methods: Two from machine learning (artificial neural network [ANN] and decision trees) and one from statistics (logistic regression) and aimed to choose the best model through the efficiency of each model and with the most effective variables to these models and the most common important predictor. We defined the three main modeling aims and used by demonstrating the purpose of the modeling. By using data mining, we can begin to characterize and describe trends and patterns that reside in data and information. The preprocessed dataset contents were of 87 variables and the total of the records are 457,389; which became 93 variables and 90308 records for each variable, and these datasets were from the SEER database. We have achieved more than three data mining techniques and we have investigated all the data mining techniques and finally, we find the best thing to do is to focus about these data mining techniques which are ANN, Decision Trees, and Logistic Regression using SAS Enterprise Miner 5.2 which is in our view of point are the suitable system to use according to the facilities and the results are given to us. Several experiments have been conducted using these algorithms. The achieved prediction implementations are comparison-based techniques. However, we have found out that the neural network has a much better performance than the other two techniques. Finally, we can say that the model we chose has the highest accuracy which specialists in the breast cancer field can use and depend on.**

***Keywords*—Predicting breast cancer, Data mining, SEER database, Artificial neural network.**

I. Introduction

In their worldwide end-user business analytics forecast, IDC, a world leader in the provision of market information, divided the market, and differentiated between "core" and "predictive" analytics (Vesset and Chua, 2017). Breast cancer is cancer that forms in breast tissues and is classed as a malignant tumor when cells in the breast tissue divide and grow without the normal controls on cell death and cell division. We know from looking at breast structure that it contains ducts (tubes that carry milk to the nipple) and lobules (glands that make milk) (Holland, 2008). Breast cancer can occur in both men and women, although breast cancer in men is rarer and so breast cancer is one of the common types of cancer and major causes of death in women in the UK. In the past 10 years, breast cancer rates in the UK have increased by 12%. In 2004, there were 44,659 new cases of breast cancer diagnosed in the UK: 44,335 (99%) in women and 324 (1%) in men. Breast cancer risk in the UK is strongly related to age, with more than (80%) of cases occurring in women over 50 years old. The highest number of cases of breast cancer is diagnosed in the 50–64 age groups. Although very few cases of breast cancer occur in women in their teens or early 20s, breast cancer is the most commonly diagnosed cancer in women under 35. By the age of 35–39 years, almost 1500 women are diagnosed each year. Breast cancer incidence rates continue to increase with age, with the greatest rate of increase before the menopause.

As the incidence of breast cancer is high and 5-year survival rates are over 75%, many women are alive who have been diagnosed with breast cancer (Holland, 2008). The most recent estimate suggests around 172,000 women are alive in the UK having had a diagnosis of breast cancer. Even though in the last couple of decades, with their increased emphasis toward cancer-related research, new and innovative methods of detection and early treatment have developed which help to reduce the incidence of cancer-related mortality (Edwards et al., 2002), cancer in general and breast cancer to be specific is still a major cause of concern in the United Kingdom.

Although cancer research is in general clinical and/or biological in nature, data-driven statistical research is becoming a widespread complement in medical areas where data and statistics driven research is successfully applied.

For health outcome data, an explanation of model results becomes really important, as the intent of such studies is to get knowledge about the underlying mechanisms.

Cihan University-Erbil Journal of Humanities and Social Sciences
Volume IV No. 1(2020); 14 pages
DOI: 10.24086/cuejhss.vol4n1y2020.pp17-30
Received 26 August 2019; Accepted 26 October 2019; Regular research paper: Published 10 February 2020
*Corresponding author's e-mail: omead.hussain@cihanuniversity.edu.iq

Problems with the data or models may indicate a common understanding of the issues involved, which is contradictory. Common uses of the models, such as the logistic regression model, are interpretable. We may question the interpretation of the often inadequate datasets to predict. Artificial neural networks (ANN s) have proven to produce good prediction results in classification and regression problems. This has motivated the use of ANN) on data that relate to health results such as death from breast cancer disease or its diagnosis. In such studies, the dependent variable of interest is a class label, and the set of possible explanatory predictor variables – the inputs to the ANN – may be binary or continuous (Allison, 2001).

Predicting the outcome of an illness is one of the most interesting and challenging tasks in which to develop data mining applications. Survival analyses are a section in medical speculation that deals with the application of various methods to historic data to predict the survival of a specific patient suffering from a disease over a particular time period.

With the rising use of information technology powered with automated tools, enabling the saving and retrieval of large volumes of medical data, this is being collected and being made available to the medical research community who are interested in developing prediction models for survivability (Chow et al., 1994).

II. Methodology

A. Research Aims and Objectives

The objective of the present presentation is to significantly enhance the efficiency of the accuracy of the three models we chose. Considering the justification of the high efficiency of the models, it was decided to embark on this research study with the intended outcome of creating an accurate model tool that could both build calculate and depict the variables of overall modeling and increase the accuracy of these models and the significance of the variables.

For the purposes of this study, we decided to study each attribute individually and to know the significant of the variables which are strongly built into the models. Furthermore, for the first iteration of our simulation for choosing the best model (Intrator and Intrator, 2001), we decided to focus on only three data mining techniques which were mentioned previously. Having chosen to work exclusively with SAS systems, we also felt it would be advantageous to work with SAS rather than other software since this system is most flexible.

After duly considering feasibility and time constraints, we set ourselves the following study objectives:

a. Propose and implement the three models which are selected and applied, and their parameters are calibrated to optimal values and to measure and predict the target variable (0 for not survive and 1 for survive)
b. Propose and implement the best model to measure and predict the target variable (0 for not survive and 1 for survive)
c. To be able to analyze the models and to see which variables have the most effect on the target variable
d. To visualize the aforementioned target attributes through simple graphical artifacts
e. Built the models that appear to have high quality from a data analysis perspective.

B. Activities

The steps taken to achieve the above objectives can be summarized as below. As mentioned, the study consisted of building the model, which has the highest accuracy and analyzing the three models we chose.

Points (a) and (b) relate to the data preparation of the study, points (c) and (d) relate to the build of the model, and points (e) through (g) relate to the analyze of the models:

a. To characterize and describe trends and patterns that reside in data and information about the data
b. To choose the records, as well as evaluating these transformation and cleaning of data for modeling tools. Cleaning of data contains the estimate of missing data by modeling (mean, mode, etc.)
c. Selecting modeling techniques and applying their parameters, requirements on the form of data, and applying the dataset of our choosing
d. Evaluation of the model and review of the steps executed to construct the model to achieve the business objectives
e. To be able to analyze the models and to see which variables are more applicable to the target variable
f. Decide on how the decision on the use of the data mining result should be reached
g. SAS software to be able to get the best results and analyze the variables which are most significant to the target variable.

C. Data Source

We decided to use a dataset which is a compatible with our aim; the data mining task we decided to use was the classification task.

One of the key components of predictive accuracy is the amount and quality of the data (Burke et al., 1997).

We used the dataset contained in the SEER cancer incidence public-use database for the years 1973–2001. The SEER is the surveillance, epidemiology, and end results in data files that were requested through web site (http://www.seer.Cancer.gov). The SEER Program is part of the Surveillance Research Program at the National Cancer Institute and is responsible for collecting incidence and survival data from the participating 12 registries (Item Number 01 in SEER user file in the cancer web), and deploying these datasets (with the descriptive information of the data itself) to institutions and laboratories for the purpose of conducting analytical research (SEER cancer).

The SEER public use data contain nine text files, each containing data related to cancer for specific anatomical sites (i.e., breast, rectum, female genital, colon, lymphoma, other digestive, urinary, leukemia, respiratory, and all other sites). In each file, there are 93 variables (the original dataset before changing) which became 33 variables, and each record in the

file relates to a specific incidence of cancer. The data in the file are collected from 12 different registries (i.e., geographic areas). These registries consist of a population that is representative of the different racial/ethnic groups living in the United States. Each variable of the file contains 457,389 records (observations), but we are making some changes to the total of the variables adding some extra variables according to the variables requirements in the SEER file, for instance, the variable number 20 which is (extent of disease) contains (12-digits), the variable field description is denoted to SSSEELPNEXPE and we describe those letters to: SSS are the size of tumor, EE are the clinical extension of tumor, L is the lymph node involvement, PN are the number of positive nodes examined, EX are the number nodes examined, and PE are the pathological extensions for 1995+ prostate cases only. We have had some problems when we converted data into SAS datasets, but we recognized the problem which was with some names of the variables, for instance, the variable "Primary Site" and "Recode _ICD_O_I" are actually character variables: They, therefore, need to be read in using a "$" sign to indicate that the variable is text, we have also read in the variable "Extent_ of_Disease." There are two types of variables in the dataset which are categorical variables and continuous variables.

