العنوان:	Development of important patterns in student database using statistical classifier
المصدر:	مجلة جامعة الأنبار للعلوم الاقتصادية والإدارية
الناشر:	جامعة الأنبار - كلية الإدارة والاقتصاد
المؤلف الرئيسي:	محمد، أنهار خير الدين
مؤلفين آخرين:	داؤد، سهير عبد(م.مشارك)
المجلد/العدد:	مج 4, ع 9
محكمة:	نعم
التاريخ الميلادي:	2012
الصفحات:	425 - 452
رقم MD:	401828
نوع المحتوى:	بحوث ومقالات
قواعد المعلومات:	EcoLink
مواضيع:	الجامعات والكليات ، التعليم الجامعي، طلبة الجامعات ، قواعد البيانات ، الإحصاء التحليلي ، التسرب في التعليم ، جامعة الموصل
رابط:	http://search.mandumah.com/Record/401828

Development of important patterns in student database using statistical classifier

م. أنهار خير الدين محمد   م. سهير عبد داؤد

كلية الإدارة والاقتصاد/ جامعة الموصل

تطوير أنماط هامة في قاعدة بيانات الطلاب باستخدام المصنفات الإحصائية

Abstract

The research aims to use some data mining techniques to predict the phenomenon of leakage of undergraduate students using a number of risk factors (important) (the gender , the attendance , the former grade of students , the educational level of parents , friends , first child , working) by using the closet nearest neighbor algorithm (KNN).

350 survey forms were distributed among the students of all the stages of the college of administration and economy department of administrational information systems. The forms contained 19 questions.

The variables of the forms and their data took the classification of (C4.5) and CART to compare the results and to choose the best to predict the rate of success and failure of the students who have failed before .

It has also been designed a database for the students of college of Administration & economics , with a computer system related on concerning with all registration affairs.

المستخلص:

يهدف البحث إلى استخدام بعض تقنيات تنقيب البيانات للتنبؤ بظاهرة تسرب طلبة المرحلة الجامعية باستخدام عدد من عوامل الخطورة (المهمة) (الجنس، الدوام، الدرجات السابقة للطالب، المستوى التعليمي للوالدين، لديه أصدقاء، أول طفل، يعمل) وباستخدام خوارزمية الجار الأقرب (*KNN*)

تم توزيع (٣٥٠) استمارة استبيان على جميع المراحل لطلبة كلية الإدارة والاقتصاد/ قسم نظم المعلومات الإدارية والتي تتضمن (١٩) سؤال وأخذت متغيرات الاستمارة وبياناتها كإدخال لخوارزميتي التصنيف (C4.5) و (CART) للمقارنة بين نتائج الخوارزميتين واختيار الأفضل للتنبؤ بنسبة النجاح والرسوب للطلبة غير المتسربين.

تم تصميم قاعدة بيانات خاصة بطلبة كلية الإدارة والاقتصاد/ قسم نظم المعلومات الإدارية، مع نظام حاسوبي متكامل وخاص بكل ما يتعلق بأمور التسجيل.

1.   Introduction

As a result Of rapid advances in the field of information technology there are huge information stored in databases that are in education and increasing rapidly and significantly , these databases contain a vast wealth of data , which constitute a potential mine of information value .

Due to changes in the structure of the databases emerges a new educational decisions in the ocean.

And find valuable information and hidden in these databases and the establishment of appropriate models and the distinguish them is difficult task. Technology has played an important role exploration data in each step of the discovery of information .

At the present time may be compiled from higher education and in the vicinity of competition high , which aims to get the benefits of competition , as these organizations must improve the quality of their services and have to satisfy customers who are students and teachers as they make up the source of useful and valuable essential as they wanted to prove indicators of their operations by means of use effective sources of power .

I have been using the nearest neighbor method (KNN) in this study for the purpose of discovery and analysis of the problem of drop out in the faculty of management and Economics/Department of Management Information Systems for the study sample , and resolution (CART) and (C4.5) to predict the rate of success and failure in the final exam for students who are not dropouts, one of the exploration data and techniques that can be used in many departments of education.

In this study , using data mining techniques to identify the cause of the leak of students and teachers to guide the intensification of the advantages and features of appropriate and related to students and monitoring them with financial aid.

2.   Objectives of the study

1- Building software for the computerization of administrative procedures in the management of student files the preliminary studies in the Faculty of Management and Economics adoption of the language (VB 6.0) and connection with two statistical software XLMiner 3 and Weka 3.6.0.

2- A software application with real data for students of primary studies in the faculty of Management and Economics and subjected to software designed to adapt to the nature of work in the unit register with the provision of information needs of the beneficiaries and of the reports and statistics required in this unit.

3- Predict the probability of the phenomenon of drop out for students and repeated using the nearest neighbor algorithm(KNN) .

4- Using the algorithm(CART) and (C4.5) to predict the rate of success and failure of non-dropouts and students choose the best algorithm to predict.

3.   The study sample

The College of Administration and Economic has been selected as one of the colleges that suffer from the phenomenon of students dropout their seats to the school where they were taking the information and data collection from registration unit in college .

4.   **Data Mining (الفخري، الدباغ، عبد العزيز,2006: 119-120)**

Data Mining is the effectiveness of access to knowledge to achieve the goal of which is the basic of the discovery of hidden facts contained in databases and through the use of multiple techniques include artificial intelligence . Statistical analysis , and data models , the process of data mining models and generate relationships and clear in the data , which helps to predict future results .

In general it can be argued that any of the following important relations in the field of data mining .

1- **Classes:** The commonly used to put the data stored in the field of data mining Identified in advance to build a model based on some independent variables .

2- **Clusters** : Used to put the data in the totals for the adoption of the logical relationships in other words, the algorithms used for classification in this way seeks to divide the data into groups so that the records are similar in the same group and that these groups must be different from each other as much as possible .

3- **Associations:** She knows the special relations data mining as the algorithms used to establish the rules to link the incidents that appear together in the data .

4- **Series Models** : The data mining to predict behavior and trends of models obtained .

The most widely used techniques in data mining is (الحمامي, 2008:34)

1- **Artificaial Neural Networks :** non-linear predictive modeling to learn through training and are installed in the biological neural networks .

2- **Decision Trees:** structures in the form of groups of trees are making these decisions generate rules for the classification of the data set contains methods for decision tree : classification and regression trees (CART) and automatic interaction detection to the Chi-Square (CHAID).

3- **Genetic Algorithms :** optimal techniques used treatments such as integration of genetic mating (conversion and testing natural (natural selection) in the design depends on the concepts of evolution .

4- **Nearest Neighbor Method** : techniques classifies each entry in a data set based on a combination of items to the limitations of (K) , which are more similar in the historic data set, where (K=l).
Sometimes called the nearest neighbor technique .

5- **Rule Induction :** also called extraction rules as adopted extraction rules (If- Then) on the statistical significance of many of these techniques, which use more than ten years of analysis tools are specialized and working directly with the mines of data (Data Warehouse) standard industrial and with the rules OLAP.

