High Cardiovascular Risk Models for a city in the Central Region of Cuba

Gladys M. Casas-Cardoso[1], Jorge Luis Morales-Martínez[1], Lisset Denoda-Pérez[1], Leidys Cabrera-Hernández[1], Lucía Argüelles-Cortés[1], Emilio González-Rodríguez[1]

[1]Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba
{gcasas, jmm, ldenoda, leidysc, largue, eglez}@uclv.edu.cu

Abstract

This research is based on the study of a real database, in order to determine the factors of the High Cardiovascular Risk in the city of Santa Clara, in the Central Region of Cuba. Mathematical models of the high risk index, considering the characteristics of the population of the city are obtained. Four multivariate statistical techniques are applied to obtain the mathematical models: Linear Discriminant Analysis, Logistic Regression, Classification Trees and a Fuzzy Logic based model. Diversity measures are calculated for the base classifier. As a result, a set of classifiers are combined into a multiclassifier. This model is better than the others. The statistical package SPSS, Microsoft Excel and EFuzzy software were used to perform all the calculus.

Keywords: High Cardiovascular Risk, Fuzzy Logic, multiclassifier, diversity measures

1. Introduction

Over the years, man has used different branches of science to control and improve processes. An example of this is the Statistics. Moreover, computer technology available today has made possible to obtain extraordinary advances in the data analysis, which is reversed in remarkable advances in Medicine, Bioinformatics and production among many other areas.

Medical professionals should base a large part of its requirements in the quantitative analysis results, but some of their decisions are based on diagnoses that are associated with the experience and intuition of the experts themselves, which have a subjective nature, so it is appropriate for applying the fuzzy logic.

In Cuba, the specific conditions that characterize the economy and society, the limitations of material resources, changes in risk profiles, morbidity and mortality from transmissible diseases and its implications on the health of the population, have shown the need to improve surveillance systems at all levels.

Epidemiological surveillance and prevention and control programs in health depend largely on the knowledge of the risk factors associated with diseases. To do this, epidemiologists have traditionally performed case-control studies or cohort studies and contingency techniques applying the Mantel-Haenszel's methodology [1]. This analysis is necessarily supplemented with multivariate studies that are performed with standard techniques such as linear discriminant analysis (LD) and logistic regression (LR), or more modern methods, derived from Artificial Intelligence such as Classification Trees (CT) and others based on Fuzzy Logic (FL).

The classification problem is one of the first to appear in the scientific activity. It is a process inherent in almost any human activity, so that in problem-solving and decision-making, the first part of the task is precisely to sort the problem or situation, and then to apply the methodology. The same happens with the Medicine, science in which the diagnosis is a major part, being a prerequisite for the implementation phase of therapy. In Medicine, the diagnosis is equivalent to classifying a subject in a particular disease based on its recognition data, scan and tests. When it comes to classifying a subject in a particular group, from the values of some parameters measured or observed, and this classification has some degree of uncertainty, it is reasonable to consider the use of a probabilistic methodology which allows quantifying this uncertainty [1].

In a real study there are often multiple variables (predictors or independents) that can be associated with a dependent variable. Presenting many contingency tables, for example, by analyzing the case in which the variables are discrete, [2-3] it does not always reflect essential associations, and it usually becomes a huge list of tables difficult to interpret, even when Cramer's V is used to sort the strength of the associations [4]. A multivariate analysis approach is the effect of all possible variables, including possible correlations together, but it can be particularly interesting, considering also the possibility of interaction among the predictor variables that do not necessarily have to be all discrete on the dependent variable [5-6]. When the number of variables grows, the set of possible interactions grows too much, it is then almost impossible to analyze them and therefore it is of special interest an automatic detection technique of fundamental interactions [7].

In this context, the aim of this work is to obtain several mathematical models describing the rate of cardiovascular risk [8]. Besides, some diversity measures are used in order to select the better set of classifiers to be combined.