Afterward, we explored the data, preparation, and cleansing the dataset, the final dataset which contained 93 variables 92 predictor variables and the dependent variable.

The dependent variable is a binary categorical variable with two categories: 0 and 1, where 0 representing to did not survive and 1 representing to survive. The types of variables are as follows:

The categorical variables are: (1) Race (28 unique values), (2) marital status (6 values), (3) primary site code (9 values), (4) histology (123 values), (5) behavior (2 values), (6) sex (2 values), (7) grade (5 values), (8) extent of disease (36 values), (9) lymph node involvement (10 values), (10) radiation (10 values), (11) stage of cancer (5 values), and (12) site-specific surgery code (11 values).

While the continuous variable is: (1) Age, (2) tumor size, (3) number of positive nodes, (4) number of nodes, and (5) number of primaries.

The dataset is divided into two sets: Training set and testing set. The training set is used to construct the model, and the testing set is employed to determine the accuracy of the model built.

The position of the tumor in the breast may be described as the positions on a clock, as shown in (Fig. 1) (Coding Guidelines Breast, 2007; SEER Program Quality Control Section, 2007; SEER Program Code Manual, 1998).

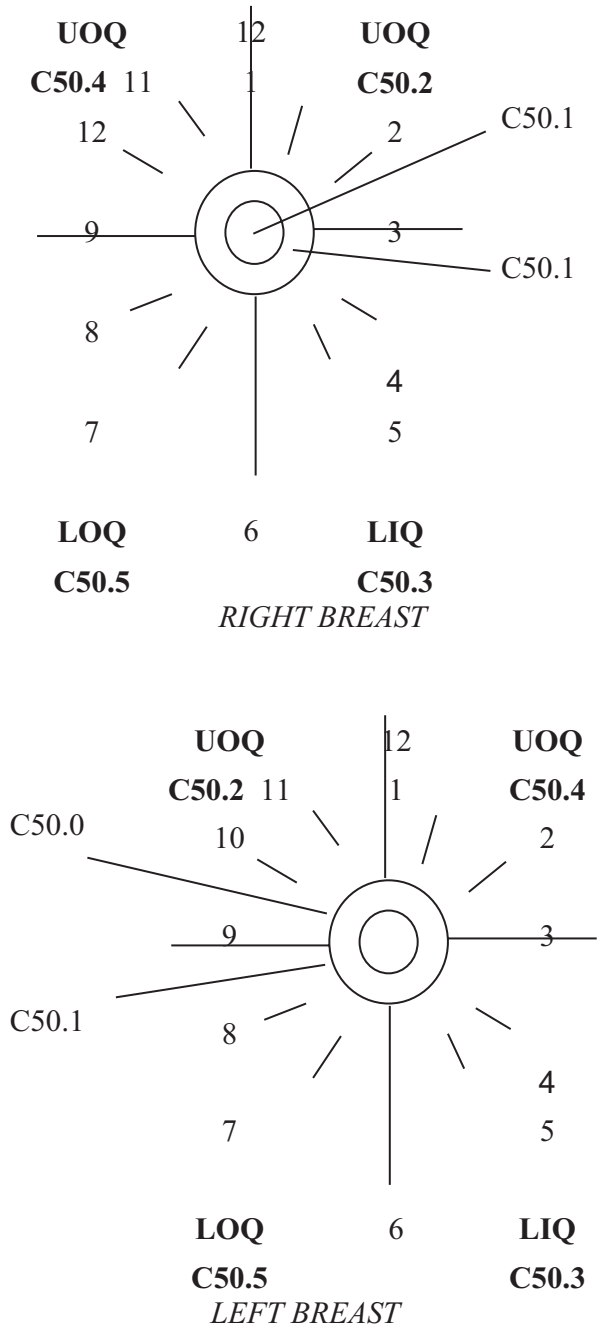

Fig. 1. O'clock positions and codes quadrant of breasts.

III. Overview

A. Background

We can explain here some research studies which carried out regarding the prediction of breast cancer survivability.

The first paper is "predicting breast cancer survivability: A comparison of three mining methods" (Delen et al., 2004). They have used three data mining techniques, which are decision tree (C5), ANNs, and logistic regression. They have used the data contained in the SEER cancer incidence public-use database for the years 1973–2000 and obtained the results using the raw data which were uploaded into the MS access database, SPSS statistical analysis tool, statistical data miner, and clementine data mining toolkit. These software packages were used to explore and manipulate the data. The following section describes the surface complexities and the structure of the data. The results indicated that the decision tree (C5) is the best predictor from which they found an accuracy of 93.6% and they found it to be better than the ANNs which had an accuracy of about 91.2%. The logistic regression

model was the worst of the three with 89.2% accuracy.

The models for the research study were based on the accuracy, sensitivity, and specificity and evaluated according to these measures. These results were achieved using ten-fold cross-validations for each model. They found according to the comparison between the three models that the decision tree (C5) performed the best of the three models evaluated and achieved a classification accuracy of 0.9362 with a sensitivity of 0.9602 and a specificity of 0.9066. The ANN model achieved accuracy 0.9121 with a sensitivity of 0.9437 and a specificity of 0.8748. The logistic regression model achieved a classification accuracy of 0.8920 with a sensitivity of 0.9017 and a specificity of 0.8786; the detailed prediction results of the validation datasets are presented in the form of confusion matrixes (Hosmer and Lemeshow, 1994).

The second research study was "predicting Breast Cancer survivability using data mining techniques" (Bellaachia and Guven, 2005). In this research, they have used data mining techniques: The Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms (Huang et al., 2003). The data source which they used was the SEER data (period of 1973–2000 with 433,272 records named as Breast.txt); they pre-classified into two groups of "survived" 93,273 and "not survived" 109,659 depending on the survived time records field. They have calculated the results using the Weka toolkit. The conclusion of the research study was based on calculations dependent on specificity and sensitivity. They also found that the decision tree (C4.5) was the best model with accuracy 0.0867, then the ANN with accuracy 0.865 and finally the Naïve Bayes with accuracy 0.0845. The analysis did not include records with missing data. This research study did not include the missing data, but our research does include the missing data, and this is one of the advances we made when comparing to the previous research.

The third research study was "ANN improve the accuracy of cancer survival prediction" (Burke et al., 1997). They have focused on the ANN and the tumor nodes metastasis (TNM) staging, and they used the same dataset SEER, but for new cases collected from 1977 to 1982. Based on this research study, the extent of disease variables for the SEER dataset was comparable to the TNM variables but not always identical to it. If considering accuracy, they found when the prognostic score is not related to survival and the score is 0.5, indicates a good chance for the accuracy, but if the score is from 0.5, that means this is better on average for the prediction model is at predicting which of the two patients will be alive.

The fourth research study was "prospects for clinical decision support in breast cancer based on neural network analysis of clinical survival data" (Kates et al., 2000). This research study used a dataset for patients with primary breast cancer who were enrolled between 1987 and 1991 in a prospective study at the Department of Obstetrics and Gynecology of the Technische University of Munchen, Germany. They have used two models (neural network and multivariate linear cox). According to the conclusion, the neural network in this dataset does not prove that the neural nets are always better than cox models, but the neural environment used here tests weights for significance, and removing too many weights usually reduces the neural representation to a linear model and removes any performance advantage over conventional linear statistical models.

Fig. 2 shows the survival rates of breast cancer among the states where the lowest rate is highlighted with red color, while, the highest rate where highlighted by yellow color.

B. Data Mining Techniques

What is data mining? Why use data mining

Nowadays, data mining is the process of extracting hidden knowledge from large volumes of raw data. Data mining is the main issue at the moment; the main problems these days are how we can to forecast any kind of data to find the best predictive result for predicative our information. Unfortunately, many studies fail to consider alternative forecasting techniques, the relevance of input variables, or the performance of the models when using different trading strategies.

