Introduction

Knowledge discovery differs from traditional information retrieval and databases. In traditional DBMS, database records are returned in response to a query; while in knowledge discovery, what is retrieved is not explicit in the database. Rather, it is implicit patterns. The Process of discovering such patterns is termed data mining.

Data mining finds these patterns and relationships using data analysis tools and techniques to build models. There are two main kinds of models in data mining. One is predictive models, which use data with known results to develop a model that can be used to explicitly predict values. Another is descriptive models, which describe patterns in existing data. All the models are abstract representations of reality, and can be guides to understanding business and suggest actions.

Data Mining for Medical Management has been instrumental in detecting patterns of diagnosis, decisions and treatments in Medical. Data mining has aided in several aspects of Medical management including disease diagnosis, decision-making for treatments, medical fraud prevention and detection, fault detection of medical devices, Medical quality improvement strategies and privacy. Data Mining for Medical Management is an emerging field where researchers from both academia and industry have recognized the potential of its impact on improved Medical by discovering patterns and trends in large amounts of complex data generated by Medical transactions.

Breast cancer accounts for 22.9% of all cancers (excluding non-melanoma skin cancers) in women. In 2008, breast cancer caused 458,503 deaths worldwide (13.7% of cancer deaths in women).Breast cancer is more than 100 times more common in women than in men, although men tend to have poorer outcomes due to delays in diagnosis.

Prognosis and survival rates for breast cancer vary greatly depending on the cancer type, stage, treatment, and geographical location of the patient. Survival rates in the Western world are high; for example, more than 8 out of 10 women (85%) in England diagnosed with breast cancer survive for at least 5 years. In developing countries, however, survival rates are much poorer.

1.2 Problem Statement

The prediction of breast cancer survivability has been a challenging research problem for many researchers, many researchers have tried many algorithms using seer data set[1] [1] [15].

1.3 The Research Objective

The main aim of this research is to compare the performance of three of classifications techniques (K-nearest neighbor, MLP and C4.5 in predicting breast cancer survivability.

1.4 Research Methodology

KDD road map will be followed in this research [4]. Since the first step in KDD is data selection and in this work the data is already selected, our work will go directly to the next step in KDD namely preprocessing. After preprocessing steps relevant to our classification task in KDD will be figured out and followed. This will be done with each one of the three classifiers. Comparison the discusser of the three experiments conduce the work.

1.5 Organization of the Thesis

Beside this chapter, this thesis consists of three chapters as follows chapter 2 provides literature review and previous works, and chapter 3 is about the experiments and the discussion of the results. Finally, chapter 4 contains conclusion and future works.

Chapter Two

Literature Review

2.1 Introduction

Relevant literature is reviewed in this chapter. We start from the definition of KDD and Data Mining and KDD steps in section 2.2.Data Mining Tasks was reviewed in section 2.3. Section 2.4 which explain the theory of C4.5 in 2.4.1, MLP in section 2.4.2 and K-nearest neighbor in section 2.4.3 algorithms, in section 2.5 we showed how to estimate model performance. Review of previous work relevant to the problem also discussed in section 2.6

2.2 Definition Discovery and Data Mining

KDD employs methods from various fields such as machine learning, artificial intelligence, pattern recognition, database management and design, statistics, expert systems, and data visualization. It is said to employ a broader model view than statistics and strives to automate the process of data analysis, including the art of hypothesis generation.

KDD has been more formally defined as the non-trivial process of identifying valid, over potentially useful and ultimately understandable patterns in data. The KDD Process is a highly iterative, user involved and multistep process [16].

According to definition above, the KDD is an interactive and iterative process. It means that at any stage the user should have possibility to make changes (for instance to choose different task or technique) and repeat the following steps to achieve better results. In table are listed particular steps of the KDD where we compared the terms of different sources. Table is organized on the way that the terms in the row refer to the same action.

Table 2.1 the KDD Steps

Process	Description	Step of Process	Result
	understanding the domain	learning the application domain	task discovery
data selection		creating a target dataset	data discovery
data transformation	preparing the data set	data cleaning and preprocessing	data cleaning
		data reduction and projection	model development
		choosing the function of data mining	
data mining	discovering patterns (data mining)	choosing the data mining algorithm(s)	data analysis
		data mining	
result interpretation	discovered patterns	interpretation	output generation
	putting the results into use	using discovered knowledge	

Task Discovery is one of first steps of KDD. Client has to state the problem or goal, which often seems to be clear. Further investigation is recommended such as to get acquainted with customer's organization after spending some time at the place and to sift through the raw data (to understand its form, content, organizational role and sources of data). Then the real goal of the discovery will be found.

Data Discovery is complementary to step of task discovery. In the step of data discovery, we have to decide whether quality of data is satisfactory for the goal (what data does or does not cover).

Data Cleaning is often necessary though it may happen that something removed by cleaning can be indicator of some interesting domain phenomenon (outlier or key data point?). Analyst's background knowledge is crucial in data cleaning provided by comparisons of multiple sources. Other way is to clean data before loaded into database by editing procedures.

Data reduction Finding useful features to represent the data depending on the goal of the task, using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.

2.3 Data mining tasks

Several data mining problem types or analysis tasks are typically encountered during a data mining project. Depending on the desired outcome, several data analysis techniques with different goals may be applied successively to achieve a desired result. For example, to determine which customers are likely to buy a new product, a business analyst may need first to se cluster analysis to segment the customer database, and then apply regression analysis to predict buying behavior for each cluster. The data mining analysis tasks typically fall into the general categories listed below. For each data analysis task, an example of a useful data analysis technique is presented.

Again, there is a continuum of data analysis techniques and the two disciplines of statistics and machine learning often overlap. Table 1 is a matrix that summarizes the data mining analysis tasks and the techniques useful for performing these tasks. The table is representative of the many possibilities since the permutations and combinations of data analysis tasks and techniques are numerous.

Data Summarization gives the user an overview of the structure of the data and is generally carried out in the early stages of a project. This type of initial exploratory data analysis can help to understand the nature of the data and to find potential hypotheses for hidden information. Simple descriptive statistical and visualization techniques generally apply.

Segmentation separates the data into interesting and meaningful sub-groups or classes. In this case, the analyst can hypothesize certain subgroups as relevant for the business question based on prior knowledge or based on the outcome of data description and summarization.

Automatic clustering techniques can detect previously unsuspected and hidden structures in data that allow segmentation. Clustering techniques, visualization and neural nets generally apply.

Classification assumes that a set of objects—characterized by some attributes or features—belong to different classes. The class label is a discrete qualitative identifier; for example, large, medium, or small. The objective is to build classification models that assign the correct class to previously unseen and unlabeled objects. Classification models are mostly used for predictive modeling. Discriminated analysis, decision tree, rule induction methods, and genetic algorithms generally apply.