The paper is structured as follows: next section is dedicated to do an initial data analysis. Cases and variables used in the research are introduced. Then, there are a section dedicated to present the mathematical models to diagnose the High Cardiovascular Risk: Linear Discriminant Analysis, Logistic Regression, Classification Tress and a Fuzzy Logic based models. Afterwards, another section explains some diversity measures. The main ideas to build a multiclassifier better than individuals models are also presented. Finally, results and conclusions are provided.

2. Initial Data Analysis

The database used in this research was provided by PRODEC project of the UCLV, Cuba [9]. The sample has 849 cases, which were classified by an Expert Committee of Medical Doctors as: 220 patients with elevated values of arterial tension (hypertenses), 219 with moderated values of arterial pressure (pre-hypertenses) and 410 with normal values (normotenses). The random sample was selected from the supposedly healthy individuals of the city of Santa Clara. Table 1 shows the discrete variables used in this analysis [10].

TABLE 1 Predicted Discrete Variables used in the analysis

Variable	Values	Percentage
Smoking	Present (Yes) Not present (No)	38.7% 61.3%
Mellitus Diabetes (MD)	Present (Yes) Not present (No)	10.3% 89.7%
Hypertense Diagnostic (HypDiag)	Hypertense Pre-hypertense Normotense	25.9% 25.8% 48.3%
High Cardio-vascular Risk (CHRisk) Dependent Variable	High Not high	4.8% 95.2%

High Cardiovascular Risk (CHRisk) is the dependent variable. Its two categories are: "high" and "not high". Each patient in the analysis was diagnosed by a High Qualified Expert Committee composed by relevant Medical Doctors belonging to the PRODEC project [9]. It was very important for these specialists to obtain a good mathematical index of High Cardiovascular Risk. That is the main objective of this investigation.

Table 2 illustrates the minimum and the maximum values for the predicted continuous variables used in the research.

TABLE 2 Predicted Continuous Variables used in the analysis

Variable	Minimum	Maximum
Age	18	78
Body Mass Index (BMI)	15.5	44.5
Systolic Basal Pressure (SBP)	80	220
Diastolic Basal Pressure (DBP)	50	130
Systolic Pressure after 1 minute of stress (SP1)(mmHg)	80	230
Diastolic Pressure after 1 minute of stress (DP1)(mmHg)	48	140
Average Arterial Pressure (AAP)	66.7	170
Triglycerides (mmol/L)	0.42	7.95
Total Cholesterol (TChol) (mg/dL)	88.9	421.5
High Density Lipoproteins Cholesterol (HDLChol) (mg/dL)	13.9	270.7

The first problem emerges when we analyze the data, because the two classes are very unbalanced. Traditional classification problems do not work appropriately with this unbalance. Table 3 shows the initial data:

TABLE 3 Unbalanced data

CHRisk categories	Frequencies	Percentage
Not high	808	95.2
High	41	4.8
Total	849	100

As it can be seen, less than the 5.0% of the sample has High Cardiovascular Risk. Numerous papers solve the problem of the unbalanced classes [11-13]. For this research we decided to compute a weight variable to eliminate the unbalance problem.

The "Not high" class is almost 20 times (808/41=19.7) higher than the "High" class. Therefore, we decide to use a weight variable with the value of 20 for the high category of CHRisk and the value 1 for the not high category. Table 4 shows the same data, now balanced.

TABLE 4 Balanced data

CHRisk categories	Frequencies	Percentage
Not high	808	49.60
High	820	50.40
Total	1628	100

With the weighted data we obtain most of the mathematical models.

3. Mathematical Models

In this topic, the more relevant results of the mathematical models are presented. All the calculus was done with SPSS, Excel and the EFuzzy software. [14-15]

3.1 Linear Discriminant Model

Linear Discriminant analysis is a multivariate technique to solve a classification problem. It works with data that is already classified into groups to derive rules for classifying new (and as yet unclassified) individuals on the basis of their observed variable values. Fisher's suggestion was to seek a linear transformation of the predictor variables such that the separation between the group means would be maximized. The dependent variable is CHRisk and the predictor variables are the others defined in tables 1 and 2 [5-6, 16-18].