5.   **Nearest Neighbor Algorithm :**

is a method of predictive appropriate models for classification as a K number of cases is similar , or the number of elements in the group is the data training in the method of closest neighbor is the model not being built when it is providing a new case of the model searching algorithm in all data to find a subset (subset) of the cases that are more similar, and use it to forecast the output as shown in Figure(l) .

Training Group

droup out	work s	friends hip	Fir st chi ld	Educati on level of Parents	Previo us grades of student	alwa ys	sex	N o.
yes								1
-	-	-	-	-	-	-		-
								76

Learning Algorithm

Inductio n

Learnin g Moidel

Model

Application Model

conclusio --

Test group

droup out	work s	friend ship	Fir st chil d	Educat ion level of Parents	Previo us grades of student	alwa ys	sex	N o.
؟								1
؟	-	-	-	-	-	-		-
؟								51
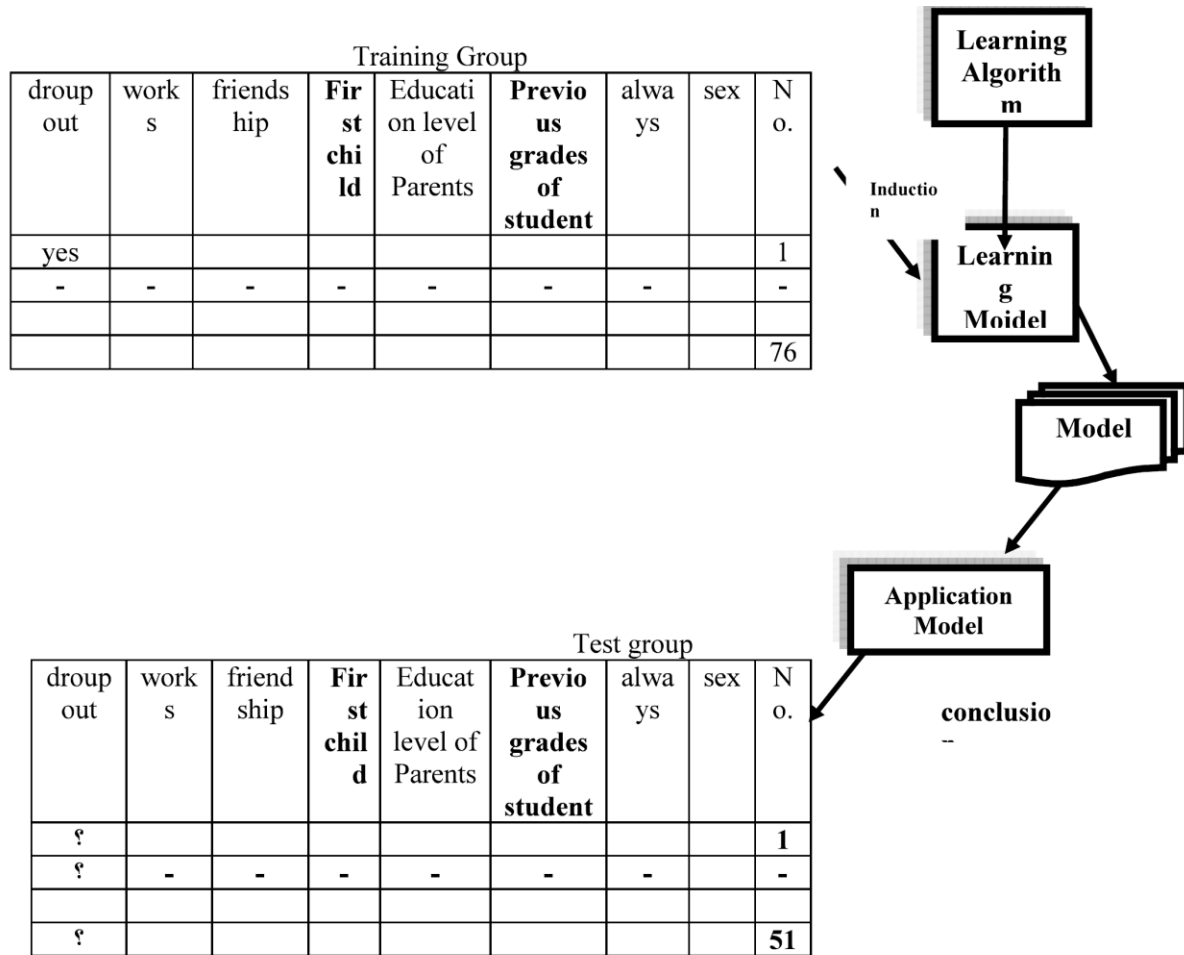
Figure(1) General method for building classification Model

There are key determinants in the nearest neighbor algorithm

1- Closest to the number of cases are used (K).
2- The unit of measurement (Metric) to measure the similarity.

Requires all use of the nearest neighbor algorithm that is specified as a positive value (K), and this determines the number of existing cases that are searched for when making a new case of the algorithm as an example of 4-NN indicates that the algorithm will use the four cases closest to the predicted output to the new situation Figure(2) neighbor, tow neighbors, three neighbors closest to the point of the data instance to be classified or based in the center of each circle

(c) 3 neighbor   (b) 2 neighbor   (a)1 neighbor
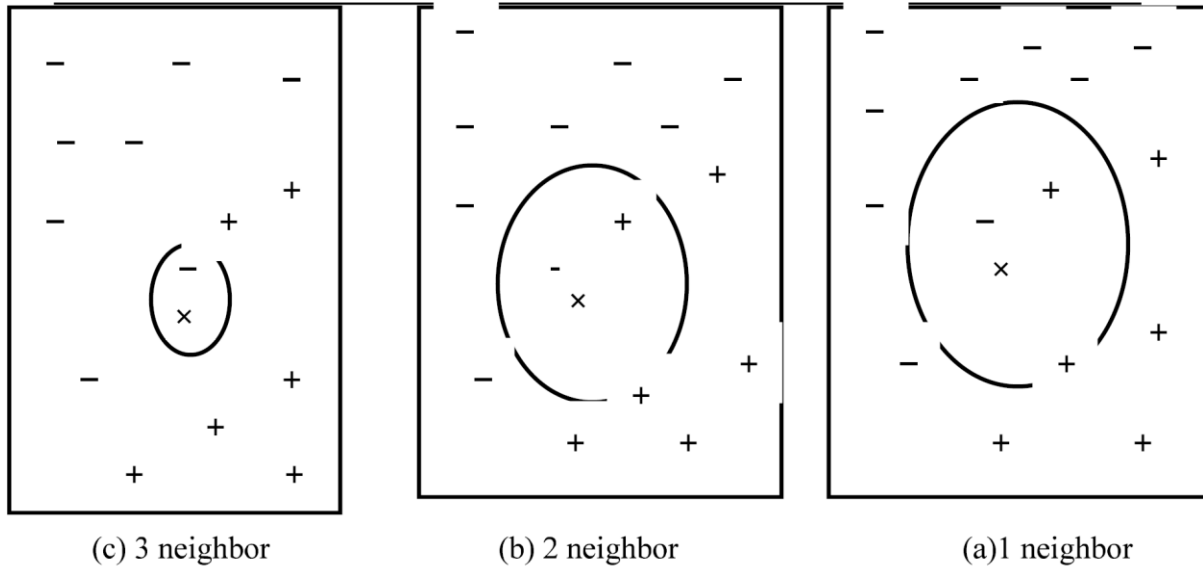
Figure (2) neighbor 1.2.3 nearest neighbor