The concept of data mining is often defined as the process of discovering patterns in larger databases that means the data are largely opportunistic, in the sense that it was not necessarily got for the purpose of statistical inference. A significant part of a data mining exercise is spent in an iterative cycle of data investigation, cleansing, aggregation, transformation, and modeling. Another implication is that models are often built on data with large numbers of observations and/or variables. Statistical methods must be able to execute the entire model formula on separately acquired data and sometimes in a separate environment, a process referred to as scoring. Data mining is the process of extracting knowledge hidden from large volumes of raw data. Powerful systems for collecting data and managing it in large databases are in place in all large and mid-range companies (Agilent Technologies, 2005). However, the bottleneck turning this data into valuable information is the difficulty of extracting knowledge about the system studied from the collected data. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support system or assessed by a human analyst (Witten and Frank, 2005). Fig. 3 shows data mining process model:

Data mining is a practical topic and involves learning in a practical, not theoretical; sense (Witten and Frank, 2005). Data mining involves the systematic analysis of large datasets

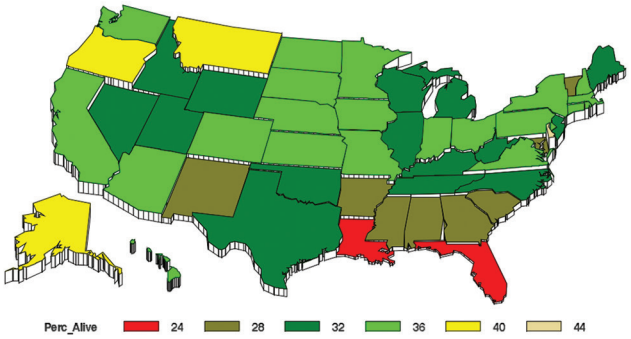

Fig. 2. Breast cancer survival rates by state.

using automated methods. By probing data in this manner, it is possible to prove or disprove existing hypotheses or ideas regarding data or information, while discovering new or previously unknown information. In particular, unique or valuable relationships between and within the data can be identified and used proactively to categorize or anticipate additional data (McCue, 2007). People always use data mining to get knowledge, not just predictions gaining knowledge from data certainly sounds like a good idea if we can do it.

Fig. 4 shows the graph is a 3-D vertical bar chart of 'Laterality', with a series variable of 'Grade', and a subgroup variable of 'Alive', and a frequency value, and shows the details of the values by clicking the arrow on the chart as showing in Fig. 5.

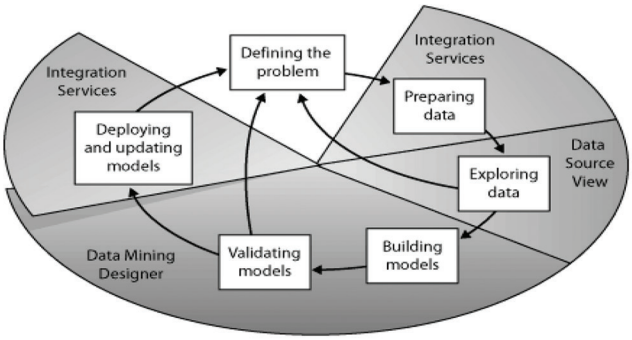

Fig. 3. Data mining process model.

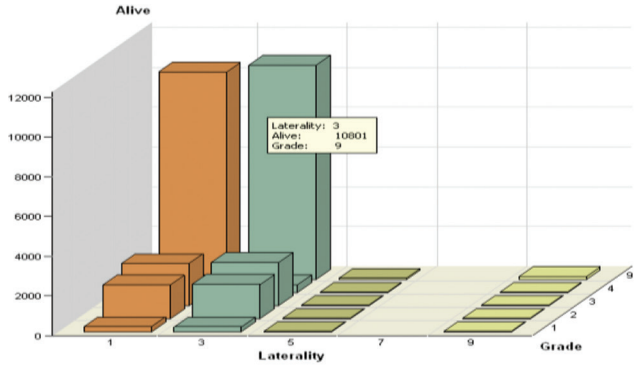

Fig. 4. The graph is a 3-D vertical bar chart of "Laterality," with a series variable of "Grade," and a subgroup variable of "Alive," and a frequency value, and shows the details of the values by clicking the arrow on the chart.

Classification

Classification is a key data mining technique whereby database tuples, acting as training samples, are analyzed to produce a model of the given data which we have used to predict group outcomes for dataset instances, and we used it to predict whether the patient will be alive or not alive as our project. It predicts categorical class labels classifies data (constructs a model) based on the training set and the values (class labels) in a classification attribute and uses it in classifying new data. The predictions are the models continuous-valued functions that means predicts unknown or missing values (Chen, 2007). In the classification, each list of values is supposed to belong to a predefined class which considered by one of the attributes, called the classifying attribute. Once derived, the classification model can be used to categorize future data samples and also to provide a better understanding of the database contents. The classification has numerous applications including credit approval, product marketing, and medical diagnosis (Allison's, 2003).

IV. Testing and Results

Table I. Some statistical information about the interval variables:

As we know the SAS enterprise miner doing all the necessary imputation and transformation to the dataset, then we do not want to be very worried about the data if it is not distributed normally as we said before (Aster, 2005).

Tables II-IV showing the important variables to the Alive (Target Variable):

The SEER historic Stage A Cramer's V is 0.29 which means the association between SEER historic Stage A and Alive is 0.29 which means there is a relationship between them, Clinical_Ext_of_Tumor_New and Alive is 0.28 and so on, but the association between Alive and First_malignant_prim_ind is almost non-existent because it is close to 0.

Form the basic analysis to the dataset, we see the important variable to the target varaiable (Alive) is the SEER historic Stage A (Stages 0, 1, 2, 4, or 8), for instance, if the stage

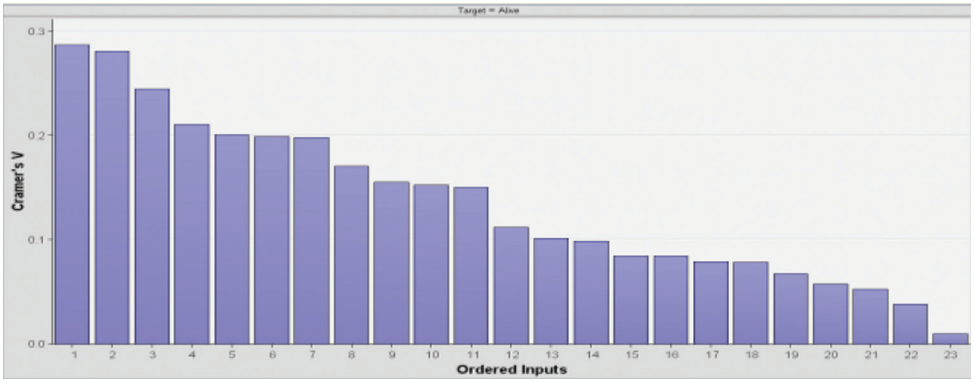

Fig. 5. Chi-square plot.

TABLE I
INTERVAL VARIABLES

Obs.	Name	Mean	Std. Dev.	Skewness	Kurtosis
1	Age_recodeless	12.67	2.909	-0.08295	-0.7114
2	Decade_at_Diagnosis	55.95	14.989	-0.00715	-0.5608
3	Decade_of_Birth	1919.47	16.077	0.13791	-0.369
4	Num_Nodes_Examined_New	11.8	16.768	3.45426	15.0212
5	Num_Pos_Nodes_New	40.2	45.521	0.45785	-1.7509
6	Number_of_primaries	1.21	0.464	2.22851	5.2614
7	Size_of_Tumor_New	92.4	230.732	3.61947	11.2935

TABLE II
CHI-SQUARE AND CRAMER'S V

Input	Cramer's V	Chi-square	DF	Ordered inputs	Plot	Group count
SEER_historic_stage_A	0.2872	7447.801	4	1	1	1
Clinical_Ext_of_Tumor_New	0.2808	2445.714	26	2	1	2
Site_specific_surgery_I	0.2445	5164.87	23	3	1	3
Reason_no_surgery	0.2106	4004.076	6	4	1	4
Tumor_Marker_I	0.2005	3631.661	5	5	1	5
Conversion_flag_I	0.1991	3581.374	5	6	1	6
Tumor_Marker_II	0.1982	3549.276	5	7	1	7
Sequence_number	0.1707	2630.806	6	8	1	8
Lymph_Node_Involvement_New	0.1551	617.9745	8	9	1	9
Grade	0.1525	2099.662	4	10	1	10
Histologic_Type_II	0.1502	2037.371	79	11	1	11
Diagnostic_confirmation	0.112	1132.812	7	12	1	12
Recode_I	0.1012	921.4222	17	13	1	13
Marital_status_at_diagnosis	0.0986	877.4522	5	14	1	14
PS_Number	0.0841	639.0324	8	15	1	15
Race_ethnicity	0.0841	638.2951	23	16	1	16
Radiation	0.0791	564.559	9	17	1	17
Birthplace	0.0784	555.4019	198	18	1	18
ICD_ Number	0.0675	411.894	5	19	1	19
Laterality	0.0576	300.0729	4	20	1	20
Behavior recode for analysis	0.0526	250.0972	1	21	0	21
Radiation_sequence_with_surgery	0.0385	133.5344	5	22	0	22
First_malignant_prim_ind	0.0097	8.478975	1	23	0	23

is 1 that means the localized stage of an invasive neoplasm confined entirely to the organ of origin.