Wilks lambda was calculated and its p-value was 0.000. Due to this we reject the fundamental hypothesis and we conclude that the groups come from different populations. Table 5 shows the standardized coefficients of the discriminant function.

TABLE 5 Standardized Linear Discriminant Model

Predicted	Function
Age	0.783
SBP	0.638
Smoking	0.412
MD	0.497
HDLChol	-0.200
BMI	-0.177
SP1	-0.397
Triglycerides	0.113
HypDiag	0.188
AAP	0.556
DBP	-0.246
TChol	0.079

The structured matrix shows the correlation between each variable with the discriminant function. Age is the more important variable; notice its high correlation (0.689) in table 6.

TABLE 6 Structured Matrix of LD Model

Predicted	**Function**
Age	0.689
SBP	0.397
Smoking	0.392
MD	0.319
HDLChol	0.315
BMI	-0.289
SP1	0.286
Triglycerides	0.267
HypDiag	0.221
AAP	0.118
DBP	-0.115
TChol	0.098

Finally, table 7 presents the classification results of the model. The accuracy represents the percentage of correctly classified cases. Its value is 94.5%.

TABLE 7 Classification table for the LD model

		Predicted CHRisk using LD model	
		Not high	High
Real CHRisk	Not high	88.9% (718)	11.1% (90)
	High	0% (0)	100% (820)
Accuracy = 94.5%			

Table 8 summarizes the analogous information for the original data, without weight cases. Observe that the accuracy is now 96.1%.

TABLE 8 Classification table for the LD model with real data

		Predicted CHRisk using LD model	
		Not high	High
Real CHRisk	Not high	97,6% (789)	2,4% (19)
	High	34,1% (14)	65,9% (27)
Accuracy = 96.1%			

3.2 Logistic Regression Model

Binary Logistic Regression (LR) is a statistical multivariate technique. It is useful for situations in which we want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model. LR is applicable to a broader range of research situations than discriminant analysis.

Logistic regression does not rely on distributional assumptions in the same sense that discriminant analysis does. The dependent variable should be dichotomous. Independent variables can be continuous or categorical; if categorical, they should be dummy [5, 19-20].

The purpose of the analysis is to predict the probability of the occurrence of an event: high cardiovascular risk or not. All the calculus was done with the SPSS software. Table 9 shows the obtained LR model for our data [6, 16-18].

TABLE 9 LR Model

Variables	B	Exp(B)
Age	0.376	1,457
SBP	0.024	1,024
Smoking (1)	-4,079	0.017
MD(1)	-5,456	0.004
HDLChol	-0.073	0.930
BMI	-0.180	0.835
SP1	-0.021	0.979
Triglycerides	0.871	2,388
HypDiag(1)	-5,600	0.004
HypDiag(2)	-3,557	0.029
AAP	0.219	1,245
DBP	0.074	1,076
TChol	0.008	1,008
Constant	-36,16	0.000

Table 10 explains the final results of the classification process. The LR model is better than the LA because of its accuracy 96.9%, which is superior to 95.4%.

TABLE 10 Classification table for the LR model

		Predicted CHRisk using LR model	
		Not high	High
Real CHRisk	Not high	95.7% (757)	4.3% (51)
	High	0% (0)	100% (820)
Accuracy = 96.9%			

Table 11 shows the classification values without the weight variable, that is the results with real data.

TABLE 11 Classification table for the LR model with real data

		Predicted CHRisk using LR model	
		Not high	High
Real CHRisk	Not high	99.5% (804)	0.5% (4)
	High	36.6% (15)	63.4% (26)
Accuracy = 97.8%			

The accuracy is very good: 97.8%, but 15 patients with high CHRisk are classified as not high. This result is not so good.

3.3 Classification Tree Model

The Classification Tree procedure creates a tree-based model. It classifies cases into groups or predicts values of a dependent (CHRisk) variable based on values of predictor variables [2, 7, 16-17, 21]. We used the Classification and Regression Trees (CRT) method to obtain the tree. CRT splits the data into segments that are as homogeneous as possible with respect to the dependent variable. A terminal node in which all the cases have the same value for the dependent variable is a homogeneous, "pure" node.