Prediction is very similar to classification. The difference is that in prediction, the class is not a qualitative discrete attribute but a continuous one. The goal of prediction is to find the numerical value of the target attribute for unseen objects; this problem type is also known also known as regression, and if the prediction deals with time series data, then it is often called forecasting. Regression analysis, decision trees, and neural nets generally apply [5]

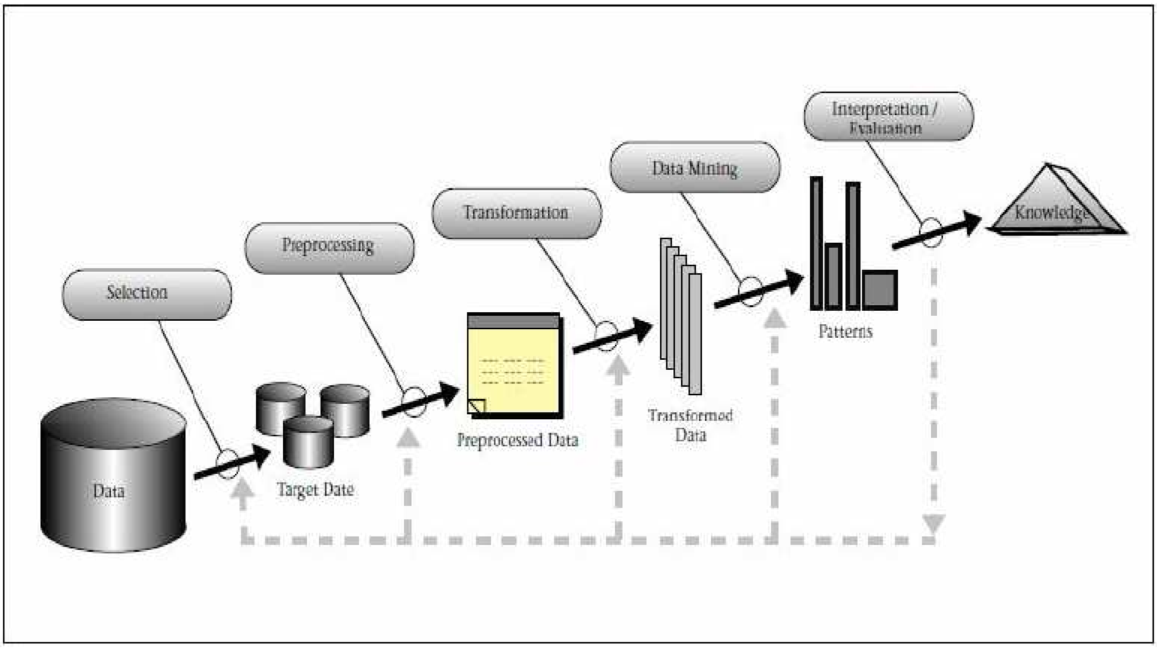


Figure 2.1 the description KDD Steps [14]

2.4 C4.5 decision trees

A complete description of C4.5, the early 1990s version, appears as an excellent and readable book (Quinlan 1993), along with the full source code.

The problem of constructing a decision tree can be expressed recursively. First, select an attribute to place at the root node and make one branch for each possible value. This splits up the example set into subsets, one for every value of the attribute. Now the process can be repeated recursively for each branch, using only those instances that actually reach the branch. If at any time all instances at a node have the same classification, stop developing that part of the tree, The method we have described only works when all the attributes are nominal [12].

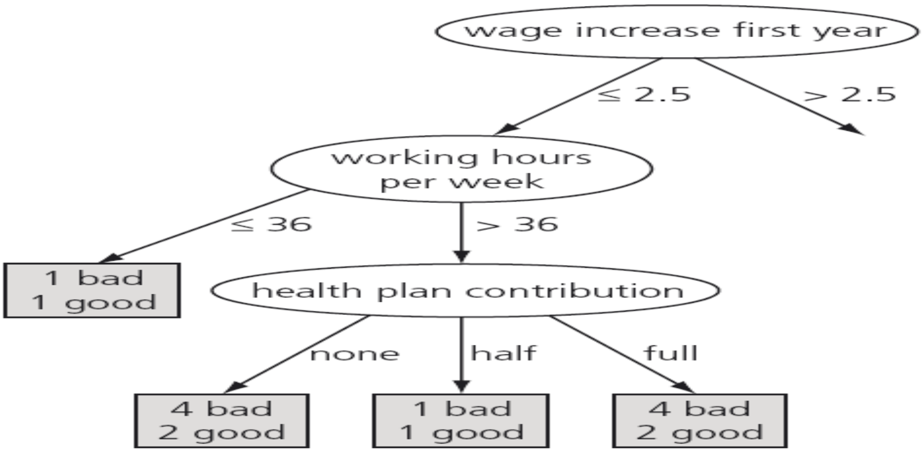


Figure 2.2 the description of C4.5 decision tree [1]

Decision tree construction process performed on very large datasets leads to bushy or meaningless results [13]. There are three steps to making C 4.5 decision tree

- Generalization compresses training data. This includes storage of generalized data in data cube to allow fast accessing
- Relevance analysis, that removes irrelevant attributes, thereby, further compacting training data.
- Multi-level mining, which combines the induction of decision trees with knowledge in concept hierarchies [13].

C4.5 disadvantages

- Decision-tree learners can create over-complex trees that do not **generalize** the data well. This is called **over fitting**. Mechanisms such as pruning (not currently supported), setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such

as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.

- There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.
- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

2.5 The Multi-Layer Perceptron (MLP) classifier

The multi-layer perceptron (MLP) training are based on the idea of changing network parameters (weight and biases) and checking the influence of such change on the mean-square error (MSE), or another error measure [11]. Building on the algorithm of the simple Perceptron, the MLP model not only gives a perceptron structure for representing more than two classes, it also defines a learning rule for this kind of network. The MLP is divided into three layers the input layer, the hidden layer and the output layer, where each layer in this order gives the input to the next. The extra layer gives the structure needed to recognize non-linearly separable classes [7].

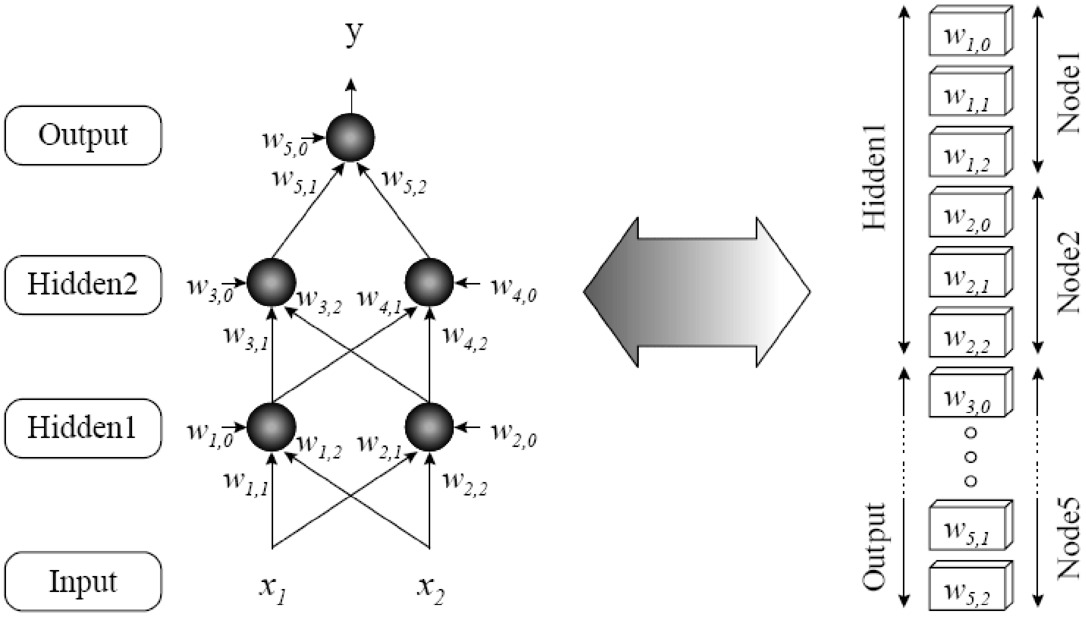


Figure 2.3 The description of MLP classifier [1]

MLP disadvantages

The limitations faced by neural network (MLP algorithm is kinds of ANN) are that it is unable to handle linguistic information and cannot manage imprecise or vague information. The inability to combine numeric data with linguistic or logical data is another major disadvantage of neural network. It is difficult to reach global minimum even by complex BP learning and relays on trial-and-errors to determine hidden layers and nodes [10].

