Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, Volume: 1, Issue: 1, JAN. 2025

Machine Learning Classification Methods Performance Comparison in Liver Cancer Cohort

Soran Husen Mohamad[1]

[1]Statistics and informatics department - College of Administration and Economics -University of Sulaimani- Sulaimani city – Iraq

[*]Corresponding author E-mail: (soran.abdulrahman@univsul.edu.iq)

Abstract. *The goal of this paper is to compare the performance of the statistical and machine learning classification methods in diagnosing the death of liver cancer patients based on demographic characteristics, risk factors, and medical interventions. For this purpose, five methods; include random tree, C4.5, random forest, support vector machine (SVM) and logistic regression, all of which are supervised methods, were selected. The data used in this research are the real data of 165 patients diagnosed with liver cancer in a hospital in Portugal. The aim variable is the patient's death during the trial period, and the aforementioned group was monitored for a year. There are twenty-six qualitative and twenty-three quantitative variables in this diverse dataset. In total, 10.22% of the dataset is missing data, and just eight patients have full information in every field (4.85%). Additionally, there is some class disparity (63 cases classified as "Dead" and 102 as "Alive"). With 73.33% accurate detection, the SVM approach was found to be the most effective approach. After that, the random forest method with 71.52% had a more correct identification ratio than the others. The treesC4.5 method had the lowest correct diagnosis with 58.18%. Although, based on the ROC Area index, the random forest method performed better (with an area under the curve = 0.789) than the SVM method (with an area under the curve = 0.711). In total, SVM and random forest methods worked with a large difference compared to others in diagnosing the death of patients with liver cancer.*

Keywords: *Random tree, C4.5, Random forest, SVM, Logistic regression, Liver Cancer.*

1. INTRODUCTION

The three most popular forms of machine learning are supervised learning, unsupervised learning, and unsupervised learning. In the contemporary world, every significant practical scientific sector has been impacted by machine learning, artificial intelligence, and the abundance of study on these subjects..

Many types of machine learning methods have been introduced, so that each one is used in specific situations. Also, machine learning methods have different applications in the field of data mining, such as prediction, classification, etc.

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, Volume: 1, Issue: 1, JAN. 2025

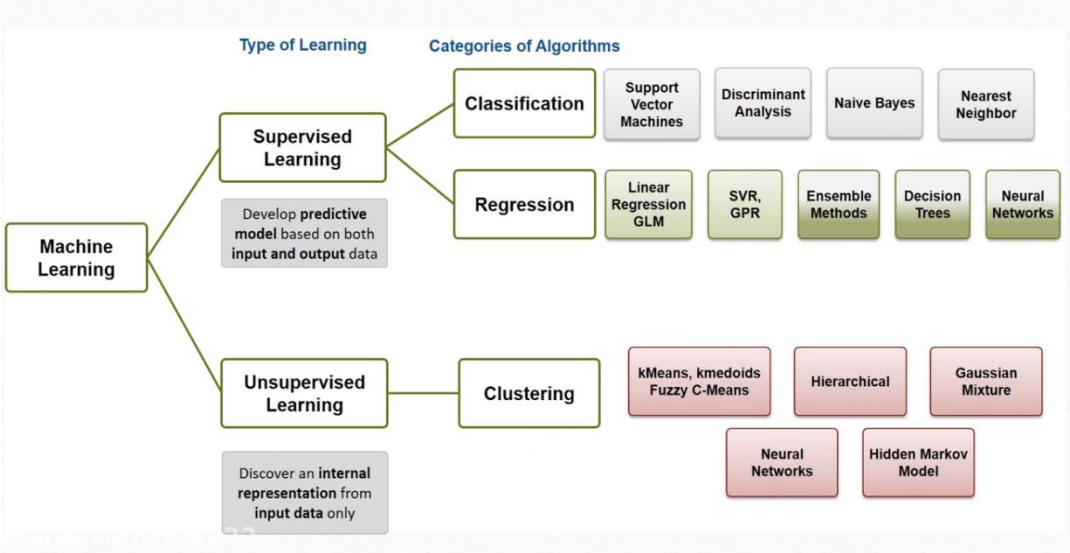

Figure 1. some machine learning methods in tow category, show those applications

The methods examined in this research are all supervised.