Figure 1 show the tree model; notice that the most important variable is Age.

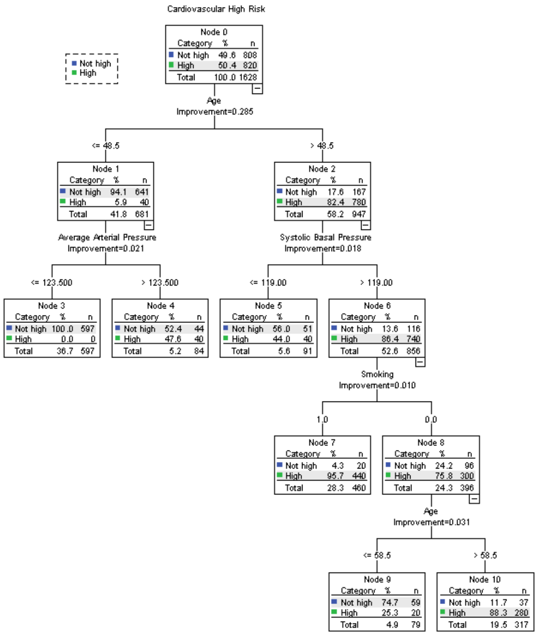

Fig. 1: Classification tree model with the CRT method

The root node has the total of cases: 50.4% represents high cardiovascular risk patients and 49.6% correspond to no high risk cases. Remember that cases are weighted.

The first variable in the tree is Age. It is the more important predictor of the high risk.

Now we will explain the information of the six terminal nodes.

Node 3: Subset composed by 597 young patients. Their age is inferior to 49 years old and their average arterial pressure is lower than 123.5 mmHg. All are healthy people.

Node 4: Subset composed by 84 patients. They have less than 49 years old and their average arterial pressure is higher than 123.5 mmHg. The 52.4% of the total does not have high cardiovascular risk. The two percents are quite similar, and then the group is risky.

Node 5: Subset composed by 91 patients with their age superior to 49 years old and with systolic basal pressure lower than 119 mmHg. The 56% of these patients does not have high cardiovascular risk.

Node 7: Subset composed by 460 patients with age higher than 49, systolic basal pressure higher than 119 mmHg and smokers. Almost all patients of this group (95.7%) have high cardiovascular risk. The interaction of these three factors constitutes a severe risk factor for the disease.

Node 9: Subset composed by 79 patients with age between 49 and 59 years old, systolic basal pressure higher than 119mmHg and not smokers. The cases belonging to this group do not have high cardiovascular risk.

Node 10: Subset composed by 317 old patients (more than 59 years old), systolic basal pressure higher than 119 mmHg and not smokers. The most of these patients (88.3%) present high cardiovascular risk.

The CT technique does not produce a mathematical model in the same way that LD and LR, but with its information it is possible to define rules for classification [17]. Now we show a summary of rules obtained from the tree model:

A patient has High Cardiovascular Risk when:

- His age is higher than 48 years old; his systolic basal pressure is superior to 119 mmHg and he smokes.
- His age is higher than 58 years old; and his systolic basal pressure is superior to 119 mmHg.

Table 12 presents the classification results of the CT model:

TABLE 12 Classification table for the CT model

Predicted CHRisk using CT model			
		Not high	High
Real CHRisk	Not high	92.9% (751)	7.1% (57)
	High	12.2% (100)	87.8% (720)
Accuracy = 90.4%			

This is the worst model of the three. Its accuracy is just 90.4%, lower than the other two. Besides, there are 100 cases with high risk reported as not high risk. This result is not desirable in medical diagnostic problems because these patients will not be treated.

Table 13 shows the same results considering real data:

TABLE 13 Classification table for the CT model with real data

Predicted CHRisk using CT model			
		Not high	High
Real CHRisk	Not high	99.3% (802)	0.7% (6)
	High	70.7% (29)	29.3% (12)
Accuracy = 95.9%			

The accuracy is now 95.9%, but there are still 29 patients with high risk reported with no high risk.