2.6 K-nearest neighbor classifier

K-nearest neighbor (KNN) classification, finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood. There are three key elements of this approach a set of labeled objects, e.g., a set of stored records, a distance or similarity metric to compute distance between objects, and the value of k, the number of nearest neighbors. To classify an unlabeled object, the distance of this object to the labeled objects is computed, its k-nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the object.

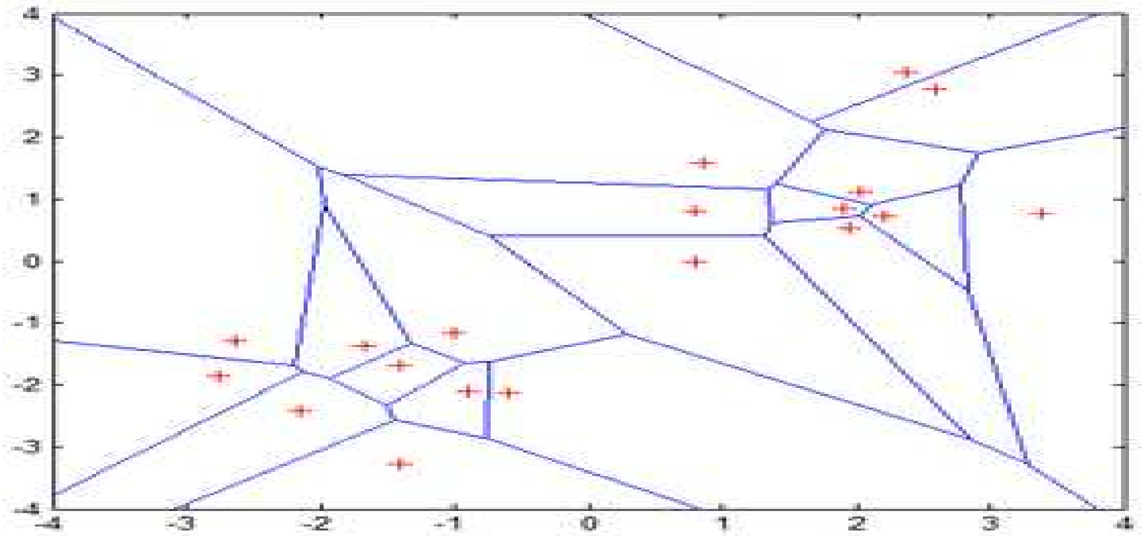


Figure 2.4 the description of K-nearest neighbor classifier [15]

The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters [12]

KNN disadvantages

The traditional KNN text classification algorithm has three limitations:

(i) Calculation complexity due to the usage of all the training samples for classification.

(ii) The performance is solely dependent on the training set (number of K).

(iii) There is no weight difference between samples.

2.7 Estimation for model performance

2.7.1 Accuracy, sensitivity, specificity

To estimate performance in the model, we used three performance measures. The accuracy, sensitivity, specificity are calculated as the following formulas

$$\mathrm{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \qquad (1)$$

$$\mathrm{sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (2)$$

$$\mathrm{specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \qquad (3)$$

True Positive (TP) means patients who are predicted as survived among survived patients at 5 years. True Negative (TN) means patients who are predicted as death among Not-survived patient. False Positive (FP) means patients who are predicted as survived among death patients. False Negative (FN) means patients who are predicted as death among survived patients. These values are often displayed in a confusion matrix as see in Table 2.2 .

Table 2.2 Confusion matrix

	Predicted class		
Actual class		Survived	Not-survived
	Survived	TP	FN
	Not-survived	FP	TN

2.7.2 K-Fold Cross-Validation

In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the predictive accuracy of two or more methods, we used k-fold cross-validation. In *k*-fold cross-validation, also called rotation estimation, the complete dataset (*D*) is randomly split into *k* mutually exclusive subsets (the folds D_1, D_2, . . ., D_k) of approximately equal size. The classification model is trained and tested *k* times. Each time (t ϵ {1, 2 . . . *k*}), it is trained on all but one folds (D_t) and tested on the remaining single fold (D_t). The cross-validation estimate of the overall accuracy is calculates as simply the average of the *k* individual accuracy measures

In this research, to estimate the performance of all selected classifiers a *10*-fold cross-validation approach is used. Empirical studies showed that 10 seem to be an optimal number of folds

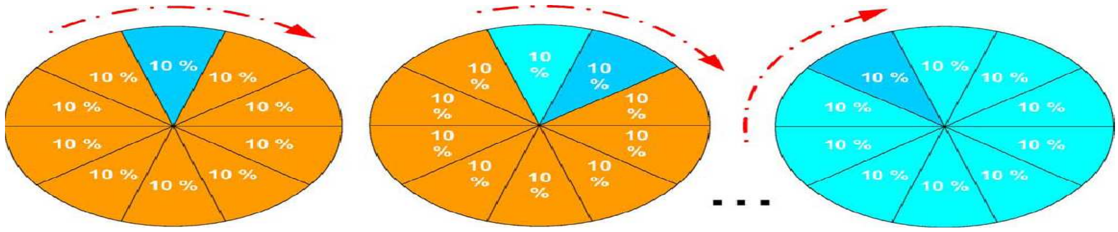


Figure 2.5 the description of cross validations

(That optimizes the time it takes to complete the test while minimizing the bias and variance associated with the validation process) [1]. In 10-fold cross-validation the entire dataset is divided into 10 mutually exclusive subsets (or folds). Each fold is used once to test the performance of the classifier that is generated from the combined data of the remaining nine folds, leading to 10 independent performance estimates (figures 2.5). Specifically, 10-fold cross-validation is calculated by following ways. First, the dataset is divided in 2 sections randomly. One section is included 90% of all dataset and called as learned dataset. Another section is included 10% of all dataset and called as validation dataset. Second, the process is repeated 10 times.

2.8 Related Work

A literature survey showed that there have been several studies on the survivability prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and survivability using data mining approaches like decision trees.

Abdelghani Bellaachia, Erhan Guven et al. [1]. They have investigated three data mining techniques the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms using seer data set (1973-2002). The achieved prediction performances are comparable to existing techniques. However, they found out that C4.5 algorithm has a much better performance than the other two techniques.

Dursun Delen, Glenn Walker and Amit Kadam et al. [1]. In this study,

Delen et al. preprocessed the SEER data (period of 1973-2000 with 433,272 records named as breast.txt) for breast cancer to remove redundancies and missing information. The resulting data set had 202,932 records, which then pre-classified into two groups of “survived” (93,273) and “not survived” (109,659) depending on the Survival Time Recode (STR) field. The “survived” class is all records that have a value greater than or equal 60 months in the STR field and the “not survived” class represent the remaining records. After this step, the data mining algorithms are applied on these data sets to predict the dependent field from 16 predictor fields. The results of predicting the survivability were in the range of 93% accuracy.

They used two popular data mining algorithms (artificial neural networks and decision trees) along with a most commonly used statistical method (logistic regression) to develop the prediction models using a large dataset, The results indicated that the decision tree (C4.5) is the best predictor with 93.6% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), artificial neural networks came out to be the second with 91.2% accuracy and the logistic regression models came out to be the worst of the three with 89.2% accuracy.