A. The Artificial Neural Network

From the results, (Figs. 6 and 7) displays the iteration plot with an average squared error at each iteration for the training and validation datasets. The estimation process required 100 iterations. The weights from the iteration were selected. Around iteration, the average squared error flattened out in the validation (the red line) dataset, although it continued to drop in the training dataset (the green line).

As we knew, the objective function is the average error. The best model is the model that gives the smallest average error for the validation data. Table V shows some statistics label, both targets are range normalized. Values are between

TABLE III
CLASS VARIABLE SUMMARY STATISTICS

Variable	Number of unique values
Behavior_recode_for_Analysis	2
Birthplace	199
Clinical_Ext_of_Tumor_New	28
Conversion_flag_I	6
Diagnostic_confirmation	8
First_malignant_prim_ind	2
Grade	5
Histologic_Type_II	80
ICD_Number	6
Laterality	5
Lymph_Node_Involvement_New	10
Marital_status_at_diagnosis	6
PS_Number	9
Race_ethnicity	24
Radiation	10
Radiation_sequence_with_surgery	6
Reason_no_surgery	7
Recode_I	19
SEER_historic_stage_A	5
Sequence_number	7
Site_specific_surgery_I	25
Tumor_Marker_I	6
Tumor_Marker_II	6
Alive	2

TABLE IV
INTERVAL VARIABLE SUMMARY STATISTICS

Variable	Mean	Std. Dev.	Min.	Median	Max.
Age_recodeless	12.67	2.909	4	13	18
Decade_at_Diagnosis	55.95	14.989	10	60	100
Decade_of_Birth	919.47	16.077	1870	1920	1970
Num_Nodes_Examined_New	11.8	16.768	0	10	98
Num_Pos_Nodes_New	40.2	45.521	0	9	98
Number_of_primaries	1.21	0.464	1	1	6
Size_of_Tumor_New	92.4	230.732	0	30	998

0 and 1. The root mean square error for Target 1 is about 43.5%, mean square error is 18.9%. The following table shows that:

B. Decision Trees

The decision trees technique repetition separated observations in branches to make a tree for the purpose of evolving the prediction accuracy. Using mathematical algorithms (Gini index, information gain, and Chi-square test), to identify a variable and corresponding threshold for the variable that divides the input values into two or more subgroups. This step is repetition at each leaf node until the complete tree is created (Neville, 1999).

The aim of the dividing algorithm is to identify a variable-threshold pair that maximizes the homogeneity of the two results or more subgroups of samples. The most mathematical algorithm used for splitting contains entropy-based information gain (used in C4.5, ID3, and C5), Gini index (used in classification and regression tree), and the Chi-squared test (used in Chi-square Automatic Interaction Detector).

Fig. 6. Iteration plot with an average squared error.

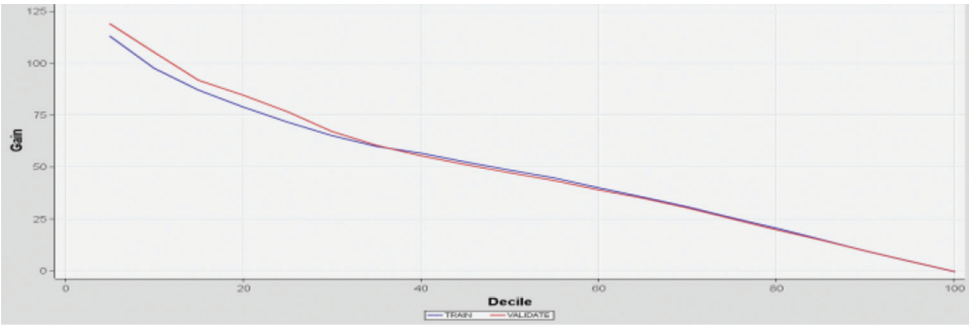

Fig. 7. Score rankings overlay Alive (gain chart).

TABLE V
Fitted Statistics

Target	Fit statistics	Statistics label	Train	Validation	Test
Alive	_DFT_	Total degrees of freedom	30167	0	0
Alive	_DFE_	Degrees of freedom for error	29831	0	0
Alive	_DFM_	Model degrees of freedom	336	0	0
Alive	_NW_	Number of estimated weights	336	0	0
Alive	_AIC_	Akaike's information criterion	33753.85	0	0
Alive	_SBC_	Schwarz's Bayesian criterion	36547.52	0	0
Alive	_ASE_	Average squared error	0.187201	0.1868483	0.187468
Alive	_MAX_	Maximum absolute error	0.987512	0.99525055	0.990725
Alive	_DIV_	Divisor for ASE	60334	45190	45048
Alive	_NOBS_	Sum of frequencies	0.191418	NaN	22524
Alive	_RASE_	Root average squared error	0.18931	0.1868483	0.432976
Alive	_SSE_	Sum of squared errors	0.437513	NaN	8445.057
Alive	_SUMW_	Sum of case weights times freq.	0.435097	0.43225953	45048
Alive	_FPE_	Final prediction error	0.548312	0.54876668	NaN
Alive	_MSE_	Mean squared error	33081.85	24798.7662	0.187468
Alive	_RFPE_	Root final prediction error	0.30066	0.29161319	NaN
Alive	_RMSE_	Root mean squared error	9070	6589	0.432976
Alive	_AVERR_	Average error function	0.18931	0.1868483	0.550722
Alive	_ERR_	Error function	0.437513	NaN	24808.94
Alive	_MISC_	Misclassification rate	0.435097	0.43225953	0.293598
Alive	WRONG	Number of wrong classifications	0.548312	0.54876668	6613

We have used the entropy technique and summarized the results according to the most common variables to choose the most and important predictor variables. In appendix (4), the decision tree property criterion is entropy, one of the results examples is as follows:

If Site_specific_surgery_I = 09 and SEER_historic_stage_A = 4 and Lymph_Node_Involvement_New = 0 and Clinical_Ext_of_Tumor_New = 0 then node: 140, N (number of values in the node): 1518, not survived (0): 94.8%, survived (1): 5.2%, or if the decision tree property criterion is Gini, one of the example is; IF Site_specific_surgery_I = 90 and SEER_historic_stage_A = 4 AND Lymph_Node_Involvement_New = 0 and Clinical_Ext_of_Tumor_New = 0 then node: 130, N: 1272, survived: 85.4% and not survived: 14.6%. and finally if the Decision tree properity criterion is ProbChisq, one of the exaplme is; Grade is one of: 9 or 2 and Sequence_number is one of: 00, 02, or 03 and Reason_no_surgery is one of: 0 or 8 and SEER_historic_stage_A = 4 then node: 76,

for the number of the values is 2310, survived is 86.3%, and not survived is 13.7%.

The most important variables participate for the largest numbers of the observations to the target variable if used Entopy are:

Clinical_Ext_of_Tumor_New, Site_specific_surgery_I, Histologic_Type_I, Size_of_Tumor_New, Grade, Lymph_Node_Involvement_New, Sequence_number, SEER_historic_stage_A, Age_recodeless, Conversion_flag_I, Decade_of_Birth and Age_recodeless.

We can say the most important variables to the target variables are: Grade, Size of Tumor New, SEER historic stage A, Clinical Ext of Tumor New Lymph Node Involvement New, Histologic Type II, Sequence number Age recodeless, Decade of Birth and Conversion flag I.

Table VI view displays a list of variables in the order of their importance in the tree.

These results from the (autonomous decision tree) icon when we used the interactive property, the table shows that the prognosis factor "SEER historic stage A" is by far the most important predictor, which is not consistent with the previous research, the previous research was the prognosis factor "Grade" the most important predictor and "Stage of cancer" secondly! But from our table, we see the second most important factor is "clinical extension of tumor new," then "Decade (Age) at diagnosis" and "Grade." However, we noticed that the size of the tumor in the eighth in the standings.

C. The Logistic Regression

Firstly, let us start with the logistic regression figure as shown in Fig. 8:

The intercept and the parameters in the regression model shows that bar number one represents the intercept with value (–1.520597), bar 2, the value of the parameter which represent the variable (SEER historic stage A) with value (–1.378877), the second bar is and so on.