3.4 Fuzzy Logic based Model

This method does not need to weight data. It works appropriately with unbalanced data.

We will follow some steps. The first step is to build a fuzzy set for each predictor variable. The Experts Committee selects a cutoff point to define an ideal pattern for each predictor variable. For example for the total cholesterol variable the cutoff points are: 200mg/dL and 240 mg/dL. A patient with a total cholesterol lower than 200mg/dL has 0 membership to the fuzzy cholesterol set. A patient with total cholesterol higher than 240mg/dL has the maximum membership to the fuzzy cholesterol set. Finally the membership of a patient with total cholesterol between 200 mg/dL and 240mg/dL is computed by using a triangular membership function.

A fuzzy set is defined as $\bar{A} = \{(x, \mu_{\bar{A}}(x)) \mid x \in X\}$ and the membership function is:

$$\begin{aligned} &\mu_{\bar{A}} : X \to [0,1] \\ &x \in X \to \mu_{\bar{A}}(x) \in [0,1] \end{aligned} \tag{1}$$

where 0 indicates the "no membership" to the set $\bar{A}$ and 1 indicate the whole membership. If $\mu_{\bar{A}}(x) = 0.9$, then the membership of the element x to the fuzzy set $\bar{A}$ is elevated, but if $\mu_{\bar{A}}(x) = 0.1$, the membership of the element x to the fuzzy set $\bar{A}$ is low. The membership function quantifies a degree of belonging of some element to a given set. These functions can be triangular, trapezoidal and so on.

The second step is to build a table with m rows (one for each variable) and n column (one for each patient). Each cell have a membership value ($f_C(a_i)$) representing the evaluation of the a_i patient in the C_j variable. The last column corresponds to the pattern for each variable ($f_P(c)$).Table 14 shows the results of evaluation of the three first patients with all variables.

TABLE 14 Evaluation table for the three first patients

	$f_C(a_1)$	$f_C(a_2$	$f_C(a_3)$	…	**Pattern ($f_P(c)$)**
Age	0	0.6	0	⋯	**0.2**
BMI	0.5	0	0.85	⋯	**0.6**
Smoking	1	1	1	⋯	**1**
MD	0	0	0	⋯	**1**
SBP	0.69	1	0.33	⋯	**0.67**
DBP	0.02	0.5	1	⋯	**0.5**
SP1	1	1	1	⋯	**0.67**
AAP	0	1	1	⋯	**0.67**
Trig	1	0	1	…	**0.58**
TChol	1	1	0.8	⋯	**0.5**
HDCh	1	0	0.09	⋯	**0.25**
HypDg	0	1	1	⋯	**1**

The function $f_{P \Rightarrow A}(c,a)$ is used to determine the degree of incidence of each variable in relation with its pattern. P represents the set of the values of the pattern and A is the set of variables.

To compute $f_{P \Rightarrow A}(c,a)$ we use the formula:

$$\begin{aligned} &f_{P \Rightarrow A}(c,a) = S_L(1 - f_P(c), f_C(a)) \\ &= \min(1, 1 - f_P(c) + f_C(a)) \end{aligned} \tag{2}$$

The third step consists in compute the formula (2) for each value in each row. Numerical results can be organized in a table. We used this formula because:

- If $1 - f_P(c) + f_C(a)$ is higher than 1, then $-f_P(c) + f_C(a)$ is higher than 0, that is $f_P(c)$ is lower than $f_C(a)$ and $f_{P \Rightarrow A}(c,a) = S_L(1 - f_P(c), f_C(a)) = \min(1, 1 - f_P(c) + f_C(a)) = 1$
- If $1 - f_P(c) + f_C(a)$ is lower than 1, then $-f_P(c) + f_C(a)$ is lower than 0, that is $f_P(c)$ is higher

than $fc(a)$ and $f_{P\Rightarrow A}(c,a) = S_L(1-f_P(c), f_C(a)) = \min(1,1-f_P(c)+f_C(a))$ is lower than 1.