Chapter Three

Experiments & Results

3.1 Introduction

In this chapter, section 3.2 data description section 3.3 provides a detailed description of our preprocessing method to the data used for implementing the algorithms. Section 3.4 shows the experiments and results. The final section 3.5 discussions the results.

3.2 Data Description

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (NCI) is an authoritative source of information on cancer incidence and survival in the United States. SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 28% of the US population. SEER coverage includes 26% of African Americans, 38% of Hispanics, 44% of American Indians and Alaska Natives, 50% of Asians, and 67% of Hawaiian/Pacific Islanders

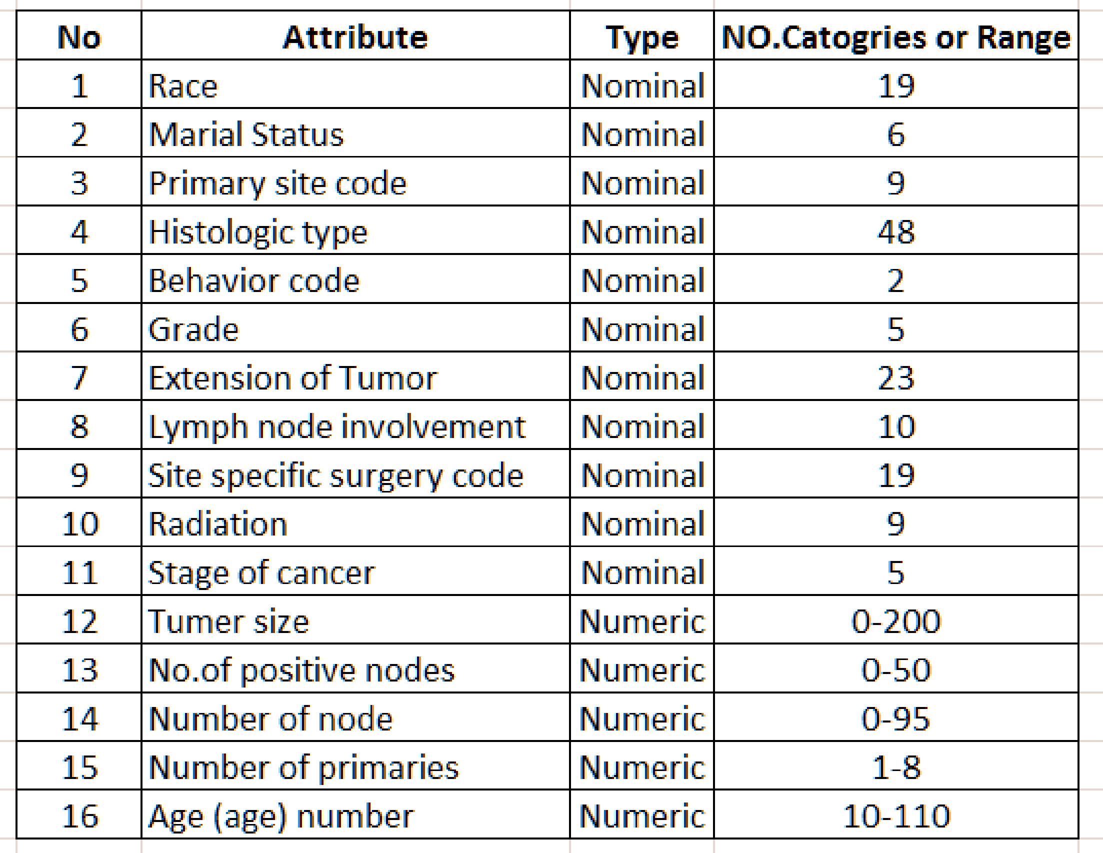
The SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. The SEER Program is the only comprehensive source of population-based information in the United States that includes stage of cancer at the time of diagnosis and patient survival data. The mortality data reported by SEER are provided by the National Center for Health Statistics. The population data used in calculating cancer rates is obtained periodically from the Census Bureau. Updated annually and provided as a public service in print and electronic formats, SEER data are used by thousands of researchers, clinicians, public health officials, legislators, policymakers, community groups, and the public.

NCI staff work with the North American Association of Central Cancer Registries (NAACCR) to guide all state registries to achieve data content and compatibility acceptable for pooling data and improving national estimates. The SEER team is developing computer applications to unify cancer registration systems and to analyze and disseminate population-based data. Use of surveillance data for research is being improved through Web-based access to the data and analytic tools, and linking with other national data sources. The data set use in this research is breast cancer data set from 1973 to 2009 contains of 657,712 cases and 135 fields.

3.3 Data Preprocessing

Before using KDD roadmap we found there are some variables not related to the disease according to previous research [1]. Then selected 16 variables.

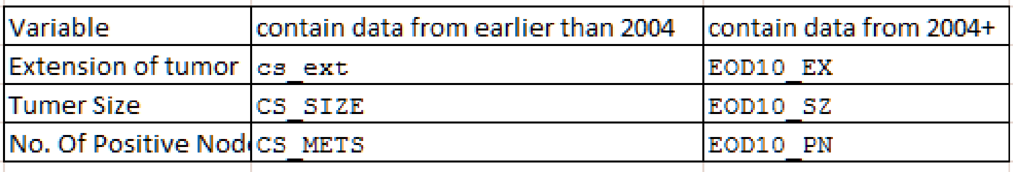
Table 3.1 attributes using to data mining



No	Attribute	Type	NO.Catogries or Range
1	Race	Nominal	19
2	Marial Status	Nominal	6
3	Primary site code	Nominal	9
4	Histologic type	Nominal	48
5	Behavior code	Nominal	2
6	Grade	Nominal	5
7	Extension of Tumor	Nominal	23
8	Lymph node involvement	Nominal	10
9	Site specific surgery code	Nominal	19
10	Radiation	Nominal	9
11	Stage of cancer	Nominal	5
12	Tumer size	Numeric	0-200
13	No.of positive nodes	Numeric	0-50
14	Number of node	Numeric	0-95
15	Number of primaries	Numeric	1-8
16	Age (age) number	Numeric	10-110

Also there is some fields change in 2004 to anther fields (The fields contain the same data but changed the categories), we merged this fields with new categories:

Table **Error! No text of specified style in document.**3.2 attributes merging



Variable	contain data from earlier than 2004	contain data from 2004+
Extension of tumor	cs_ext	EOD10_EX
Tumer Size	CS_SIZE	EOD10_SZ
No. Of Positive Nod	CS_METS	EOD10_PN

Also we need to calculate survival field that is calculated from this algorithms:

```
If STR $\geq$ 60 months and VSR is alive then
    The record is pre-classified as “survived”
Else if STR $<$ 60 months and COD is breast cancer, then
    The record is pre-classified as “not survived”
Else
    Ignore the record
End if
```

This algorithm divided cases (records) to live who alive 5 or more years after diagnose and not alive for cases live less than this time and ignore other cases.