If the neighbor closest to the point of data is the example with reference negative , as in Figure (2-a) , the point of the data that belongs to the class negative and the status of the number of neighbors is three , as in Figure(2-c) then has a neighbor (are two examples nearby and an example of a negative using the scheme of majority voting (Majority voting scheme) , the point data belong to the class the positive in the case of the presence of neighbors as in Figure(b — 2) are tested one grade randomly are classified as point data on the basis of being tested as to overcome the structures when they are a very small value due to the presence of noise(the values of abnormal) in the data training , but be very large who shall work the classification of examples (examples of the test) is wrong in the current study was to use this algorithm to detect leaks is done using the information on cases of leakage that were previously in order to identify students who drop now by selecting a number of experimental records and then use it to predict the required value **.(WU & Kumar,2009 :154)**

- **We can summary The Steps of nearest neighbor algorithm Chakraborty,2008: 10) (الطويل 2010 :233)(S & S,1993:2).**

The Algorithm calculate the distance(or similarly) between each test example z=(x',y') and each training examples(x,y) $\in$ D to detect the list of nearest neighbor Dz

1: let k be the number of nearest neighbors and D be the best of training examples.
2: for each test example z=(x',y') do
3: compute d(x',x), the distance between z and every example , (x,y) $\in$ D.

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{t=1}^{T} [x_{ti} - x_{tj}]^2}$$

4: Select Dz $\subseteq$ D the set of k closest training examples to z .
5: y =argmax $\sum_{(xi,yi) \in DZ} I(V=Yi)$
6: end for

* Risk factors that were used in the prediction and taken from the historical records for the student are :
- Sex
- Consistently
- The previous grades for students
- The educational level of parents
- The first child
- Friendships
- works

6- Decision Trees

Is a structure like a tree , as it carried out many tests to get to the best sequence (sequence) in order to predict what is needed in the sense that each test is either a cause other tests , or end node and paper (Leaf) and in the path from the root to the node of the paper target , the rule that classifies the target (predicted to be) and interpreted this rule (IF..Then) as in Figure(3).
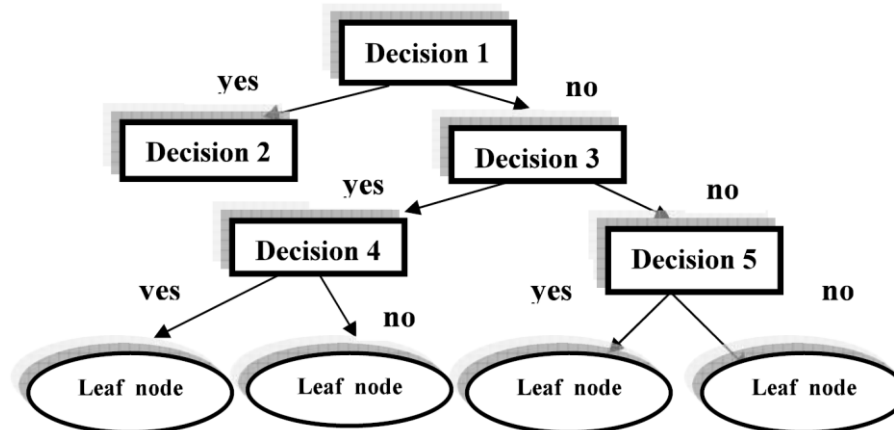
Figure(3) Decision Tree

Decision Tree starting from the root node and divided the data in each node to from new branches , even access to the node that can not be divided into any other branch (leaf node) traverse the tree from the root to the node of the leaf is the rule that best classifies the desired goal. (Rokach, 2008:1)

Example of a base (IF.. Then) are as follows :

If the decision is Yes, 1 = end node and the test leaf.

If the cam resolution 1 = does not apply to resolution 3 .

And so on up to the decision node(node resolution 4, or a decision node 5), and the process continues until the node determine the classification of leaf to the new situation.(خلوف،
رزوق، شميس, 2010: 90)

6-1 chain of the decision tree
1. building the tree
2. pruning
3. choose the perfect tree

1-   building the tree

Training data is fragmented in duplicate to become all the examples in each part leads to one class or section be small enough .

__And clarify the possible tree-building phase__(الحمامي, 2008 :167).

General growth of the tree algorithm (binary tree. Fragmentation(data S)
if (all point in (S) are of the same product) then go back .
Is (A) retail recipe
For each Class A Implemented .
Better use of fragmentation for the apportionment of S to SI and S2 .
SI
S2.

Form depends on the distribution of such section , we divide the attributes in the form of decisive A<= V where the V is a subset of all possible values of a . There are other experimental value of the recipe using a value comparable to the index .

Entropy (entropy(T) = - ∑Pj * LOG2 (Pj)
Gini index (gini(T)=l- ∑Pj2)
Where (Pj) is the frequency appropriate for the class j is T.

suppose that the index divisor is I(S) and parts divided (S) into S1,S2 , The best division is the division that maximizes the following value :

I(S) - |S1| / |S2| + I(S1) + |S2| | S |*I(S2)

2-   Decision Tree pruning ;-

Pruning is the removal of tree leaves and branches to improve the performance of the decision tree when moving from training data with known classification to applications in the real world, an algorithm to build a tree makes the best division in the root node, where the largest number of records and therefore there is a lot of information, where each division later the number of records becomes smaller and less representative with which to work toward the end .

__And objectives of pruning are :-__
Simplify the tree division .
Delete the sub tree. Replacement Securities.
Reduce the number of securities to the rules of the tree.
Tree subject to the understanding and decision Tree.