If the evaluation of a patient in a variable is higher or equal to the cutoff point, then the evaluation is maxima. The function $f_{P\Rightarrow A}(c,a)$ can be considered as an indicator of the risk for each patient in each variable. These values are averaged by column. This is the final step of the algorithm. Table 15 shows the final results for the three first patients.

TABLE 15 Main results for the three first patients

$f_{P\Rightarrow A}(c,a)$ $= \min(1,1-$ $f_P(c)+f_c(a))$	**a$_1$**	**a$_2$**	**a$_3$**	**…**
Age	0.8	1	0.8	…
BMI	0.42	0.4	1	…
Smoking	1	1	1	…
MD	0	0	0	…
SBP	0.83	1	0.67	…
DBP	1	1	1	…
SP1	1	1	1	…
AAP	1	1	1	…
Triglycerides	0.42	0.42	1	…
TChol	0.5	1	1	…
HDLChol	1	0.75	0.84	…
HypDiag	1	1	1	…
Average	**0.75**	**0.8**	**0.86**	**…**

The last row is an estimation of the risk. It can be classified in high or not high using a cutoff point. After a meeting, the Expert Committee decided to use a cutoff point of 0.85.

With this information, we build a classification table, see table 16.

TABLE 16 Classification table for the FL based model with real data

		Predicted CHRisk using FL based model	
		Not high	High
Real CHRisk	Not high	97.6% (807)	2.4% (33)
	High	34.1% (1)	65.9% (8)
Accuracy = 96.0%			

The accuracy is 96.0%.

4. Better Model Selection

In this research we obtained four models to diagnose patients with High Cardiovascular Risk: LD, LR, CT and FL. The accuracy is calculated in all cases. LR has the better value (97.8%) and the CT the worst one (95.9%).

Frequently, in medical diagnostic problems, it is more important to obtain mathematical models that minimize the probability to classify a sick patient as healthy.

Table 17 summary the classification results presented in Tables 8, 11, 13 and 16.

TABLE 17 Models Summary

Mathematical Model	Accuracy	Cases with high risk classified as healthy
LD	96.1%	14
LR	97.8%	15
CT	95.9%	29
FL	96.0%	1

Common intuition suggests that the classifiers in ensemble should be as accurate as possible and should not make coincident errors [22]. In order to prove this affirmation, we calculated diversity measures for all combinations of our four classifiers.

5. Diversity in Classifier Ensembles

In this section we presented the statistics used to assess the diversity among individual classifiers. There are a lot of diversity measures reported in [23]. We used the disagreement measure, the double fault measure, the correlation coefficient measure and the Q statistic measures.

The Disagreement Measures (D)

In 1966 Skalak proposed the disagreement measure to evaluate the diversity between two classifiers. This measure is based in the idea that the behavior of two classifiers is different on the same data. Diversity within the three or more base classifiers are calculated by averaging all pair of classifiers [24].

The Double Fault Measure (DF)

Giacinto and Roli [25] proposed the double fault measure to select diverse classifiers. This measure arose from the idea that two classifiers perform different to be diverse. The diversity decreases when the value of the double fault measure increases.

The Correlation Coefficient Measure (ρ)

The correlation coefficient (ρ) between the results of two classifiers is calculated. The diversity of two predictors is inversely proportional to the correlation between them. Two classifiers with low correlation are more diverse that two classifiers with high correlation.

To get a single value, the correlation can be averaged across all pairs of classifiers.

The Q Statistic

Q statistic is also a paired measure to determine diversity between two classifiers. Q varies between -1 and 1. Classifiers that tend to recognize the same objects correctly will have positive values of Q. It can be proved that $|\rho| \leq |Q|$

5.1 Calculating Diversity Measures

Four diversity measures were calculated for all combinations of base classifiers. Table 18 shows the obtained results.