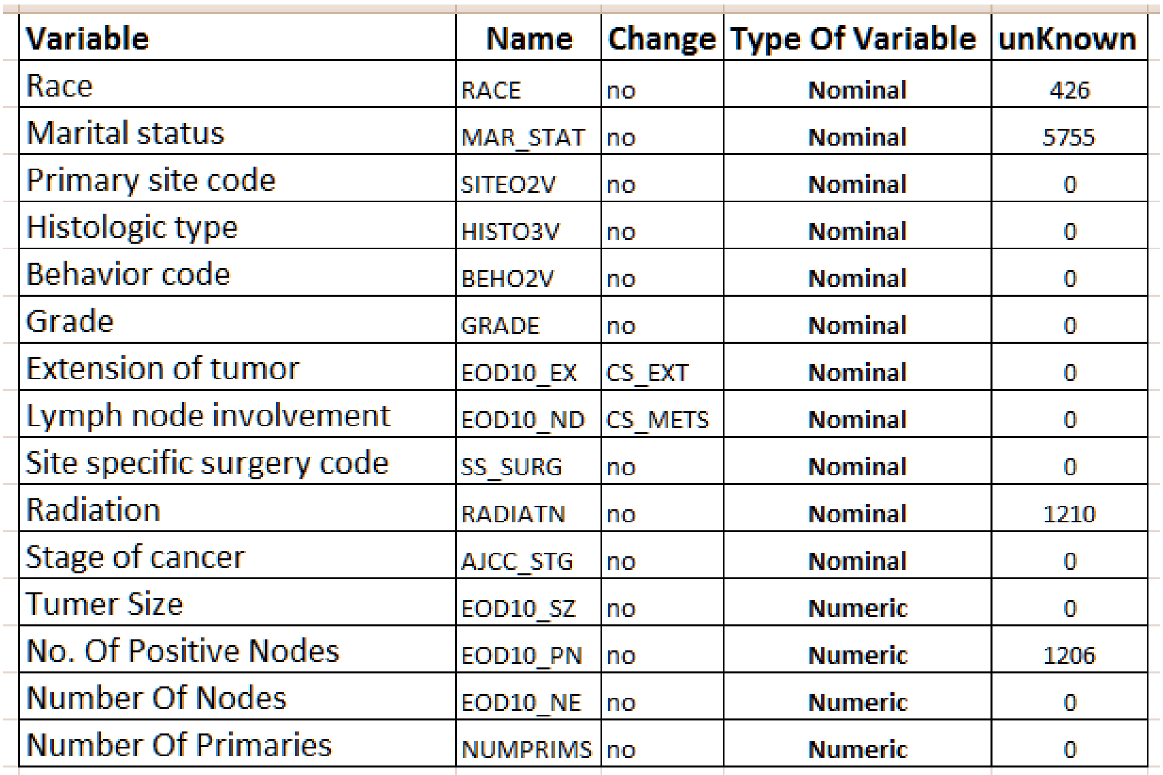
Now we follow KDD roadmap:

1- The first step is data selection and this is already done in the data description sections.

2- The second step is data transformation which contains cleaning, reductions and choosing the functions.

2.1 Cleaning: To cleaning data set we used SQL Server, to delete record contain missing values.

Table **3.3** Deleted missing and outlier values



Variable	Name	Change	Type Of Variable	unKnown
Race	RACE	no	**Nominal**	426
Marital status	MAR_STAT	no	**Nominal**	5755
Primary site code	SITEO2V	no	**Nominal**	0
Histologic type	HISTO3V	no	**Nominal**	0
Behavior code	BEHO2V	no	**Nominal**	0
Grade	GRADE	no	**Nominal**	0
Extension of tumor	EOD10_EX	CS_EXT	**Nominal**	0
Lymph node involvement	EOD10_ND	CS_METS	**Nominal**	0
Site specific surgery code	SS_SURG	no	**Nominal**	0
Radiation	RADIATN	no	**Nominal**	1210
Stage of cancer	AJCC_STG	no	**Nominal**	0
Tumer Size	EOD10_SZ	no	**Numeric**	0
No. Of Positive Nodes	EOD10_PN	no	**Numeric**	1206
Number Of Nodes	EOD10_NE	no	**Numeric**	0
Number Of Primaries	NUMPRIMS	no	**Numeric**	0

2.2 Reduction: We already selected 16 fields according to previous researches [1]

2.3 Mining: to predicting the survivability three classification models C4.5 decision tree, MLP and KNN are selected.

Table **3.4** Data description after Preprocess

Class	Number of records	Percentage
Survived=L	171,953	95%
Not-survived=D	8,349	5%
Total	180302	100%

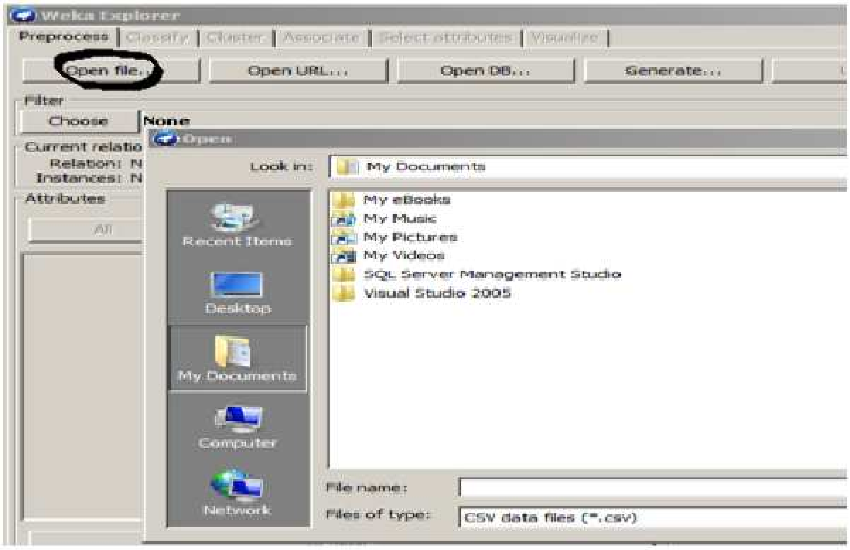
3.4 Experiments and results

To implement the three modules, weka tools 3.7.1 has been used Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes [10]. Weka is open source software issued under the GNU General Public License.

To enter the data into weka, it has been converted to csv format.

Database => task =>export (choose file .csv)

- Insert data into Weka tools



Look in: My Documents

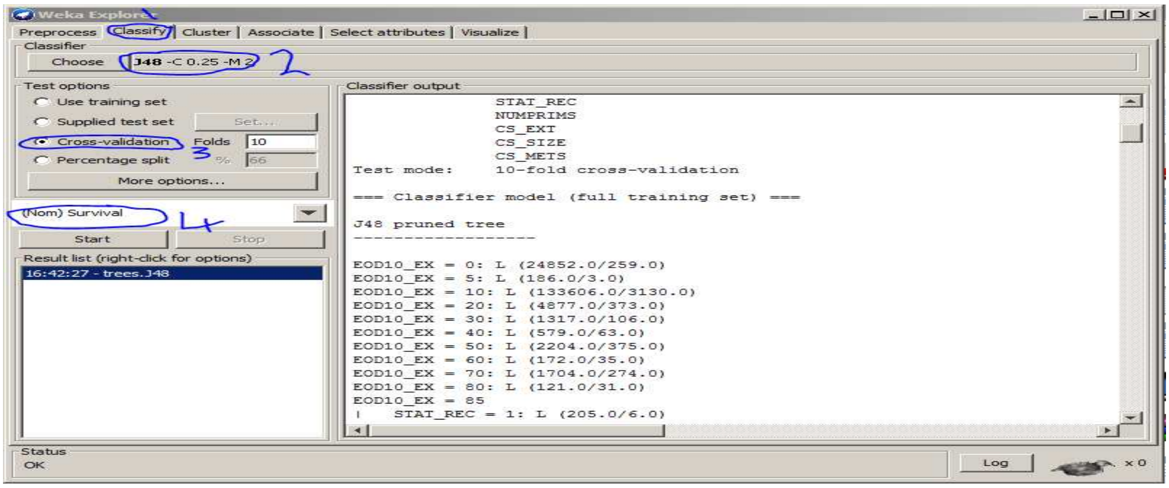
File name:

Files of type: CSV data files (*.csv)

Figure 3.1 convert data to Weka tool from CSV file

3.4.1 C4.5 experiments and results

C4.5 was applied in Weka tools by choosing **Classify** in Weka tool and choose **J48 –C0.25 –m2, Cross-Validation** type of validations**,** and survival **variable Test.**



Classifier: J48 -C 0.25 -M 2

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

More options...