3-   Selection of Parts

The next step essential in the analysis of tree classification is tested parts of the variables predicted which are used to predict membership in the categories of the variables adopted for the cases or materials in the analysis due to the natural of the hierarchical tree classification , the parts are tested each part at a time starting from the section in the root node , and continuing with the parts construction of the node to the ends of the resulting retail, and construction node that is not divisible because they hold the ends , there are three types of selection methods :

1. Retail second variable-based discrimination , is used to determine the best two-node to the current fragmentation in the tree , and any prediction variable should be used to complete the process fragmentation .
2. Linear Retail merger approved on excellence: to predict the sequence variables (Prediction imposed be measured on the lower measurement period .This method works by dealing with the ongoing prediction, which consist of integrating a linear manner similar to the method of critical fasteners that have been deal with in the previous method, you use the hash value of the individual methods to convert continuous stabilizers to anew set of invariants is extra (duplicate).. (الحمامي **2008: 186**)

7-C4.5 Tree

Decision tree using the method of divide and conquer , as it is sufficient to divide the complex issue matters , and is the simplest recursion of the function of all parts of the same issue , can gather the parts to produce solution to resolve the complex issue(مختبر البرمجة/ الذكاء ,2008: 2)

Can be illustrated by the work of C4.5 decision tree the following steps **(WU & Kumar, 2009 :3), (Moertini, 2003:109).**

Input: an attribute-valued dataset D

    1: Tree = { }
    2: if D is "pure" OR other stopping criteria met than
    3: terminate
    4: end if
    5: for all attribute a € D do
    6: Compute information-theoretic criteria if we split on a
    7: end for
    8: abest = Best attribute according to above computed criteria
    9: Tree = Create a decision node that tests abest in the root
    10: Dv = Induced sub-datasets from D based on abest
    11: for all Dv do
    12: Treev = C4.5(Dv)
    13: Attach Treev to the corresponding branch of Tree
    14: end for
    15: return Tree

The C4.5 algorithm uses two types of pruning
1- Reduced error pruning:

It uses a separate test dataset, for but it directly uses the fully induced tree to classify instances in the test dataset. For every non leaf sub tree in the induced tree, this strategy evaluates whether it is beneficial to replace the sub tree by the best possible leaf, if the pruned tree would indeed give an equal or small number of errors than the un pruned tree and the replaced sub tree does not itself contain another sub tree with the same property, then the sub tree is replaced. This process until further replacements actually increase the error over the test dataset(Wu & Kumar,2009:7)

2- pessimist pruning :

Is an innovation in C4.5 that does nor require a separate test set rather it estimates the error that might occur based on the amount of misclassification in the training set. This approach recursively estimates the error rate associated with a node based on the estimated error rates of its branches. For a leaf with N instances and E errors(i.e The number of instances that do not belong to the class predicated by the Leaf), pessimistic pruning first, determines the empirical error rate at the leaf. As the ratio$(E+0.5)/N$.

For a sub tree with L leaves and $\sum E$ and $\sum N$ corresponding errors and number of instances over these leaves , the error rate for the entire sub tree is estimated to be $(\sum E+0.5*L)/ \sum N$ Now, assume that the sub tree is replaces by its best Leaf and that J is the number of cases from the training set that it misclassifies. .pessimistic pruning replaces the sub tree with this best Leaf if $(J+0.5)$ is within one standard deviation $(\sum E+0.5 * L)$
(WU & Kumar,2009 :7) (Oguz,2008: 13)

8- Classification and Regression tree(CART)

Called trees of resolution used to predict the variables critical trees classification (classification Trees) as she puts moments (Cases) in the classification or species trees, the resolution used to predict the variables continuous(continuous variables) is called (trees, regression), was described by (Olshen) and (Stone) in 1984. **(Kohavi &Quinlan, 1999 :8)**

And classification and regression trees are a type of decision tree algorithm and the pruning process that has been through the process of auditing and other technologies . Can be a tree derived from the database that contains hundreds of pages , variable answer with dozens of items recovered , be like this tree is difficult to understand despite the fact that each track to the paper is clear and understandable , in this sense , a decision tree describes the forecasts , which are of interest.

All regression techniques require the presence of variable and one or more of (predicted variables), the variable digital output.

Decision tree allows input variables to be a combination of continuous and categorical variables , the regression tree is created as each node in the decision tree contains the value of the test on some of the input variable , the final node of the tree contain variable and predicted production values .

Rules to stop the public is simply specify maximum depth which can grow to tree , the base limit, another option for the rules to stop is to prune the tree , allowing the tree to grow to full size and then using either a building extension or intervention of the beneficiary , is trim the tree back to the size smaller is not a rigorous process , for example , vacuum or sub - tree user not to feel important(Inconsequential) because they have very few cases you may delete the (CART) prune trees through cross-checking to see if the improvements contract extra precision balance . (الحمامي, **2008 :170**).

8-1 CART Algorithm

Decision tree using the method of divide and conquer, as they count the complex issue dividing them into questions of the simplest is then recursion of the function the same for each question, you can collect solutions to parts for the production of solving the complex issue , and this is the basic idea in algorithms based on decision tree such as the

The CART algorithmic (الفخري, 2003:15)

The Classification and Regression Tree (CART) algorithm can be summarized as follows :-

1. Create a set of questions that consists of all possible questions about the measured variables(Phonetic context) .
2. Select a splitting criterion(LikeLihood) .
3. Initialization : Create a tree with one node containing all the training data .
4. Splitting : Find the best question for splitting each terminal node. Split the one terminal node that results in the greatest increase in the LikeLihood .
5. Stopping : If each leaf node contains data samples from the same class , or some pre-set threshold is not satisfied stop , otherwise, continue splitting .
6. Pruning : Use an independent test set or cross-validation to prune the tree .

The CART algorithm uses what is known as **COST- COMPLEXITY PRUNING** where a series of trees are grown, each obtained from the previous by replacing one or more sub trees with a leaf. The last complexity is a metric that decides which sub trees should be replaced by a leaf predicating the best class value, each of the trees one then evaluated on a separate test dataset and based on reliability measures derived from performance on the test dataset, a "best" tree is selected **(Oquz,2008:32):-**

MINIMAL COST- COMPLEXITY PRUNING (X,K)

1   Input: **X** : a set of N labeled instances .

          K : maximum number of trees

2   **T**   ⟵—— LEARNER(X,attrs,0)

     ma

2   **T1 = T(αmin) where αmin = 0 :**

3   **R(T1) = R(Tmax)**

4   **For I⟵—— 1 to k**

       do

5   **for   t   Ti gi(t)=**

$$gi(t)= \begin{cases} \dfrac{R(t) - R(Tt)}{/Tt/ -1} & \text{if } t\ \check{T} \in\ Ti \\ \infty, & \text{else} \end{cases}$$

6   **Choose the weakest link ti   :**

       ti=argmin gi(t)

         t∈Ti

      gi(ti) = min gi(t)

         t ∈Ti

   and the set of weakest links { ti }

     {ti }={ t'i   : gi(ti) = gi(t'i) }

7   **αi+1   ⟵——— gi(t'i)**

8   **Ti+1   ⟵—— Ti − Tti , ti ∈ { t i}**

9 return The Sequence of Pruned Trees and Their Complexity parameters

     T1 > T2 >...> Tk   and {αi : αi+1 , K ≥1, α1=0}

     T(αi) = Ti , for I ≥1 , αi ≤ α <αi+1

9- system components

    Scheme(l) of Appendix(3) shows the main menus in the system screens of these statements set out in Appendix (4) - Appendix (10) .