The Table VII shows the regression model explanation, and it's very clear in this model as the variable (SEER historic stage A) one of the most important variables to the target variable, the intercept of Alive = 1 is equal to −1.5206 which means the amount of change for the target variable (Alive = 1), the coefficient of the variable (SEER historic stage A) is −1.38 which means the amount of change in this variable on the Alive by −1.38; also, the t-test is to calculate the significance of the independent variable with the target variable, t = −28.66 means (SEER historic stage A = value 4) is insignificant because if we compare it with level of statistical significance equal to −0.05 > −28.66 that means to reject the null hypothesis and accept the alternative hypothesis instead, and this depends to the hypothesis that we want to test it, might be we want to use this hypothesis:

H_0: μ=0 against H_1: $\mu\neq0$ or H_0: $\sigma\neq1$ against H_1: $\sigma\neq1$.

However, this difference if we choose another value of SEER historic stage A = value 0 because the t value = +9.31, at this stage, the variable is significant to the target variable.

TABLE VI
THE MOST IMPORTANT VARIABLES BY USING ENTROPY CRITERION

Variable	Nodes	Training	Validation
SEER_historic_A	1	1	1
Clinical_Ext_of_turnor_New	1	0.938	0.929
Decade_at_Diagnosis	12	0.477	0.389
Grade	24	0.47	0.399
Sequence_number	5	0.41	0.409
Histologic_Type_II	14	0.381	0.261
Num_Pos_Nodes_New	15	0.309	0.183
Size_of_Turnor_New	20	0.303	0.151
Site_specific_Surgery_I	11	0.296	0.15
Raeson_no_surgery	3	0.295	0.309
PS_Number	36	0.286	0.106
Num_Nodes_Examined_New	11	0.244	0.129
Birthplace	5	0.241	0.028
Radiation	14	0.189	0.088
Turnor_Marker_I	2	0.183	0.183
Conversion_flag_I	2	0.168	0.168
Laterality	28	0.141	0.091
Number_of_primaries	9	0.139	0.137
Turnor_Marker_II	2	0.1	0.014
Lymph_Node_Involvement_New	4	0.094	0
Recode_I	3	0.089	0
Decade_of_Birth	3	0.083	0.038
Age_recodeless	3	0.073	0.007
Marital_status_at_diagnosis	1	0.062	0
Diagnostic_confirmation	2	0.043	0.012
Behaviour_recode_for_Analysis	0	0	0
ICD_Number	0	0	0
Race_ethincity	0	0	0
SEEr_modified_ICCC_Site_recode	0	0	0
Scheme	0	0	0
Radiation_sequence_with_surgery	0	0	0
First_malignant_prim_ind	0	0	0

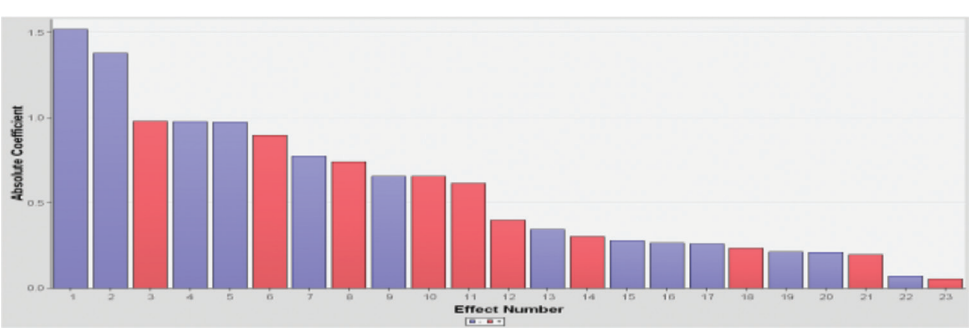

Fig. 8. Bar charts for logistic regression.

10.24086/cuejhss.vol4n1y2020.pp17-30

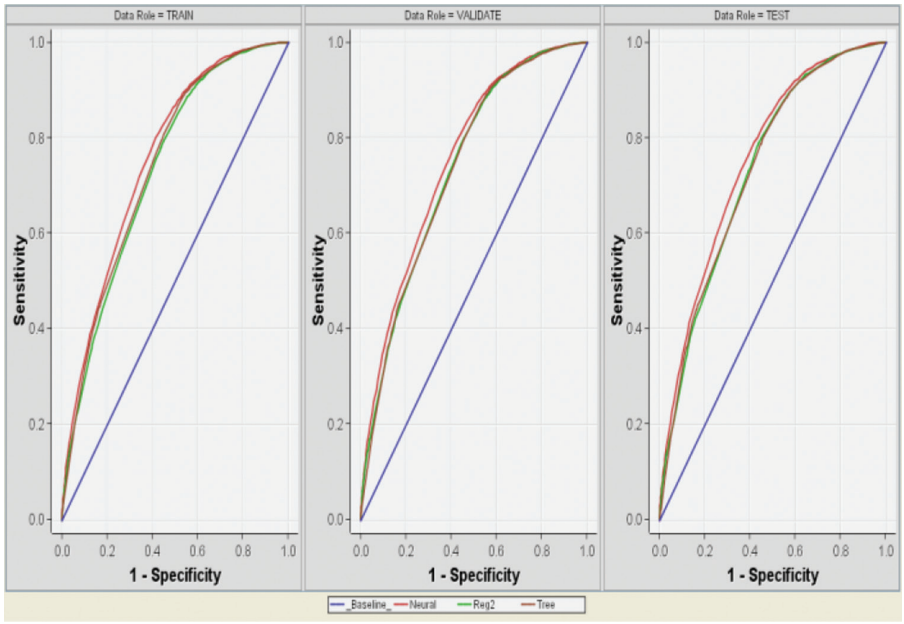

Fig. 9. Model comparison chart.

TABLE VII
Regression Most Important Variables

Variable	Level	Effect	Effect label
Intercept	1	Intercept	Intercept:Alive=1
SEER_historic_stage_A	4	SEER_historic_stage_A4	SEER_historic_stage_A 4
IMP_Site_specific_surgery_I	2	IMP_Site_specific_surgery_I02	Imputed Site_specific
IMP_Site_specific_surgery_I	0	IMP_Site_specific_surgery_I00	Imputed Site_specific _surgery_I 00
IMP_Site_specific_surgery_I	9	IMP_Site_specific_surgery_I09	Imputed Site_specific _surgery_I 09
Tumor_Marker_I	2	Tumor_Marker_I2	Tumor_Marker_I 2
Grade	3	Grade3	Grade 3
Tumor_Marker_I	8	Tumor_Marker_I8	Tumor_Marker_I 8
Sequence_number	0	Sequence_number00	Sequence_number 00
Grade	4	Grade4	Grade 4
Tumor_Marker_I	0	Tumor_Marker_I0	Tumor_Marker_I 0
IMP_Site_specific_surgery_I	40	IMP_Site_specific_surgery_I40	Imputed Site_specific_surgery_I 40
SEER_historic_stage_A	2	SEER_historic_stage_A2	SEER_historic_stage_A 2
IMP_Site_specific_surgery_I	58	IMP_Site_specific_surgery_I58	Imputed Site_specific_surgery_I 58
SEER_historic_stage_A	0	SEER_historic_stage_A0	SEER_historic_stage_A 0
IMP_Site_specific_surgery_I	20	IMP_Site_specific_surgery_I20	Imputed Site_specific_surgery_I 20

TABLE VIII
Event Classification

Obs.	Model	FN	TN	FP	TP
1	Step.Reg TRAI	5867	16131	3224	4945
2	Step.Reg VALI	4368	12174	2470	3583
3	Back.Reg TRAI	6624	16490	2865	4188
4	Back.Reg VALI	4815	12564	2080	3136
5	Forw.Reg TRAI	6624	16490	2865	4188
6	Forw.Reg VALI	4815	12564	2080	3136
7	Neural TR	6124	16409	2946	4688
8	Neural VA	4375	12430	2214	3576
9	Tree TRAI	7469	20477	3270	4907
10	Tree VALI	5527	15491	2485	3589

D. Model Comparison using SAS

The model comparison node belongs to the assessment category in the SAS data mining process of the sample, explore, modify, model, and assess. The model comparison

TABLE IX
Confusion Matrix

Obs.	Model	FN	TN	FP	TP	Accuracy	Sensitivity	Specificity
1	Step.Reg TRAI	5867	16131	3224	4945	0.69864	0.45736	0.83343
2	Step.Reg VALI	4368	12174	2470	3583	0.69737	0.45064	0.83133
3	Back.Reg TRAI	6624	16490	2865	4188	0.68545	0.38735	0.85198
4	Back.Reg VALI	4815	12564	2080	3136	0.69484	0.39442	0.85796
5	Forw.Reg TRAI	6624	16490	2865	4188	0.68545	0.38735	0.85198
6	Forw.Reg VALI	4815	12564	2080	3136	0.69484	0.39442	0.85796
7	Neural TR	6124	16409	2946	4688	0.69934	0.43359	0.84779
8	Neural VA	4375	12430	2214	3576	0.70839	0.44975	0.84881
9	Tree TRAI	7469	20477	3270	4907	0.70271	0.39649	0.8623
10	Tree VALI	5527	15491	2485	3589	0.70427	0.3937	0.86176