(Nom) Survival

Result list (right-click for options)
16:42:27 - trees.J48

Classifier output

```
                  STAT_REC
                  NUMPRIMS
                  CS_EXT
                  CS_SIZE
                  CS_METS
Test mode:        10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

EOD10_EX = 0: L (24852.0/259.0)
EOD10_EX = 5: L (186.0/3.0)
EOD10_EX = 10: L (133606.0/3130.0)
EOD10_EX = 20: L (4877.0/373.0)
EOD10_EX = 30: L (1317.0/106.0)
EOD10_EX = 40: L (579.0/63.0)
EOD10_EX = 50: L (2204.0/375.0)
EOD10_EX = 60: L (172.0/35.0)
EOD10_EX = 70: L (1704.0/274.0)
EOD10_EX = 80: L (121.0/31.0)
EOD10_EX = 85
|   STAT_REC = 1: L (205.0/6.0)
```

Status
OK

Figure 3.4 C4.5 module Implement

The result:

Table **3.5** show C4.5 result of this experiment

The Result	The Value	Percentage of Value
Correctly Classified Instances	172383	95.6079 %
Incorrectly Classified Instances	7919	4.3921 %
Kappa statistic	0.2128	
Mean absolute error	0.0726	
Root mean squared error	0.1911	
Relative absolute error		82.1443 %

Root relative squared error		90.9318 %
Coverage of cases (0.95 level)		98.0882 %
Mean rel. region size (0.95 level)		55.9411 %
Total Number of Instances		18030

Table **3.6** show C4.5 confusion matrix of this experiment

L	D	<-- classified as
171225	728	L
7191	1158	D

3.4.2 MLP experiments and results

Implement MLP algorithm in Weka tools by choosing **Classify** in Weka tool and choose **Multilayer Perceptron, Cross-Validation** type of validations, and survival **variable Test.**

The result:

Table **3.7** she MLP result of this experiment

The Result	The Value	Percentage Of Value
Correctly Classified Instances	171965	95.3761 %
Incorrectly Classified Instances	8337	4.6239 %
Kappa statistic	0.187	

Total Cost	8337	
Average Cost	0.0462	
K&B Relative Info Score		-6235398.0321 %
K&B Information Score	-16867.5479 bits -0.0936 bits/instance	
Class complexity | order 0	34130.1311 bits 0.1893 bits/instance	
Class complexity | scheme	34130.1311 bits 0.1893 bits/instance	
Complexity improvement	14639.9925 bits 0.0812 bits/instance	
Mean absolute error	0.0719	
Root mean squared error	0.1887	
Relative absolute error		81.3782 %
Root relative squared error		89.7747 %
Total Number of Instances	180302	

Table **3.8** show MLP confusion matrix of this experiment

a	b	<-- classified as
170909	1044	a = L
7293	1056	b = D

3.4.3 K-nearest neighbor experiments and results

Implement KNN model in Weka tools by choosing **classify** in Weka tool and choose**,** **cross-validation** type of validations**,** and KNN is k values**.**

The Results

Figure 3.9 show KNN result of this experiment

The Result	The Value	Percentage Of Value
Correctly Classified Instances	172049	95.4227 %
Incorrectly Classified Instances	8253	4.5773 %
Kappa statistic	0.2605	
Mean absolute error	0.0671	
Root mean squared error	0.2015	
Relative absolute error		75.9464 %
Root relative squared error		95.8884 %
Coverage of cases (0.95 level)		98.1315 %
Mean rel. region size (0.95 level)		58.4023

Table **3.10** show KNN confusion matrix of this experiment

L	D	<-- classified as
170452	1501	L
6752	1597	D

3.5 Discussion of the results

The Result indicted that C4.5 is best model to predicting survivability with accuracy 95.6%, also we found that Kappa statistic is 0.187 and this shows that the agreement of the accuracy is less, also Number of Leaves is 294 and Size of the tree is 310 and this is big tree that main complex result, The second model is KNN with accuracy of 95.4 also we find Kappa statistic is 0.2605 and this value more than C4.5 and MLP, the third models MLP with accuracy 95.3%, and Kappa statistic is 0.187 and this show that the agreement of result is little. We had shown that the accuracy in three models have been approximated.

Chapter Four

Conclusion & Future work

4.1 Conclusions

This research has outlined, discussed and resolved the issues, algorithms, and techniques for the problem of breast cancer survivability prediction in SEER database. We used three popular data mining methods ANN (MLP- Multilayer Perceptron), decision trees (C4.5) and clustering K-nearest neighbor. We acquired a quite large dataset (657,712 cases with 138 attributes) from the SEER program and after going through a long process of data cleansing and transformation used it to develop the prediction models. In this research, we defined survival as any incidence of breast cancer where person is still alive after 5 years (60 months) from the date of diagnosis. We used a 10-fold cross-validation procedure. That is, we divided the dataset into 10 mutually exclusive partitions (a.k.a. folds) using a stratified sampling technique. Then, we used 9 of 10 folds for training and the 10th for the testing and split test for K-nearest neighbor model. The results show that the classification models C4.5 with accuracy 95.6% is better than the K-nearest neighbor with accuracy 95.4% which is better than MLP with accuracy 95.3%.

4.2 Future work

There are many classifications models that are not tried. Also there are many problems facing the preprocessing, there are missing data in the EOD field from the old EOD fields prior to 1988 if we solving this might increase the performance as the size of the data set will increase considerably.

References

[1] Abdelghani Bellaachia, Erhan Guven, “Predicting Breast Cancer Survivability Using Data Mining Techniques”, feb 2005.

[2] , Am Stat. Author manuscript, “Linear Transformations and the k-nearest neighbor Clustering Algorithm applications to Clustering Curves”, Mar 16, 2007.

[3]Arabinda Nanda1 Saroj Kumar, “Data Mining & Knowledge Discovery in Databases” Proceedings of national Seminar on Future Trends in Data Mining ,10th may, 2010.

[4] B. de la Iglesia, C.M.Howard, V.J.Rayward-Smith J.C.W.Debuse, "A methodology for knowledge discovery A KDD Roadmap SYS Technical Report SYS-C99-01", 1999.

[5] Communications of the Association for Information Systems, Volume 8, 267-296, 2002.

[6] Dursun Delen*, Glenn Walker, Amit Kadam, “Predicting breast cancer survivability a comparison of three data mining methods”, 15 July 2004.

[7] Fiona Nielsen, “Neural Networks – algorithms and applications”, 4i, 12/12-2001.

[8] http//seer.cancer.gov/resources/

[9] http//www.microsoft.com/en-us/download/details.aspx?id=1695

[10] http//www.cs.waikato.ac.nz/ml/weka/

[11] Irosław Kordos, “Variable Step Search Algorithm For MLP Training”.