Following is an explanation of the list of cases to predict the leakage and a list of prediction success and failure for prediction:-

1- prediction cases of leakage by using **K-nearest neighbor algorithm!KNN)** the screen shown in the Appendix (11)

• Training algorithm

The use of information on cases of leakage that has been previously by identifying a number of experimental records (210 records) has been used the number of different K (number of nearest neighbors) , but that was the most appropriate value of K is 5 , depending on the value (Root Mean Squared error RMSE ROOT). As shown in the following table:-

Table (1) :The Value RMSE in Each Value Of K

Value Of K	RMS Error
1	0.242535625
2	0.234724422
3	0.204506053
4	0.171408282
5	0.158326482

Best K

The best value (K) is less than the value (RMSE Error) where the RMSE is used to measure the accuracy of estimate (Chakraborty, 2008: 3) (Ki-Yeol,Byoung-Jin and Gwan-Su Yi,2004:7)

• Choice of the Algorithm

To illustrate the results of research has been taking a sample composed of 350 record has been used 210 of them for the purpose of training and 140 record for the purpose of testing and the results were a predictable drop out student or not and are described in Appendix (12) purified the real value (actual value) and value of forecast and (row id) is a student number at the same record number, and (Residual) Note the amount of error and sample prediction note that 100 students from 140 predicts them not to drop out and 40 predict them drop out. Note (1. Not to dropout, 2. dropout).

2- predict the rate of success and failure (A) The application of the algorithm C4.5

C4.5 algorithm was applied to a file (Test.txt) that contains the data for a sample of students obligatory to study the error rate in predicting the right and that the ercentage is correct (94.1176%), which indicates the prediction using the algorithm (C4.5). Was a high accuracy as the error rate for this prediction .
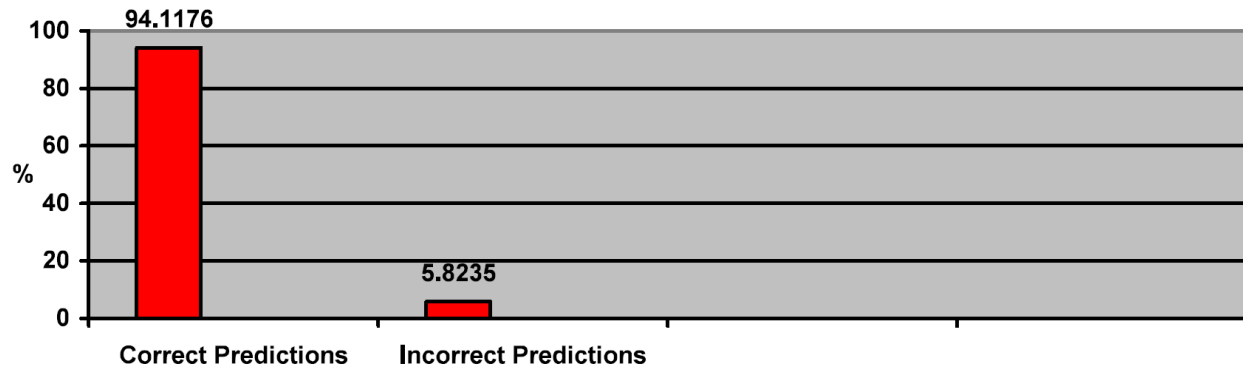
Figure(4) a histogram of the proportion of success and failure using an algorithm (C4.5).

(B) The use of Algorithm (CART)

The algorithm was applied (CART) on the file (Test. txt), which contains data on students, answers to the study sample the results of this algorithm is shown in the Appendix (14) . Figure(5) shows the percentage error in prediction and correct and that the percentage of correct are (91.1765%) indicates the prediction using the algorithm(CART) was a high accuracy as the error rate for this prediction is (6.8235%) which is small compared to the correct prediction.
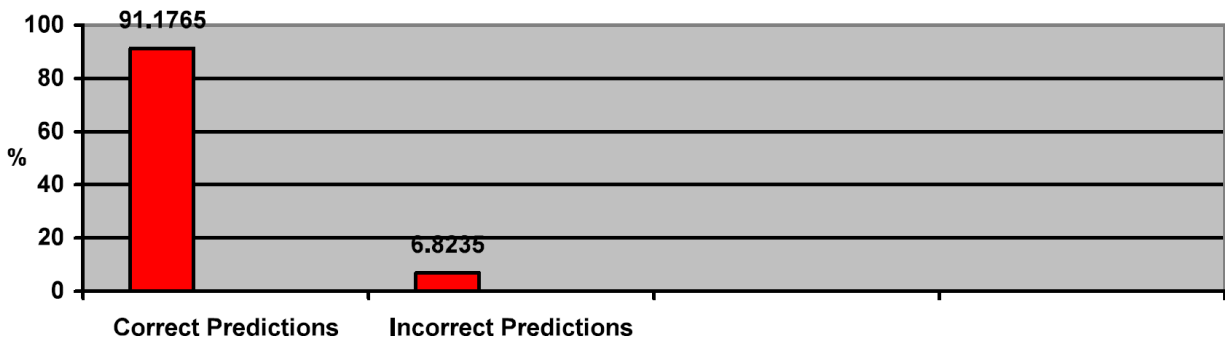
Figure(5) a histogram of the proportion of success and failure using an algorithm (CART)

And using the pruning algorithm (CART) and (C4.5) described in the appendixes (15 and 16) The form For pruning using an algorithm (C4.5) described in the Appendix(15) , The result described in the Figure (6)
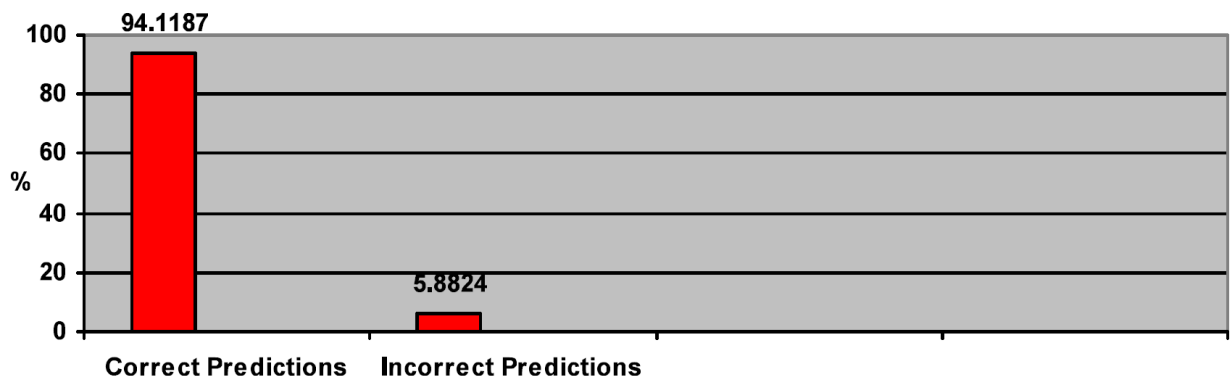
Figure(6) a histogram of the pruning using an algorithm (C4.5).
The form For pruning using an algorithm (CART) described in the Appendix(16) , The result described in the Figure (7)
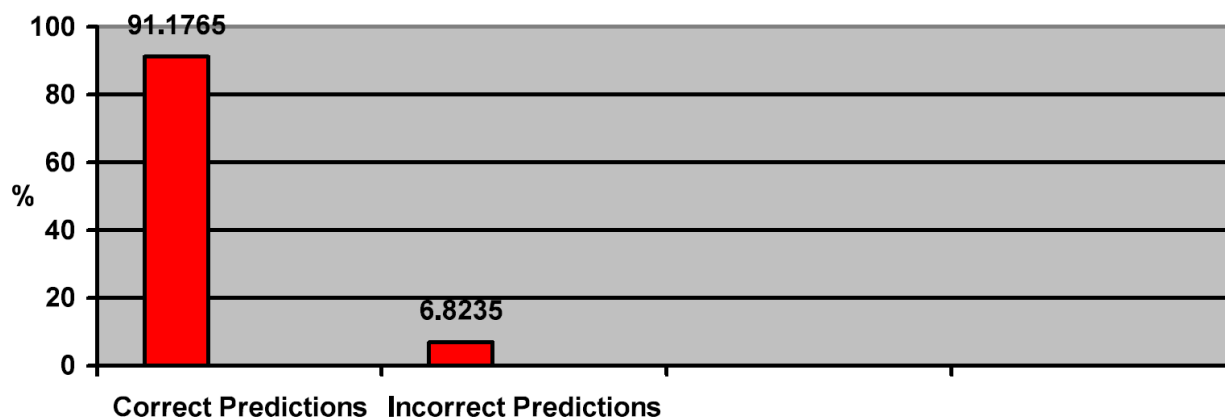
Figure(7) a histogram of the pruning using an algorithm (CART)