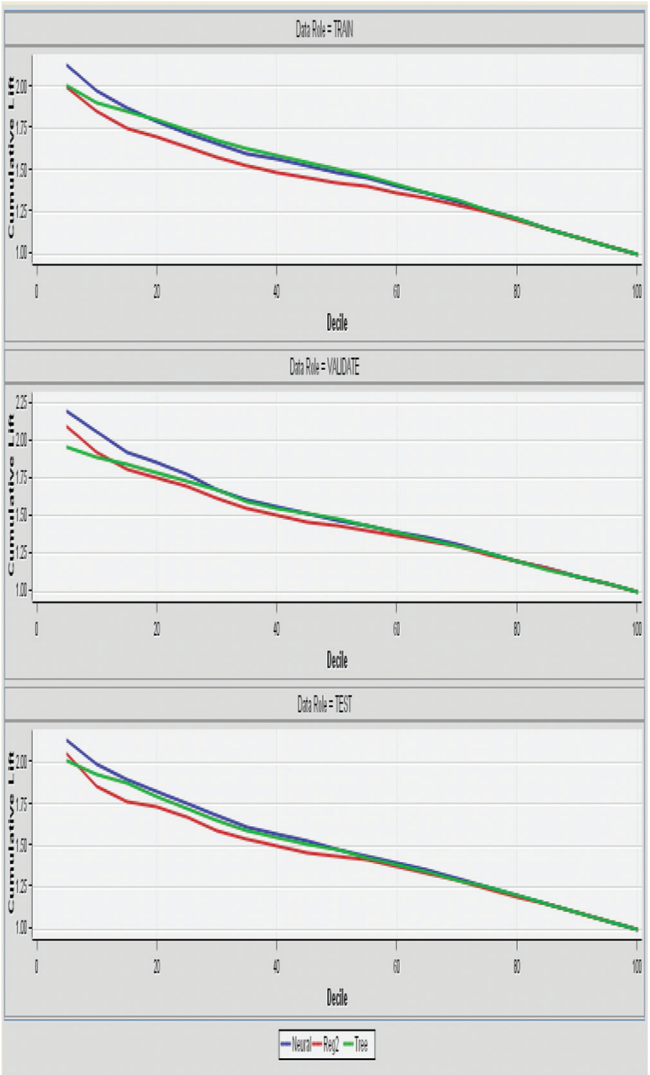

Fig. 10. Score rankings overlay: Alive (cumulative lift).

node enables us to compare models and predictions from the modeling nodes using various criteria (Han and Kamber, 2001).

A common criterion for all modeling and predictive tools is a comparison of the expected survival or not survival to actual survival or not survival getting data from model results.

The criterion enables us to make cross-model comparisons and assessments, independent of all other factors (such as sample size, modeling node, and so on).

When we train a modeling node, assessment statistics are computed on the train (and validation) data. The model comparison node calculates the same statistics for the test set when present. The node can also be used to modify the number of deciles and/or bins and recomputed assessment statistics used in the score ranking and score distribution charts for the train (and validation) dataset (Intrator and Intrator, 2001).

In addition, it computes for binary targets the Gini, Kolmogorov–Smirnov, and Bin-Best Two-Way Kolmogorov–Smirnov statistics and generates receiver operating characteristic charts for all models using the train (validation and test) datasets.

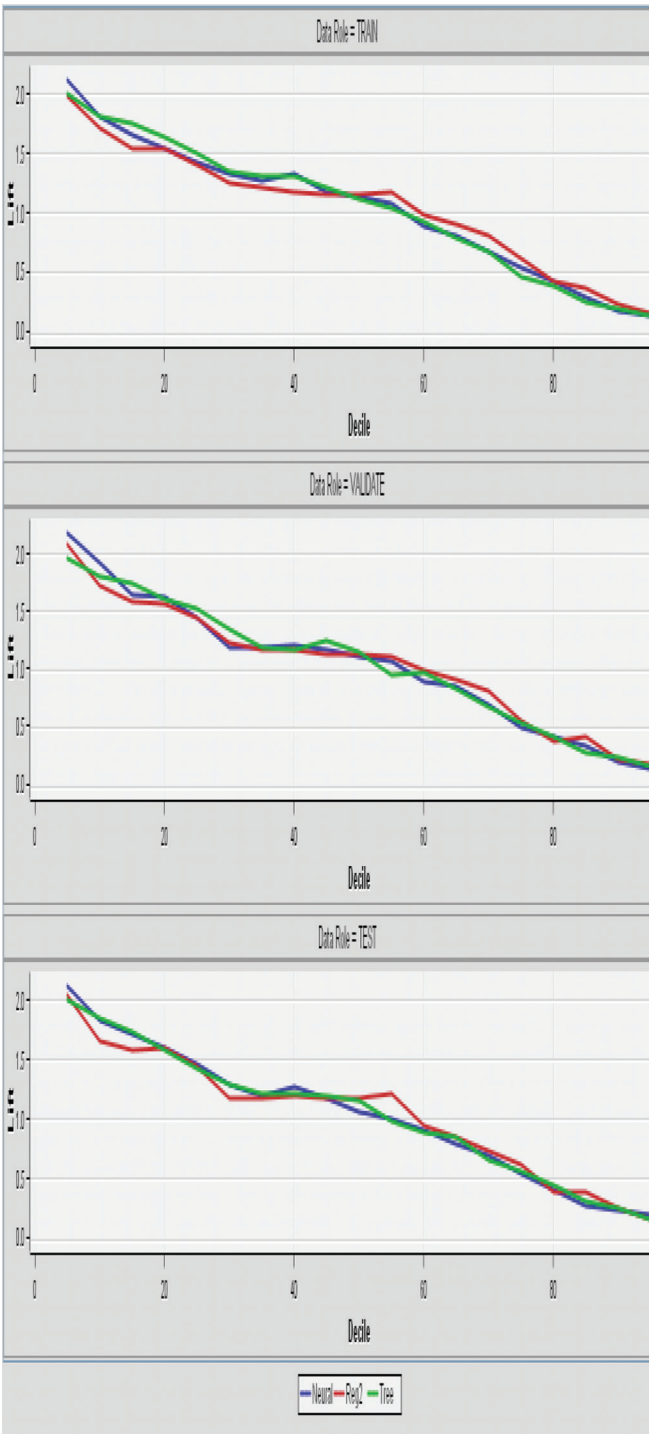

Fig. 11. Score rankings overlay: Alive (lift).

We have used the program to run the results of the accuracy, sensitivity, and specificity, between the neural network, the decision trees, and the logistic regression (stepwise, backward, and forward). The steps we will have to run, 1. We must run the model comparison to get the event classification table as shown in Table VIII.

And then, we put the results table in the program number (10) using SAS code to get the confusion matrix. Table IX shows the results of the event classification and the confusion matrix.

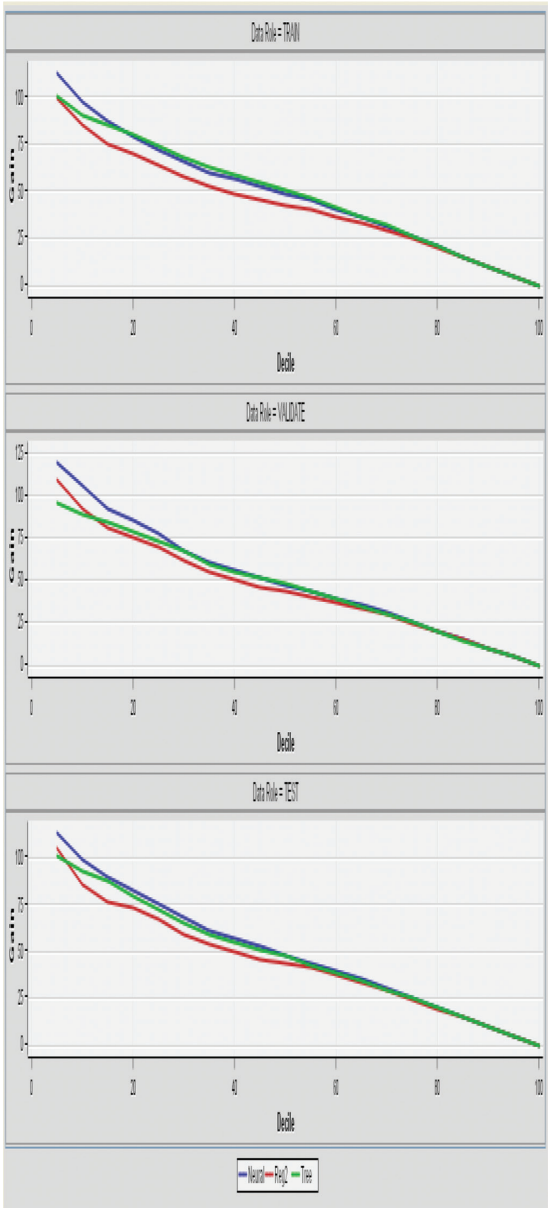

Fig. 12. Score rankings overlay: Alive (gain).