2. STATISTICAL SUPERVISED LEARNING

In statistics, the probability of output is calculated based on input. If the input is x and the output is y, $p(y|x)$ is learned from the data, in other words learning is actually finding the function p. There are two general methods for finding the function p: Generative and Discriminative. To put it very simply, in the diagnostic method, the machine learns the boundary between different classes, but in the production method, the machine learns how to produce samples of the same class. Mathematically speaking, in the diagnostic method $p(y|x)$ is learned directly, but in the production method, p(y) and $p(x|y)$ are estimated from the data and then calculated using the Bayes rule p(y|x) (Ng & Jordan, 2001).

3. MATHEMATICAL DEFINITION OF SUPERVISED LEARNING

In supervised learning, the training examples are in the form of $(x_i. y_i)$ pairs where each example is given along with its label and i is the index of each example in the set of training examples D. The goal in this learning is to get the function f that can return the appropriate patch for unseen input samples x, i.e. $f(x)$. Both instance and label can be a vector. If the label is a real number, the problem before us is called "Regression". If the label is an integer, the problem is called "Classification". The application of classification methods in clinical sciences in order to predict the survival of patients has received much attention from researchers. In the second section of this paper, the five methods reviewed in this paper will be briefly reviewed as well as the statistical indicators to compare their results. In the third section, we talk about the data and the results of implementing

different methods. In the last section, the results of the comparisons will be discussed based on the indicators mentioned in the second section.

4. MACHINE LEARNING CLASSIFICATION METHODS

Many different classification methods have been introduced in supervised machine learning. In this paper, just five methods are used, which will briefly mention their working methods and their differences. These methods are random tree, C4.5, random forest, support vector machine (SVM) and logistic regression.

4.1 Random tree

A supervised classification technique, Random Tree is an ensemble learning algorithm that produces a large number of individual learners. In order to generate a random collection of data for building a decision tree, it uses the bagging concept. The optimum split across all variables is used to split each node in a typical tree. Adele Cutler and Leo Breiman introduced this technique. The approach is applicable to both regression and classification issues. In essence, Random Trees are a hybrid of two machine learning algorithms: Random Forest concepts are mixed with single model trees. Decision trees known as "model trees" have each leaf containing a linear model that is optimal for the local subspace it describes. Random Forests, which produce tree variety through two randomization methods, have been demonstrated to significantly increase the performance of single decision trees. As in bagging, the training data is first sampled with replacement for every single tree. Second, a random subset of all characteristics is taken into consideration at each node when building a tree, and the best split for that subset is calculated rather than constantly calculating the best split for each node. These trees have been used for categorization. For the first time, random model trees merge random forests with model trees. In order to create properly balanced trees where a single global setting for the ridge value applies to all leaves, random trees use this produce for split selection, which streamlines the optimization process. (Data Mining: Practical Machine Learning Tools and Techniques, 2011 & Hall et al., 2009 & Wisaeng, K. 2013)

4.2 C4.5 decision tree

The C4.5 Algorithm, commonly known as J48 in Weka software, is a machine learning standard. In ID3 and C4.5, attribute selection is predicated on decreasing a node's information scale. A classification rule is represented by each path that leads from the root to a node. Reducing the number of tests that result in a new sample being categorized inside the database is the foundation of the idea. C4.5's attribute selection section is predicated on the observation that the decision tree's intricacy is strongly influenced by the quantity of data pertaining to that attribute. More information

is divided and separated by selecting that attribute than by selecting any other attribute. In addition to typical features in various numerical attributes, Algorithm C4.5 broadens the categorization domain. In essence, the algorithm creates the decision tree based on the property that has the greatest degree of separation across categories. The most crucial step in the C4.5 method is creating the first decision tree from the data set. Ultimately, the method generates a decision tree-shaped cluster with two different kinds of nodes. A decision node that runs tests on an attribute and generates a branch or subtree for each test outcome, and a leaf node that defines a category. Every subset of samples is constructed using a similar recursive tree construction process. Until samples from a single category are included in the subsets, this process is repeated. There are several steps involved in creating a tree. Finding the shortest decision tree from a sample dataset is a task that is regrettably NP-Complete. Thus, tree-building techniques ought to be irreversible and avaricious.