10-Conclusions
1. gives the system designer complete information about the march called for preliminary studies from the beginning to accept it and passing through the succession of years of study and the end of and all sections of the school of Management and Economics and in different classrooms and their teaching , and this without doubt will make it easier to complete follow -up process efficiently and provide to the departments of department and Dean and specialized reference a complete set for information of reports and statistics task .
2. characterized by the nearest neighbor algorithm
Training is very fast.
Can learn the functions of the target complex .
Do not lose information .
3. The results of the accuracy of prediction by your success and failure for students who drop out (Sample).

Table(2): Results of testing the accuracy of prediction

Algorithms	Correctly Classified	In Correctly Classified
C4.5	94.1176%	5.8824 %
CART	91.1765 %	6.8235 %

4. From the table(2) we observe the accuracy of the C4.5 algorithm is 94.1176 % compare with the accuracy of the Cart algorithm is 91.1765 % which mean the C4.5 algorithm is the best.
5. The (CART) where the classification in the simplest forms , as well as the solution for (C4.5) .
6. The (CART) possible to use different data types .
7. The (CART) used squat transfers fixed , where there is no need to use the logarithm of remittances , while the algorithm (C4.5) are used logarithm transfers to deal with the data .
8. The (CART) can use the same variable in different parts of the trees .

9. The (CART) is extremely robust to the effect of outliers , while the algorithm (C4.5) is finding it difficult to deal with the data of large size .

10. Use (C4.5) term is known in the information gain(Gain) [1] , through the application of the concept (entropy)[2] , while (CART) does not use it.

11. The (CART) Does not use the internal (Training data based) performance measure for tree selection Instead ,tree performance is always measured on independent test data(or via cross-validation)and tree selection proceeds only after test data based evaluation ,if testing or cross-validation has not been performed ,CART Remains agnostic regarding which tree in the sequence is best.

12. Algorithm (C4.5) generate models based on your training data on the reverse algorithm (CART) that do not use training data to measure the internal efficiency of the selected tree .

13. Algorithm (CART) used only two divisions in order to restrict the conditions of the test , the algorithm (C4.5) using a standard known as divided (Gain ratio) to determine the goodness of (division) .

14. Can be represented by a decision tree (C4.5) in a different way after the pruning process using rules(IF...Then) police and described in Appendix(15) , as in the following steps

IF Q18 > 1 Then
IF Q19 <= 1 and Q10 > 1 Then
Classification = Yes ;
Else
Classification = No ;
ElseIf Q18 >1 and Q19 > 1 Then
Classification = Yes ;
Else
Classification = No ;

   Clear from this that the contract that was which was to reduce the accuracy of the tree is :-
Q1,Q2,Q3,Q4,Q5,Q6,Q7,Q8,Q9,Q10,Q11,Q12,Q13,Q14,Q15,Q16,Q17.

   The questionnaire **of** the best contract that has been picked up in the tree , which was the basis for access **to(Optimal tree)** is **(Q10,Q18,Q19).**

Can be represented as well as the decision tree (CART) in a different way after the pruning process using rules **(IF...Then)** conditional and described in Appendix(16) and also in the steps following code :-

IF Q18 < 1.5 Then Classification = No ;
IF Q18 >= 1.5 Then

---

[1]Gain is the number of bit saved , on average ,if we transmit Y and both receiver and sender X

Gain=Entropy(x)-Entropy(x|y)-Entropy(x)- $\sum_{yeY} \frac{|S_Y|}{S} Entropy(S_{Y})$

[2]Entropy is a measure of how pure or impure a variable is Entropy(s)= $-\sum_{j=1}^{m} P_j \log_2 P_j$

IF Q19<1.5 and Q10 <1.5 Then Classification = No ;
Else
Classification = Yes ;
ElseIf Q18 >= 1.5 and Q19 >= 1.5 Then Classification = Yes ;
From this it follows that the most influential characteristics in determining the success or failure of the years the student is failing, and the years of the previous download and the time of the study , which represents **(Q10,Q18,Q19) respectively**

15. For groups of properties (C4.5) produces a branch of the division of each value in the totals characteristics .

16. (C4.5) is different from (CART) to measure the homogeneity of the node .

المصادر:

المصادر العربية

١. الحمامي، علاء حسين، (٢٠٠٨)، تنقيب البيانات، ط١، إثراء للنشر والتوزيع، عمان، الأردن.

٢. خلوف، فادي، رزوق راكان وشميس آصف (٢٠١٠)، تطوير آليات جديدة للتنقيب في المعطيات لإدارة علاقات الزبائن في بيئة مصرفية، مجلد ٢٦، العدد ١، مجلة جامعة دمشق للعلوم.

٣. الفخري، عبد الله قاسم وغيداء عبد العزيز الطالب (٢٠٠٣) استخلاص نموذج بياني من قاعدة بيانات باستخدام خوارزميتي K-means و IBK، رسالة ماجستير، كلية علوم الحاسبات والرياضيات، جامعة الموصل.

٤. الطويل، هالة (٢٠١٠)، "التنقيب عن البيانات"، شعاع للنشر والعلوم، حلب، سوريا.

٥. عبد العزيز، غيداء، الدباغ، رائد عبد القادر والفخري، نعمة عبد الله (٢٠٠٦)، استخدام المصنف C4.5 في تمييز سمة الكائن– دراسة مقارنة.