Table IX shows that the neural network model is the best model because the accuracy of the model is 0.70839 and the error rate is: 1−0.70839 = 0.29161, for sensitivity is 0.44975, and for specificity is 0.84881, these are for the validation data, and all the values for this model are bigger than the other models. The second important model is the decision tree with accuracy of 0.70427 with error rate 0.29573, sensitivity is 0.3937, and for specificity is 0.86176 and the third important model is the logistic regression (step-wise regression) with accuracy of 0.69737 with error rate 0.30263, for sensitivity is 0.45064 and for specificity is 0.83133; these results are for the validation, and so on for the backward and forward regression (Figs. 9-12).

The following table shows the results of the k-fold cross-validation:

The accuracy of the measurement model and calculated the average number of 10 times of performance. We repeated this process for each of the three prediction models. Provided us with the least bias to predict as shown in Table X the performance measures compared to the tree models. We removed two of them because of unreasonable results (The Basics of SAS Enterprise Miner, 2018).

E. Future Work

When we want to talk about future research related to our current dissertation, there are a lot of ideas and work to do in the future, one of these ideas is whether there is a relationship between breast cancer and other tumor diseases in terms of survival or response to the treatment. Using other data mining models, we could see if the new model is appropriate or not to other models. The previous models did not use the SAS system to analyses the dataset and I think SAS software has many more facilities than the other software; as a result, more useful information and results are obtained which are more efficient than the other packages.

We are thinking to do more work relate to cancer disease, because we should all be helping serve the public interest, especially when concerning Cancer. We have a lot of ideas to do more research and analysis of the data in more sectors

TABLE X
K-fold Cross-validation Results

First fold								
Obs.	Model	FN	TN	FP	TP	Accuracy	Sensitivity	Specificity
1	Tree TRAI	6569	18498	3007	4437	0.70545	0.40314	0.86017
2	Tree VALI	5084	13926	2247	3126	0.69934	0.38076	0.86106
3	Neural TR	4988	13875	2565	4268	0.70606	0.46111	0.84398
4	Neural VA	3845	10508	1919	3012	0.7011	0.43926	0.84558
5	Step.Reg.TRAI	5374	13777	2663	3882	0.68723	0.4194	0.83802
6	Step.Reg.VALI	4016	10383	2044	2841	0.68575	0.41432	0.83552
Second fold								
1	Tree4 TRA	7127	18818	2690	3876	0.69804	0.35227	0.87493
2	Tree4 VAL	5327	14030	2085	2941	0.69602	0.35571	0.87062
3	Neural4 TR	5393	14242	2646	4024	0.69439	0.42731	0.84332
4	Neural4 VA	4078	10469	2029	3057	0.68894	0.42845	0.83765
5	Step.Reg.TRAI	6230	14742	2146	3187	0.68158	0.33843	0.87293
6	Step.Reg.VALI	4698	10876	1622	2437	0.67809	0.34156	0.87022

(*Contd...*)

TABLE X
(*Countiued*)

First fold								
Third fold								
1	Neural6 TR	5316	14482	2593	4158	0.7021	0.43889	0.84814
2	Neural6 VA	4009	10638	1961	3193	0.6985	0.44335	0.84435
3	Step.Reg.TRAI	5886	14709	2366	3588	0.68918	0.37872	0.86143
4	Step.Reg.VALI	4397	10823	1776	2805	0.68825	0.38948	0.85904
5	Tree6 TRA	7305	18919	2615	3672	0.69487	0.33452	0.87856
6	Tree6 VAL	5464	14025	2020	2874	0.69306	0.34469	0.8741
Fourth fold								
1	Neural5 TR	4994	13853	2675	4197	0.70182	0.45664	0.83815
2	Neural5 VA	3788	10184	2038	3289	0.69812	0.46474	0.83325
3	Step.Reg.TRAI	5465	14016	2512	3726	0.68984	0.4054	0.84802
4	Step.Reg.VALI	4214	10402	1820	2863	0.68734	0.40455	0.85109
5	Tree5 TRA	6920	18596	2960	4035	0.6961	0.36832	0.86268
6	Tree5 VAL	5315	13775	2213	3080	0.69126	0.36689	0.86158
Fifth fold								
1	Neural7 TR	5403	14569	2586	4197	0.7014	0.43719	0.84926
2	Neural7 VA	4215	10854	1936	3065	0.69352	0.42102	0.84863
3	Step.Reg.TRAI	5878	14800	2355	3722	0.69228	0.38771	0.86272
4	Step.Reg.VALI	4515	10995	1795	2765	0.6856	0.37981	0.85966
5	Tree7 TRA	6463	17930	3536	4582	0.69244	0.41485	0.83527
6	Tree7 VAL	4916	13368	2673	3426	0.68876	0.41069	0.83336
Sixth fold								
1	Neural8 TR	5392	14367	2728	4222	0.69598	0.43915	0.84042
2	Neural8 VA	4031	10855	2004	3189	0.69944	0.44169	0.84416
3	Step.Reg.TRAI	5939	14630	2465	3675	0.68535	0.38226	0.85581
4	Step.Reg.VALI	4387	11097	1762	2833	0.69376	0.39238	0.86298
5	Tree8 TRA	7030	18598	2816	4067	0.69715	0.3665	0.8685
6	Tree8 VAL	5277	13914	2141	3051	0.69577	0.36635	0.86665
Seventh fold								
1	Neural9 TR	5093	14090	2462	4115	0.70672	0.44689	0.85126
2	Neural9 VA	3918	10560	1932	3005	0.69869	0.43406	0.84534
3	Step.Reg.TRAI	5557	14251	2301	3651	0.69495	0.3965	0.86098
4	Step.Reg.VALI	4211	10673	1819	2712	0.68942	0.39174	0.85439
5	Tree9 TRA	6373	18330	3236	4572	0.70444	0.41772	0.84995
6	Tree9 VAL	4799	13582	2562	3440	0.69811	0.41753	0.8413
Eighth fold								
1	Neural10TR	5193	14008	2669	4189	0.6983	0.44649	0.83996
2	Neural10VA	3915	10474	2055	3145	0.69524	0.44547	0.83598
3	Step.Reg.TRA	5728	14327	2350	3654	0.69001	0.38947	0.85909
4	Step.Reg.VAL	4309	10738	1791	2751	0.6886	0.38966	0.85705
5	Tree10 TR	7099	18723	2730	3959	0.69767	0.35802	0.87275
6	Tree10 VA	5359	13985	2153	2886	0.69192	0.35003	0.86659

such as financial analyses, population analysis, and health analysis ….

V. Conclusion

This research study emphasized on a dissertation effort where we developed three main prediction models for breast cancer survivability. Specifically, we used three popular data mining methods: ANN, decision trees, and logistic regression. We obtained a full and large dataset (457,389 cases with 93 prognosis factors) from the SEER program and after going through a long process of data cleansing, aggregation, transformation, and modeling by SAS, we used it to develop the prediction models. In this research, we have identified cases of breast cancer survival when a person is still alive after 5 years (60 months) from the date of diagnosis. We used a binary categorical survival variable, which was computed from the variables in the raw dataset, to assimilate the survivability where survival is represented with a value of "1" and non-survival is represented with "0." The assembly results indicated that the ANN performed the best with a classification accuracy of 70.8%, the decision tree induction method model (with multi-layered perceptron architecture) came out to be second best with a classification accuracy of 70.4%, and the logistic regression model came out to be the worst with a classification accuracy of 69.5%.