4.3 Random forest

The technique used for supervised learning is called Random Forest. This method produces a random forest, as the name implies. In reality, the artificial "forest" is a collection of "decision trees". The "bagging" approach is frequently employed to create a forest out of trees. The bagging method's basic thesis is that combining several learning models improves the model's overall performance. To put it simply, random forest creates a number of decision trees and combines them to get forecasts that are more reliable and accurate. Random forest has the benefit of being applicable to both classification and regression problems, which make up the bulk of machine learning systems today. The best of the subset of predictors selected at random from each node is used to divide each node in a random forest. Random trees are a forest, which is a group of tree predictors. The random trees classifier classifies the input vector of features using each tree in the forest and outputs the class label that earned the most "votes." This is how the classification process operates. The classification algorithm's response in a regression analysis is the mean of the answers from each tree in the forest.

4.4 Support vector machine (SVM)

One supervised learning technique for regression and classification is the support vector machine (SVM). The linear categorization of the data serves as the foundation for the SVM classifier, and we attempt to select a hyperplane with a greater confidence margin inside the linear split of the data. Linear programming techniques, which are well-known for tackling restricted problems, are used to solve the equation of determining the best line for the data. Prior to the linear division, we use the phi function to transfer the data to a much higher dimension space so that the machine can classify the data with great complexity. These techniques may be used to address the problem of extremely high dimensions by converting the desired minimization problem into its dual form using Lagrange's dual theorem, where a simpler function is used in place of the complex function phi that transports us to a high-dimensional space. We refer to the vector multiplication of the phi function as the kernel function (kernel). A variety of kernel functions, such as sigmoid, polynomial, and

exponential kernels, can be applied. V. N. Vapnik and A. Ya. Chervonenkis (Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia) created the main SVM algorithm in 1964. By adding the kernel class to SVM, Boser et al. (1992) demonstrated a method for nonlinear classification, while Cortes and Vapnik (1995) proposed soft-margin SVM.

4.4.1 Linear support vector machine

The test data set D consisting of n members (points) defined as follows:

$$D = \{(x_i, y_i) | x \in \mathbb{R}^p, y_i \in \{-1,1\}\}_{i=1}^{n}$$

where $y_i = -1$ or 1 and each x_i is a p-dimensional real vector. The goal is to find the separating hyperplane with the largest distance from the edge points that the points with $y_i = 1$ from the points with $y_i = -1$ separate. Each superplane can be a collection of points x that satisfy the following condition should be written as:

$$w.x - b = 0$$

where w is the normal vector, which is perpendicular to the hyperplane. The goal is finding w and b so that the maximum distance between the parallel superplanes separating the data is created. These hyperplanes are described using the following relation.

Any data above the separating hyperplane is marked with label 1:

$$w.x - b = 1$$

And any data at the bottom of the separating hyperplane with the label -1:

$$w.x - b = 1$$

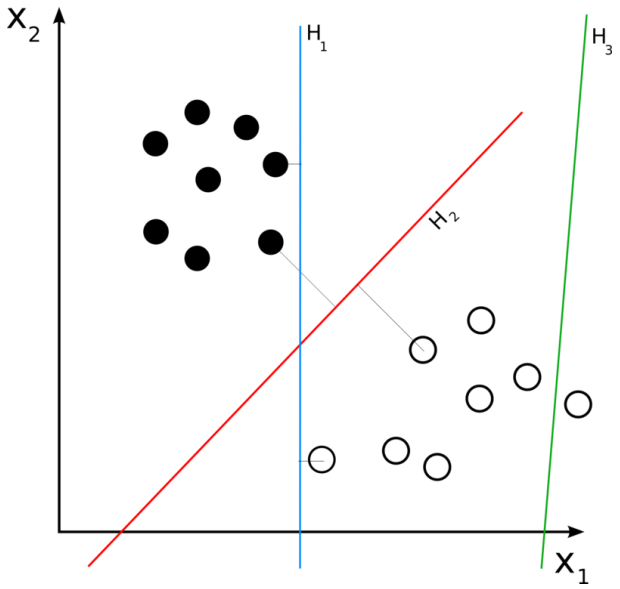

Figure 2. H3 (green) does not separate the two groups. H1 (blue) separates the two classes with a small margin, and H2 (red) separates the two groups with the maximum margin.

4.4.2 *Hard edge*

If the training data is a linear separation, we can have two planes completely at the edge of the points, consider and then try to space them. Using geometry, the distance between these two planes is$\frac{1}{\|w\|}$; so must minimize $\|w\|$. In order to prevent points from entering the margin, we add the following condition: for each i can be written as follows:

$$w.x_i - b \geq 1, \text{if } y_i = 1$$

$$w.x_i - b \leq -1, \text{if } y_i = -1$$

and

$$y_i(w.x_i - b) \geq 1, \forall 1 \leq i \leq n$$

Putting these two together yields an optimization problem:

$$\min_{(w,b)} \|w\|$$

$$\text{s.t. } \forall 1 \leq i \leq n$$

$$y_i(w.x_i - b) \geq 1.$$

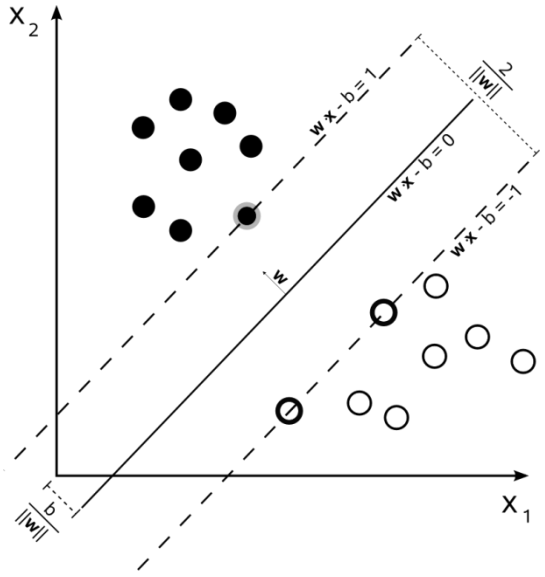

Figure 3. A maximum-margin hyperplane for a support vector machine learned with sample data from two categories. The data that are on the edge hyperplane are called support vectors.

4.4.3 *Dual form*

Using that $\|w\|^2 = w, w$ and substitution $w = \sum_{i=1}^{n} \alpha_i x_i y_i$ it can be shown that the dual SVM to the subject the following optimization is simplified:

$$\max_{\forall \alpha_i} (\tilde{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)) \text{ s.t. } \forall i, \alpha_i \geq 0, \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0.$$

Here is the kernel in the form $k(x_i . x_j) = x_i, x_j$ is defined. Phrase α forms a binary for the vector of weights of the training set:

$$w = \sum_i \alpha_i x_i y_i$$

4.4.4 Soft-margin

The use of the hard method can create limitations in some cases, one of which is the strong dependence of the separator on the border data, and if a data is placed on the border of the separating hyperplane, it will lead to a decrease in the margin and a drastic change in the desired hyperplane, this is an important point. That our data in reality has some noise in most cases and therefore the above problem can have an adverse effect on our decision boundary; therefore, hard-margin models are highly capable of overfitting.