٦. مختبر البرمجة/ الذكاء الاصطناعي، (٢٠٠٨)، "خوارزمية C4.5"، كلية الهندسة الكهربائية والإلكترونية، جامعة حلب، قسم هندسة الحواسيب.

المصادر الأجنبية:

7. Chakraborty, Dipanjan(2008), K-Nearest Neighbor Learning .

8. Ki-Yeol Kim,Byoung-Jin Kim and Gwan-Su Yi(2004), Reuse of imputed data in microarray analysis increases imputation efficiency, BMC Bioinformatics,Vol.5 .

9. Kohavi,Ron & Quinlan,Ross(1999),Decision Tree Discovery.

10. Moertini,Veronica S.,(2003),Towards the use of C4.5 algorithm for classifying banking dataset,INTEGRAL,Vol.8, No.2

11. Oquz,Gizil (2008),Decision Tree Learning For Drools

12. Rokach, Lior(2008),Decision Tree,Department Of Industrial Engineering ,Tel-Aviv University,PP: 1

13. S.Cost,S.Salzberg(1993),A weight nearest neighbor algorithm for learning with symbolic features machine learning .

14. Wu,Xindong , Kumar,Vipin,(2009),The Top Ten Algorithms in Data Mining,CRC Press,Taylor & Francis,Group Boca Raton, London.

Appendix (1)

جامعة الموصل

كلية الإدارة والاقتصاد

قسم نظم المعلومات الإدارية

المرحلة: الأولى [ ]   الثانية [ ]   الثالثة [ ]   الرابعة [ ] .

العمر: [ ] .

الجنس: ذكر [ ]   أنثى [ ] .

س١/ هل تتناول الدواء	نعم [ ] .	لا [ ] .
س٢/ هل تتعاطى الكحول	نعم [ ] .	لا [ ] .
س٣/ هل تؤدي الواجبات اليومية والمنزلية	نعم [ ] .	لا [ ] .
س٤/ تقديرك لحالتك النفسية	سيئة [ ] . متوسطة [ ] .	جيدة [ ] .
س٥/ هل دخلت دورات (دراسية خارجية)	نعم [ ] .	لا [ ] .
س٦/ الجو الأسري	سيئ [ ] . متوسط [ ] .	جيدة [ ] .
س٧/ دخل الأسرة	ضعيف [ ] . متوسط [ ] .	عالي [ ] .
س٨/ هل تمتلك حاسبة شخصية	نعم [ ] .	لا [ ] .
س٩/ حجم الأسرة	صغير [ ] . متوسط [ ] .	كبير [ ] .
س١٠/ وقت الدراسة	أقل من ساعتين [ ] . ٢ – ٥ ساعة [ ] .	٥–١٠ ساعة [ ] .
س١١/ هل لديك خط إنترنت في البيت	نعم [ ] .	لا [ ] .
س١٢/ هل تخرج مع الأصدقاء	نعم [ ] .	لا [ ] .
س١٣/ هل تأخذ راحة في عطلة نهاية الأسبوع	نعم [ ] .	لا [ ] .
س١٤/ هل أنت من سكنه الموصل	نعم [ ] .	لا [ ] .
س١٥/ الحالة الصحية	ضعيفة [ ] . متوسطة [ ] .	جيدة [ ] .
س١٦/ الخدمات المدنية في منطقتك	ضعيفة [ ] . متوسطة [ ] .	جيدة [ ] .
س١٧/ هل تصل إلى المنزل متأخرا	نعم [ ] .	لا [ ] .
س١٨/ هل لديك سنوات رسوب	نعم [ ] .	لا [ ] .
س١٩/ هل لديك تحميل من سنوات سابقة رسوب	نعم [ ] .	لا [ ] .

APPENDEXE (2)

Appendix(3)
Scheme(1)

Components System

- Record Absence
- Report & Static
  - -Report about Transfer a school curriculcem for the student
  - - List the names of student by departments & Stages & The People
  - - Report For the Percentage Of Student absent 5%
  - - Report For the Percentage Of Student Absent 10%
- Prediction Of Dropout
  - KNN Algorithm
- Prediction Of Dropout
  - J48b Algorithm
  - Algorithm. CART

- Data Processing
  - - Search of particular student within a particular Department & section
  - - Search of a particular student by using the first 3 character of student's name
  - - update File graduation
  - -Punctuate student's record
  - - Return the student registered Almrgen
  - - Distributed the student to the Classes .

- Other Events
  - Update Informatio n
  - Delete Record

- Registration &
  - Transfer Student
  - Add New Student
    - -personal Information
    - -Scientific Information
  - Deport Stage
    - Pass
      - Final Stage → File Graduation
      - Not Final → Next Stage
    - Fail
      - Final Stage → Stay in the same Stage
      - Not Final → Stay in the same Stage
    - Deffered Year
      - Stay in the same Stage

- -Methodological materials to be studied at the College Level
- -Input Coll.Name & Scientific Department
  - –The introduction of Sections & methodological materials

Appendix(4)

Personal Information & scientific student Form

Appendix(5

Transmission students Form

Appendix (6)
Leave Records student Form
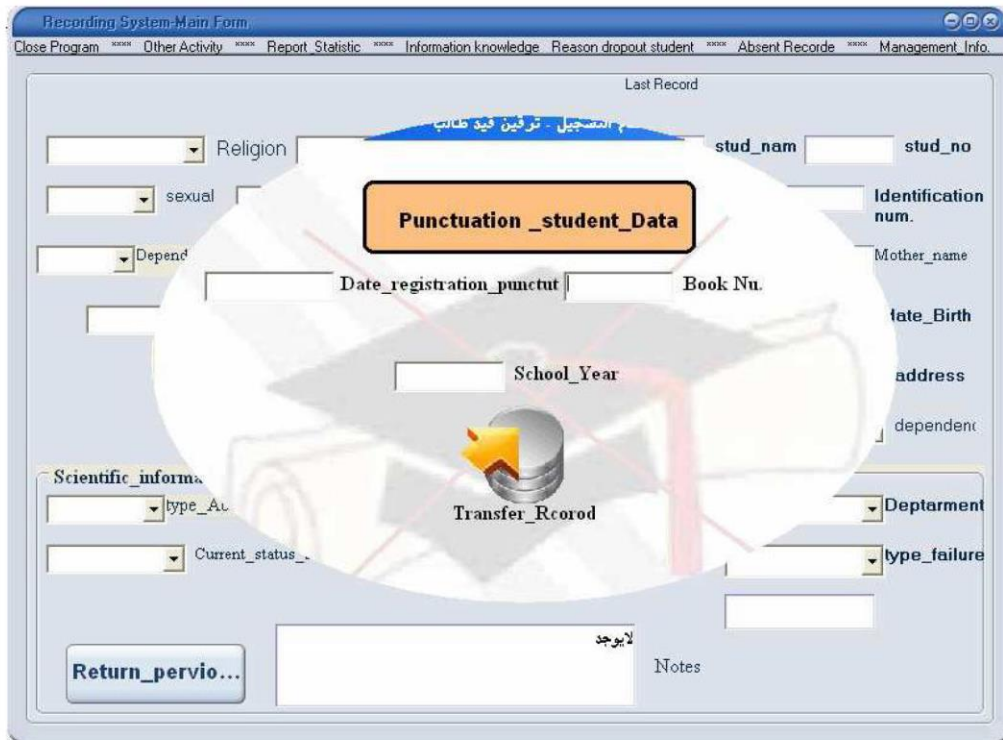
Appendix (7)
Distribute students to the Class Form