[12] Jim Gray, “Data Mining Practical Machine Learning Tools and Techniques”, second edition: Microsoft Research.

[13] Mohd. Mahmood Ali1, Mohd. S. Qaseem2, Lakshmi Rajamani3, A. Govardhan4, “Extracting Useful Rules Through Improved Decision Tree Induction Using Information Entropy”, Vol.3, No.1, January 2013.

[14] M. De Martino, A. Bertone, R. Albertoni, H. Hauska, U. Demsar, M. Dunkars” Technical Report of Data Mining”, INVISIP IST-2000-29640, 28.2.2002

[15] Rajesh1, Dr. Sheila An and PG, “Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm K”, April 2012.

[16] “Software Defects Classification Using FB-MLP Neural Network”, CHAPTER 5.

[17] Susan P. Imberman Ph.D, “Effectie Use of The KDD Process and Data Mining for Computer Performance Professionals”,

[18] XindongWu · Vipin Kumar J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng ·Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · DanSteinberg, “Top 10 algorithms in data mining”, 9 July 2007.

Appendix

SEER Breast Cancer Dataset Attributes

No	Attribute name	Description
1	Race	Patient race
2	Age	The age of the patient at diagnosis for breast cancer
3	Gender	sex of the patient at diagnosis
4	Marital status at diagnosis	This data item identifies the patient’s marital status at the time of diagnosis for the reportable tumor.
5	Primary Site	Tumor first location in the body
6	Histology	Tumor morphology
7	Behavior	
8	Grade	Indicates how the cancer cells appear and how fast they may grow and spread
9	Stage of cancer	Physical location and spread
10	Laterality	Laterality describes the side of a paired organ or side of the body on which the reportable tumor originated.
11	Sequence_Number	Sequence Number-Central describes the number and sequence of all reportable malignant, in situ, benign, and borderline primary tumors, which occur over the lifetime of a patient

12	Tumor Size	records the largest dimension of the primary tumor in millimeters
13	Tumor Extension	Identifies the growth of the primary tumor within the organ of origin or its direct extension into neighboring organs
14	Lymph node involvement	Indicates if the tumor involves lymph node chains
15	Number of positive node	Records the exact number of regional lymph nodes examined by the pathologist that were found to contain metastases
16	Number of node examined	The total number of regional lymph nodes that were removed and examined by the pathologist
17	Radiation	Radiation therapy method on first treatment
18	Radiation sequence with surgery	Sequence of administering radiation such as pre-surgery, post-surgery and pre-/post-surgery radiation
19	Site specific surgery code	The surgical procedure that removes and destroys cancerous tissue of the breast, performed as part of the initial first course of therapy
20	No surgery	The reason why surgery was not perform
21	Number of primaries	total number of tumors
22	Survival code	Denotes if patient has survived or not based on survival time and vital status recode and cause of death

23	Patient ID	This field is used in conjunction with SEER registry to uniquely identify a person
24	SEER Registry	A unique code assigned to each participating SEER registry.
25	Spanish surname or origin	This data item is used to identify patients with Spanish/Hispanic surname or of Spanish origin. Persons of Spanish or Hispanic surname/origin may be of any race.
26	NHIA Derived Hispanic Origin	Hispanic Identification Algorithm (NHIA) is a computerized algorithm that uses a combination of variables to directly or indirectly classify cases as Hispanic for analytic purposes.
27	Year of birth	
28	Place of birth	
29	Month of diagnosis	The month of diagnosis is the month the tumor was first diagnosed by a recognized medical practitioner, whether clinically or microscopically confirmed.
30	Year of diagnosis	The year of diagnosis is the year the tumor was first diagnosed by a recognized medical practitioner, whether clinically or microscopically confirmed.
31	Histology (92-00) ICD-O-2	
32	Behavior code ICD-O-2	

33	Diagnostic Confirmation	This data item records the best method used to confirm the presence of the cancer being reported.
34	Type of Reporting Source	The source documents used to abstract the case.
35	EOD 10—Prostate path ext (1995-2003)	This is an additional field for prostate cancer only to reflect information from radical prostatectomy, effective with 1995 diagnoses.
36	Expanded EOD(1) – Expanded EOD (13)	Detailed site-specific codes for EOD used by SEER for selected sites of cancer for tumors diagnosed 1973-1982, except death-certificate-only cases.
37	2-Digit NS EOD / 2-Digit SS EOD	Site-specific codes for EOD used by SEER for tumors diagnosed from January 1, 1973, to December 31, 1982, for cancer sites that did not have a 13-digit scheme.
38	EOD—Old 4 Digit	Codes for site-specific EOD used by SEER for tumors diagnosed from January 1, 1983 to December 31, 1987 for all cancer sites.
39	Coding system—EOD (1973-2003)	Indicates the type of SEER EOD code applied to the tumor.
40	Tumor Marker 1	This data item records prognostic indicators for breast cases (ERA 1990-2003), prostate cases (PAP 1998-2003) and testis cases (AFP 1998-2003).
41	Tumor Marker 2	This data item records prognostic indicators for

		breast cases (PRA 1990-2003), prostate cases (PSA 1998-2003), and testis cases (hCG 1998-2003).
42	Tumor Marker 3	This data item records prognostic indicators for testis cases (LDH 1998-2003)
43	CS mets at dx (2004+)	Information on distant metastasis.
44	CS site-specific factor 1	They can provide information needed to stage the case, clinically relevant information, or prognostic information.
45	CS site-specific factor 2 (2004+)	They can provide information needed to stage the case, clinically relevant information, or prognostic information.
46	CS site-specific factor 3 (2004+)	-
47	CS site-specific factor 4 (2004+)	-
48	CS site-specific factor 5 (2004+)	-
49	CS site-specific factor 6 (2004+)	-
50	CS site-specific factor 25 (2004+)	-
51	Derived AJCC T, 6th ed (2004+)	This is the AJCC “T” component that is derived from CS coded fields.
52	Derived AJCC N, 6th ed	This is the AJCC “N” component that is derived

	(2004+)	from CS coded fields.
53	Derived AJCC M, 6th ed (2004+)	This is the AJCC “M” component that is derived from CS coded fields.
54	Derived AJCC Stage Group, 6th ed (2004+)	This is the AJCC “Stage Group” component that is derived from CS detailed site-specific codes.
55	Derived SS1977 (2004+)	This item is the derived “SEER Summary Stage 1977”
56	Derived SS2000 (2004+)	This item is the derived “SEER Summary Stage 2000”
57	Derived AJCC—Flag (2004+)	Flag to indicate whether the derived AJCC stage was derived from CS or EOD codes.
58	Derived SS1977—Flag (2004+)	Flag to indicate whether the derived SEER Summary Stage 1977 was derived from CS or EOD codes.
59	Derived SS2000—Flag (2004+)	Flag to indicate whether the derived SEER Summary Stage 2000 was derived from CS or EOD codes.
60	CS version input (2004+)	This item indicates the number of the version used to initially code CS fields.
61	CS version latest (2004+)	This item indicates the number of the version of the CS used most recently to derive the CS output fields.
62	CS version input current (2004+)	This item indicates the number of the version of the CS after input fields have been updated or recoded.