To prevent and solve this problem, soft-margin was presented, which allows some data to be misclassified in the training process of the model and violate the set margin to avoid overfitting during testing. In other words, by doing this, a kind of generalization is given to the model so that it performs better on the test data; of course, this is done by adding a series of parameters and in a controlled manner so that the accuracy of the model does not decrease. These variables are called slack and are represented by the symbol ξ. For each data, these variables specify the extent of its violation of the determined boundary. If their value is slack, it means that the data is on the right side of the boundary, and if the value is they indicate the error value if it is greater than zero.

Next, all ξ's are added up and multiplied by the C factor and added to the expression we had in Hard-margin to get the new error expression, and we have to minimize the following expression to achieve our goal:

$$\min_{w,b,\xi} \frac{1}{2}||w||_2^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i \quad \xi_i \geq 0 \quad \forall i \in \{1, 2, \ldots, N\}$$

In error, the value of C indicates how much we allow the given model train time to be in the wrong class, i.e. the bigger the C, the less we allow the model to be wrong. Indiscriminate and uncontrolled reduction of C increases the bias and increases it and causes an increase in varicose veins. C is called regularization parameter.

4.5 Logistic regression

A statistical regression model for dichotomous dependent variables, like illness or health, death or life, is called logistic regression. Using the logit function as the link function and a polynomial distribution for its error, this model may be seen as a generalized linear model. One way to think of logistic regression is as a specific instance of both linear regression and the generic linear model. Unlike linear regression, the logistic regression model is predicated on entirely distinct hypotheses (regarding the connection between dependent and independent variables). Two features

of logistic regression show how these two models differ significantly from one another. First, because the dependent variable is binary, the conditional distribution $(y|\bar{x})$ is a Bernoulli distribution as opposed to a Gaussian distribution. The second is probabilistic prediction values, which are derived using the logistic distribution function and are restricted to values between zero and one. The output probability is predicted via logistic regression.

This model is;

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i},$$

This with a simple change will reach the logistics form as follows:

$$p = \Pr(y_i = 1|\vec{x}_i;\vec{\beta}) = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1+e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i})}}.$$

5. DATA ANALYSIS

In order to investigate the performance of these five supervised classification algorithms, real data from a cohort of cancer patients in Portugal have been used. In the next subsection, it will be explained about the cases used in this research. Also, in the second subsection, the performance results of these five methods in diagnosing the death of patients are compared. For this purpose, a number of indicators will be used. The software used in this research is Weka version 3.9.6.

5.1 Data Set Information

The HCC dataset, which was acquired from a Portuguese university hospital, includes various demographic, risk factor, laboratory, and overall survival characteristics of 165 actual patients who were given an HCC diagnosis. The EASL-EORTC (European Association for the Study of the Liver - European Organization for Research and Treatment of Cancer) Clinical Practice Guidelines, which represent the state-of-the-art in HCC care, were used to select the dataset's 49 characteristics..

There are twenty-six qualitative and twenty-three quantitative variables in this diverse dataset. In all, 10.22% of the dataset has missing data, and just eight patients have full information in every category (4.85%). One-year survival is the goal variable, and it was encoded as a binary variable with 0 denoting deaths and 1 denoting lives. There is also some class disparity (63 instances classified as "Dead" and 102 as "Alive").

A detailed description of the HCC dataset (feature's type/scale, range, mean/mode and missing data percentages) is provided in Santos et al (Santos et al., 2015).

5.2 Classification results

All five classification methods were used to predict the occurrence of death within one year after diagnosis and follow-up of these 165 patients. The predicted results were