Appendix (8)
History special student Report

Appendix (9)

List of the fourth MIS student's group (1) order by alphabetic student's name

Appendix(10)

List of the third MIS students name who rate absence 10%

Appendix(11)
Prediction of drop out students by using KNN Algorithm

Appenix(12)
Test operation Result

Appendix (13)
Predication Form by using C4.5 Algorithm
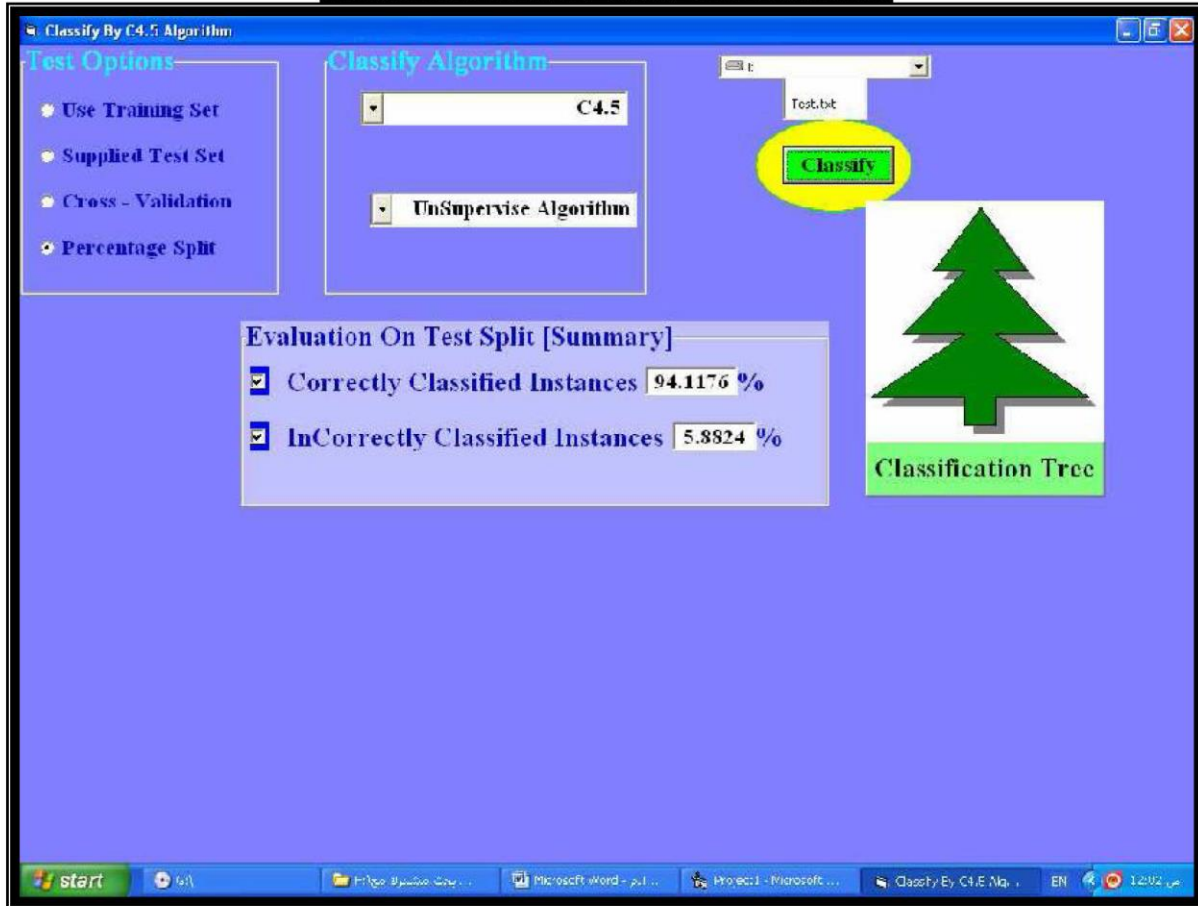
Appendix (14)
Predication Form by using CART Algorithm
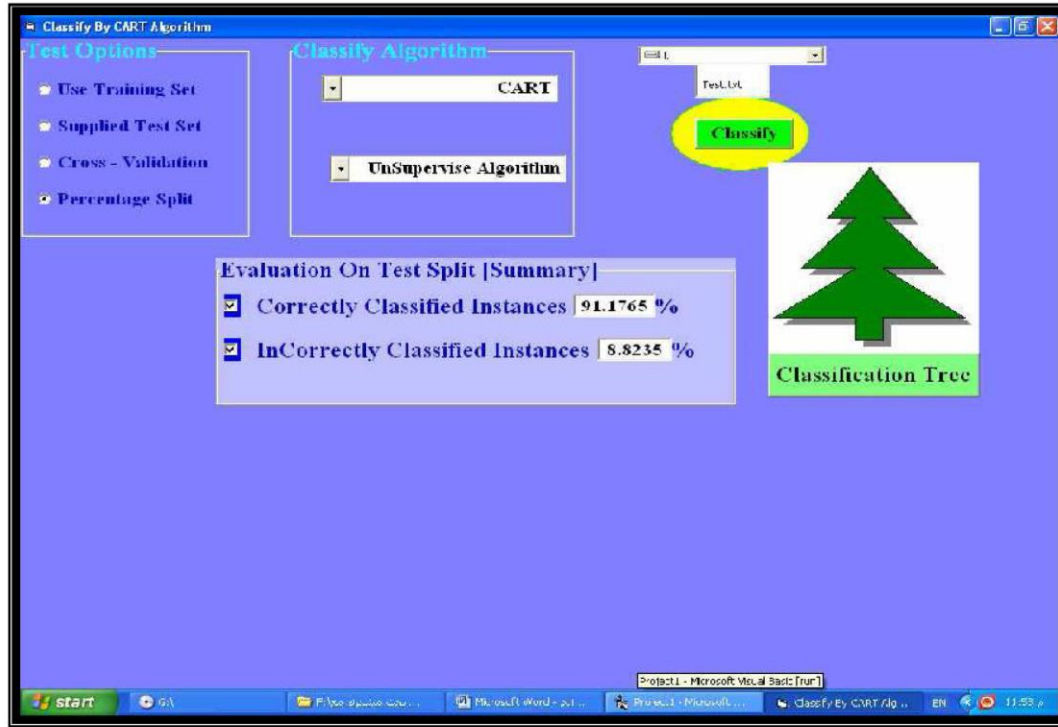
Appendix(15)
Pruning form by using C4.5 Algorithm

Appendix(16)
Pruning form by using CART Algorithm