From all the models results, the common thing between the models is that some important factors are the same effectiveness to the target variable, for instance, the prognosis factor "SEER historic Stage A" is by far the most common

important predictor, which is not consistent with the previous research, the previous research was the prognosis factor "Grade" the most important predictor and "Stage of cancer" secondly! But from our research, the second most important factor is "clinical extension of tumor new," then "decade (age) at diagnosis" and "Grade." However, we noticed that the size of the tumor has ranked eighth in the overall standings.

It will be possible to extend this research in the future and to do further research. In addition to, the most useful future results can be listed as follows: First, in the study of breast cancer survivability, we have not considered the potential relation (correlation) to other tumor sorts. It would be an interesting study to scrutinize if there is a specific cancer which has a worse survivability rating. This can be done by including all possible cancer types and their prognostic factors to investigate the correlations, commonalities and differences among them. Second, new methods as an example to support vector machines and rough sets can be used to find out if the prediction accuracy can be further improved. Another applicable option to improve the prediction accuracy would be shown that the gathering mean-square error of forecasts constructed from a particular linear combination of independent and incompletely correlated predictions is less than that of any of the individual predictions. The weights to be attached to each prediction are determined by the Gaussian method of least squares and depend on the covariance between independent predictions and between prediction and verification.

In terms of predicting accuracy in the measurement of non-biased of the three methods, we repeated this process for k (10) times so that each data point that will be used in the training and test data. We repeated this process for each of the three prediction models. This provided us with the least bias to predict performance measures compared to the tree models. If we see the Table X, the best model for most of the k-folds cross-validation is the ANN, then the decision trees and the worst is the logistic regression. The prognosis factor "SEER historic Stage A" is by far the most important predictor, which is consistent with the previous research, followed by "size of tumor," "Grade," and "lymph node involvement new."

Why these prognostic factors are more important predictors than the other is a question that can only be answered by the medical clinician and their work from further clinical studies.

We asked some specialist clinicians specializing in breast cancer and they made the following comments:

Dr. Rebecca Roylance, a Senior Lecturer and Honorary Consultant who is based at the Barts and the London (NHS Trust), comments about the most important prognosis factors:

1. Size of tumor (bigger size worse)
2. Grade of tumor, there are 3 Grades, I, II, and III, and Grade III being the worst
3. Receptor status, i.e., ER, PR, and HER2, +ve ER and PR better than ER/PR- HER2 + being the worst
4. Amount of lymph node involvement
5. Age of pt – younger worse
6. Presence of lymph vascular invasion and 5, 6 both play a role but are less important than the other predictor factors.

Increasing the accuracy of the model, for instance, increasing the accuracy of neural network classification using filtered training data, the accuracy performed by a supervised classification is to a large extent dependent on the training data provided by the analyst. The training datasets represent significant importance for the performance of all classification methods. However, this situation is more important for neural network classifiers from them to take each sample into consideration in the training stage. As we said in the neural network results, we can change the number of iterations that we want to allow during network training to give us the highest accuracy. The representation is related to the quality and size of the training data that they are very important in evaluating the accuracy. Quality analysis of training data helps to identify outlier and extreme values that can undermine the fineness and accuracy of a classification resulting from not true class limits definition. Training data selection can be thought of as a repetition process to form a representative dataset after some improvements. Unfortunately, in many applications, the quality of the training data is not required, and the dataset is directly used in the training step. With a view to increase the representativeness of the training data, a two-stage approach is applied, and completion tests are assumed for a selected region. Results show that the use of representative training data can help the classifier to make more accurate and effective results. An amendment of several percents in classification accuracy can significantly improve the reliability of the quality of the classified image.

References

Agilent Technologies, Inc. (2005). *Principal Component Analysis*. Retrieved from: http://www.chem.agilent.com/cag/bsp/products/gsgx/downloads/pdf/pca.pdf. [Last accessed on 2019 Feb 15].

Allison, P. D. (2001). *Logistic Regression Using the SAS System: Theory and Application*. SAS Publishing. Retrieved from: http://www.books.google.co.uk/books. [Last accessed on 2018 Oct 16].

Allison's, R. (2003). *SAS/Graph Examples*. Retrieved from: http://www.robslink.com. [Last accessed on 2018 Oct 07].

Aster, R. (2005). *Professional SAS Programming Shortcuts*. Retrieved from: http://www.globalstatements.com/shortcuts. [Last accessed on 2018 Nov 01].

Bellaachia, A., & Guven, E. (2005). *Predicting Breast Cancer Survivability Using Data Mining Techniques, Department of Computer Science*. Washington DC: The George Washington University.

Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E., Weinstein, J. N., Harrell, F. E Jr., Marks, J. R., Winchester, D. P., & Bostwick, D. G. (1997). Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer*, 79, 857-862. Retrieved from: http://www.info.cancerresearchuk.org/cancerstats/types/breast/incidence. [Last accessed on 2018 Sep 15].

Chen, D. (2007). *Decision Trees for Classification, in Lecture Notes in Dept of Info Systems and IT, PhD. Faculty of Business*, *Computing and Info Management*. London: South Bank University.

Chow, M., Goode, P., Menozzi, A., Teeter, J., & Thrower, J. P. (1994). *Bernoulli Error Measure Approach to Train Feed forward Artificial Neural Networks for Classification Problems, Department of Electrical and Computer Engineerin*g. Raleigh, USA: North Carolina State University.

Coding Guidelines Breast C500-C509. (2007). *SEER Program Coding and Staging Manual 2007, Coding Guidelines Breast C500-C509*. Retrieved from:

http://www.seeer.Cancer.gov. [Last accessed on 2018 Oct 14].

Delen, D., Walker, G., & Kadam, A. (2004). *Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods*. Retrieved from: http://www.journals.elsevierhealth.com. [Last accessed on 2019 Aug 01].

Edwards, B. K., Howe, H. L., Lynn, A. G. R., Thun, M. J., Rosenberg, H. M., Yancik, R., Wingo, P. A., Jemal, A., & Feigal, E. G. (2002). Annual report to the nation on the status of Cancer, 1973-1999, featuring implications of age and aging on US Cancer burden. *Cancer*, 94, 2766-2792.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*. Burlington: Morgan Kaufmann Publisher.

Holland, S. (2008). *Principal Component Analysis*. Retrieved from: http://www.uga.edu/~strata/software/pdf/pcaTutorial.pdf. [Last accessed on 2018 Dec 24].

Hosmer, W. D., & Lemeshow, S. (1994). *Applied Logistic Regression. Wiley Series in Probability and Statistics Applied Probability and Statistics Sectio*n. Retrieved from: http://www.books.google.co.uk/books. [Last accessed on 2018 Oct 20].

Huang, J., Lu, J., & Ling, C. X. (2003). *Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy*. (pp. 553-556). 3rd IEEE International Conference on 19-22. Retrieved from: https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=783455. [Last accessed on 2018 Nov 03].

Intrator, O., & Intrator, N. (2001). Interpreting neural-network results: A simulation study. *Computational Statistics and Data Analysis*, 37(3), 373-393.

Kates, R., Harbeck, N., & Schmitt, M. (2000). *Prospects for Clinical Decision Support in Breast Cancer Based on Neural Network Analysis of Clinical Survival Data*. Munich, Germany: IEEE.

McCue, C. (2007). *Data Mining and Predictive Analysis* (*Intelligence Gathering and Crime Analysis*). Oxford: Elsevier Inc.

Neville, P. (1999). *Decision Trees for Predictive Modelling*. Retrieved from http://www.sasenterpriseminer.com/documents/Decision%20Trees%20for%20Predictive%20Modeling.pdf. [Last accessed on 2018 Dec 25].

SEER Program Code Manual. (1998). *SEER Geocodes for Coding Place of Birth*. 3rd ed. Retrieved from: http://www.seeer.cancer.gov. [Last accessed on 2018 Oct 13].

SEER Program Code Manual. (1998). *Tow-digit Site Specific Surgery Codes (1983-1997)*. 3rd ed. Retrieved from: http://www.seeer.cancer.gov. [Last accessed on 2018 Oct 16].

SEER Program Quality Control Section, Suite 504. (2007). *ICD-0-3 Seer Site/Histology Validation*. Retrieved from: http://www.seeer.cancer.gov. [Last accessed on 2019 Oct 19].

The Basics of SAS Enterprise Miner 5.2. (2018). Retrieved from: http://www.support.sas.com/publishing/pubcat/chaps/59829.pdf. [Last accessed on 2018 Oct 06].

Vesset, D., & Chua, C. K. (2017). *IDC's Worldwide Big Data and Analytics Software Taxonomy*. (pp. 1-14). North Korea: Big Data.

Witten, I. H., and Frank, E. (2005). *Data Mining, Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Elsevier Inc.