63	RX Summ--Surg Prim Site (1998+)	Surgery of Primary Site describes a surgical procedure that removes and/or destroys tissue of the primary site performed as part of the initial work-up or first course of therapy.
64	RX Summ--Scope Reg LN Sur (2003+)	Scope of Regional Lymph Node Surgery describes the procedure of removal, biopsy, or aspiration of regional lymph nodes performed during the initial work-up or first course of therapy at all facilities.
65	RX Summ--Surg Oth Reg/Dis (2003+)	Surgical procedure of Other Site describes the surgical removal of distant lymph node(s) or other tissue(s) or organ(s) beyond the primary site.
66	Num of regional lym nd exam (1998-2002)	This data item records the number of regional lymph nodes examined in conjunction with surgery performed as part of the first course of treatment at all facilities. This item is only available for cases diagnosed 1998-2002.
67	First course of reconstruct (1998-2002)	The SEER program collects information in this field only for breast cancer and only for reconstruction begun as part of first course of treatment.
68	Radiation to Brain or CNS (1988-1997)	This variable was only collected for years 1988-1997 for breast and leukemia cases only. This data item codes for radiation given to the brain or central nervous system at all facilities as part of the first course of therapy.

69	Surgery of primary site (1998-2002)	Site-specific codes for the type of surgery to the primary site performed as part of the first course of treatment at all facilities for cases diagnosed 1998-2002.
70	Scope of reg lymph nd surg (1998-2002)	This field describes the removal, biopsy or aspiration of regional lymph node(s) at the time of surgery of the primary site or during a separate surgical event at all facilities for cases diagnosed 1998-2002.
71	Surgery of oth reg/dis sites (1998-2002)	This field records the removal of distant lymph nodes or other tissue(s)/organ(s) beyond the primary site given at all facilities as part of the first course of treatment for cases diagnosed 1998-2002.
72	Record number	The Record Number is a unique sequential number.
73	Age-site edit override	-
74	Sequence number-dx conf override	-
775	Site-type-lat-seq override	-
76	Surgery-diagnostic conf override	-
77	Site-type edit override	-
78	Histology edit override	-

79	Report source sequence override	-
80	Seq-ill-defined site override	-
81	Leuk-Lymph dx confirmation override	-
82	Site-behavior override	-
83	Site-EOD-dx date override	-
84	Site-laterality-EOD override	-
85	Site-laterality-morph override	-
86	Type of follow-up expected	This item codes the type of follow-up expected for a SEER case.
87	Age recode with <1 year olds	The age recode variable is based on Age at Diagnosis (single-year ages).
89	Site recode	A recode based on Primary Site and ICD-O-3 Histology in order to make analyses of site/histology groups easier.
90	Site rec with Kaposi and mesothelioma	A recode based on Primary Site and ICD-O-3 Histology in order to make analyses of site/histology groups easier.
91	Recode ICD-O-2 to 9	The primary site and morphology are recoded to ICD-9 codes using the *Conversion of Malignant*

		Neoplasms by Topography and Morphology from the International Classification of Disease for Oncology, Second Edition (ICD-0-2) to International Classification of Diseases.
92	Recode ICD-O-2 to 10	-
93	ICCC site recode ICD-O-2	A site/histology recode based on the International Classification of Childhood Cancer (ICCC) is mainly used to analyze data on children.
94	SEER modified ICCC site recode ICD-O-2	A site/histology recode based on the International Classification of Childhood Cancer (ICCC) with slight modifications is mainly used to analyze data on children.
95	ICCC site recode ICD-O-3	A site/histology recode that is mainly used to analyze data on children.
96	ICCC site recode extended ICD-O-3	-
97	Behavior recode for analysis	This recode was created so that data analyses could eliminate major groups of histologies/behaviors that weren't collected consistently over time, for example benign brain, myelodyplastic syndromes, and borderline tumors of the ovary.
98	Histology recode - broad groupings	Based on Histologic Type ICD-O-3.
99	Histology recode -	Based on Histologic Type ICD-O-3.

	1Brain groupings	
100	CS Schema – v 0203	CS information is collected under the specifications of a particular schema based on site and histology.
101	Race recode (White, Black, Other)	Race recode is based on the race variables and the American Indian/Native American IHS link variable.
102	Race recode (W, B, AI, API)	-
103	Origin recode NHIA (Hispanic, Non-Hisp)	-
104	SEER historic stage A	Derived from Collaborative Stage (CS) for 2004+ and Extent of Disease (EOD) from 1973-2003.
105	AJCC stage 3rd edition (1988-2003)	-
106	SEER modified AJCC stage 3rd (1988-2003)	-
107	SEER summary stage 1977 (1995-2000)	-
108	SEER summary stage 2000 (2001-2003)	-
109	First malignant primary indicator	Based on all the tumors in SEER. Tumors not reported to SEER are assumed malignant.
110	State-county	-

111	Survival time recode	The Survival Time Recode is calculated using the date of diagnosis and one of the following date of death, date last known to be alive, or follow-up cutoff date used for this file.
112	COD to site recode	This recode was introduced to account for several newly valid ICD-10 codes and includes both cancer and non-cancer causes of death.
113	COD to site rec KM	This is a recode based on underlying cause of death to designate cause of death into groups similar to the incidence site recode with KS and mesothelioma.
114	Vital status recode	Any patient that dies after the follow-up cut-off date is recoded to alive as of the cut-off date.
115	IHS Link	Incidence files are periodically linked with Indian Health Service (IHS) files to identify Native Americans.
116	Summary Stage 2000 (1998+)	Summary Stage 2000 is derived from Collaborative Stage (CS) for 2004+ and Extent of Disease (EOD) from 1998-2003. It is a simplified version of stage in situ, localized, regional, distant, & unknown.
117	AYA site recode	A site/histology recode that is mainly used to analyze data on adolescent and young adults.
118	Lymphoma subtype recode	A site/histology recode that is mainly used to analyze data on adolescent and young adults.
119	SEER cause-specific	-

	death classification	
120	SEER other cause of death classification	-
121	CS Tumor Size/Ext Eval (2004+)	Available for 2004+, but not required for the entire timeframe. Will be blank in cases not collected, see http//seer.cancer.gov/seerstat/variables/seer/ajcc-stage.
122	CS Reg Nodes Eval (2004+)	Available for 2004+, but not required for the entire timeframe. Will be blank in cases not collected, see http//seer.cancer.gov/seerstat/variables/seer/ajcc-stage.
123	Primary by International Rules	Created using IARC multiple primary rules. Did not include benign tumors or non-bladder in situ tumors in algorithm.
124	ER Status Recode Breast Cancer (1990+)	Created by combining information from Tumor marker 1 (1990-2003) (NAACCR Item #=1150), with information from CS site-specific factor 1 (2004+) (NAACCR Item #=2880). This field is blank for non-breast cases.
125	PR Status Recode Breast Cancer (1990+)	Created by combining information from Tumor marker 2 (1990-2003) (NAACCR Item #=1150), with information from CS site-specific factor 2 (2004+) (NAACCR Item #=2880). This field is blank for non-breast cases.

126	CS Schema- AJCC 6th ed	CS information is collected under the specifications of a particular schema based on site and histology.
127	CS site-specific factor 8 (2004+)	Each CS site-specific factor (SSF) is schema dependent. They can provide information needed to stage the case, clinically relevant information, or prognostic information.
128	CS site-specific factor 10 (2004+)	-
129	CS site-specific factor 11 (2004+)	-
130	CS site-specific factor 13 (2004+)	-
131	CS site-specific factor 15 (2004+)	-
132	CS site-specific factor 16 (2004+)	-
133	Lymph-vascular Invasion (2004+)	LVI is required for cases originally coded under CSv2 or diagnosed 2010+ for the schemas for penis and testis only.