Al-Furat Journal of Innovations in Management Sciences

ATU- AFJIMS, Volume: 1, Issue: 1, JAN. 2025

Table 1. Actucal and predictions for the first 17 patients in the data list

		Predicted				
inst#	Actual	Random tree	C4.5	Random forest	SVM	Logistic
1	Dead	Dead	Alive	Alive	Alive	Alive
2	Dead	Alive	Alive	Alive	Alive	Alive
3	Dead	Alive	Alive	Alive	Dead	Dead
4	Dead	Alive	Alive	Dead	Alive	Alive
5	Dead	Dead	Alive	Dead	Dead	Alive
6	Dead	Dead	Alive	Alive	Dead	Dead
7	Dead	Dead	Dead	Dead	Dead	Dead
8	Alive	Alive	Alive	Alive	Dead	Dead
9	Alive	Alive	Alive	Alive	Alive	Alive
10	Alive	Alive	Alive	Alive	Alive	Alive
11	Alive	Alive	Alive	Alive	Alive	Alive
12	Alive	Alive	Alive	Alive	Alive	Alive
13	Alive	Dead	Dead	Dead	Dead	Alive
14	Alive	Alive	Alive	Alive	Alive	Alive
15	Alive	Alive	Alive	Alive	Alive	Alive
16	Alive	Alive	Dead	Alive	Alive	Alive
17	Alive	Alive	Dead	Alive	Alive	Alive

As it can be seen, the algorithms have had relatively different results in diagnosing the patient's condition.

Table 2. The frequency of correct predictions of alive dead patients in two groups of patients who survived or died after one year.

Methods	Random tree		C4.5		Random forest		SVM		Logistic regression	
Prediction / Actual	Alive	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive	Dead
Alive	78	24	71	31	91	11	82	20	79	23
Dead	33	30	38	25	36	27	24	39	28	35
Correctly Classified Rate	65.4545 %		58.1818 %		71.5152 %		73.3333 %		69.0909 %	

It is evident that the SVM approach was the most effective classification technique for predicting the survival of patients with advanced liver disease, with a 73.3333 percent correct diagnostic rate. The random forest approach is then applied. Conversely, the random tree and C4.5 approaches have the lowest.

Table 3. Calculated values of comparision indices of classification algorithms

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Random tree	0.655	0.414	0.647	0.655	0.649	0.249	0.637	0.629
C4.5	0.582	0.489	0.573	0.582	0.576	0.095	0.536	0.547
Random forest	0.715	0.394	0.714	0.715	0.695	0.370	0.789	0.788
SVM	0.733	0.310	0.731	0.733	0.732	0.429	0.711	0.671
Logistic regression	0.691	0.361	0.687	0.691	0.688	0.336	0.663	0.633

The SVM method has outperformed all metrics, with the exception of the area under the ROC and POC curves, as the above table shows. Because it reduces the risk margin, this approach has proven to be highly effective in diagnosing. With its focus on large repetitions, the Random Forest approach has managed to find a circumstance that is comparatively appropriate. However, it appears that alternative methods are not appropriate for forecasting liver cancer patients' chances of survival..

6. CONCLUSIONS

As seen in the data analysis section, the SVM method has the best diagnosis and prediction of death among liver cancer patients with a one-year follow-up. This could be due to the maximum error avoidance method in this method. Minimizing the risk margins and staying away from the border of the two groups of data has provided good and accurate predictions in this method. The Random Forest algorithm, focusing on random tree iterations, has been able to provide good predictions, so that it has performed even better than SVM in two indicators. But other methods and especially the C4.5 algorithm have had a very poor performance in detecting the death of liver cancer patients.

REFERENCES

[1] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers.* Proceedings of the Fifth Annual Workshop on Computational Learning Theory.

[2] Cortes, C., & Vapnik, V. *Support-vector networks. Machine Learning*, 20(3), 273–297 (1995).

[3] Cutler, A., Cutler, D. R., & Stevens, J. R. *Random forests. Ensemble machine learning: Methods and applications*, 157-175 (2012).

[4] Data Mining: *Practical Machine Learning Tools and Techniques. (2011).* Morgan Kaufmann Publishers.

[5] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H., *The WEKA data mining software. ACM SIGKDD Explorations Newsletter*, 11(1), 10–18. (2009).

[6] Ng, A. Y., & Jordan, M. I., *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes.* Neural Information Processing Systems, 14, 841–848. (2001).

[7] Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., & Carvalho, A., *A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients.* Journal of Biomedical Informatics, 58, 49–59. (2015).

[8] Wisaeng, K., *A Comparison of Different Classification Techniques for Bank Direct* Marketing. (2013).