TABLE 18 Values of the diversity measures for all combinations

Combinations	D	DF	ρ	Q
LD-CT	0.033	0.023	0.571	0.976
LD-LR	0.023	0.019	0.629	0.992
LD-FL	**0.048**	0.015	0.363	0.922
CT-LR	0.026	0.019	0.609	0.991
CT-FL	0.02	0.031	0.743	0.993
LR-FL	0.029	**0.016**	0.537	0.982
LD-CT-LR	0.027	0.020	0.603	0.986
LD-LR-FL	**0.034**	**0.017**	**0.509**	**0.965**
LD-CT-FL	**0.034**	0.023	0.559	0.964
CT-LR-FL	0.025	0.022	0.630	0.989
LD-CT-LR-FL	0.030	0.021	0.575	0.976

Better combinations according each diversity measures are selected in table 18. The measures suggested the combination of LD, LR and FL classification methods.

5.2 Building a new Model

Probabilities of LD and LR were easily calculated with SPSS software. To compute the probability of each patient in FL based model was need to do two linear transformations:

1. To convert a low risk between 0 and 0.85 in a probability value between 0 and 0.5
2. To convert a high risk between 0.85 and 1 in a probability value between 0.5 and 1

Two transformation formulae were used to solve the problems:

1. To convert a low risk between 0 and 0.85 in a probability value between 0 and 0.5 use the transformation:
   $prob = 0.5(\frac{risk}{0.85})$
2. To convert a high risk between 0.85 and 1 in a probability value between 0.5 and 1 use the transformation:
   $$prob = 0.5 + 0.5(risk - 0.85)/0.15$$

Numerous algorithms have been proposed to combine base classifiers in order to construct a new multiclassifier. Minimum combination rule was used in this paper. Table 19 shows the multiclassifier results:

TABLE 19 Classification table for the multiclassifier

		Predicted CHRisk using multiclassifier	
		Not high	High
Real CHRisk	Not high	96.1% (808)	3.9% (33)
	High	0.0% (0)	100% (8)
Accuracy = 96.11%			

The accuracy of the multiclassifier is not so high: 96.11% in comparison with LR, but it is importance because the number of cases with high risk classified as healthy is 0. That is a very good result from the medical diagnostic point of view.

6. Conclusions

As a main result of the research, four mathematical models to diagnose the High Cardiovascular Risk Index in the Santa Clara city were obtained by applying four different multivariate statistical techniques: LD, LR, CT and FL. Due to the large unbalance present in the classes, it was necessary to weight the data to compute the three first models.

The models obtained classified data quite well. The models obtained classified data quite well. All of they agree with age as the most important risk factor for a High Cardiovascular Risk. Besides there are other risk factors for the disease, for instance: the body mass index, sex, average arterial pressure and smoking, among others.

Some diversity measures were calculated for all combination of the four base classifiers. The better set of combination has three models: LD, LR and FL. A multiclassifier was computed. The accuracy of the multiclassifier was not better than the LR accuracy, but the number of cases with high risk classified as healthy was 0. This result is a valuable tool in the hands of the Project's Committee of Experts to predict the risk of cardiovascular disease in the city of Santa Clara.

References

[1] Mantel, N. and W. Haenszel, *Statistical aspects of the analysis of data from retrospective studies of disease.* J. Natl. Cancer Inst, 1959: p. 719-748.

[2] *CHAID para SPSS sobre Windows. Técnicas de segmentación basadas en razones de verosimilitud Chi-cuadrado.* . 1994.

[3] Agresti, A., ed. *Categorical Data Analysis*. John Wiley & Sons, ed. S. Edition. Vol. 1: Florida.

[4] Press, W., *NUMERICAL RECIPES in C++. The Art of Scientific Computing. Second Edition.* 2005.

[5] Hastie, T., R. Tibshirani, and J. Friedman, *The elements of statistical learning.* 2001.

[6] Jobson, J.D. (1992) *Applied Multivariate Data Analysis Categorical and Multivariate Methods*. **Vol. II**, p. 11-54.

[7] Aytug, H., *Decision Tree Induction*. 2000: University of Florida.

[8] Molinero, L.M., *Modelos de riesgo cardiovascular. Estudio de Framingham.* Proyecto SCORE, 2003.

[9] UCLV, U.d.O.y., *"Proyección del Centro de Desarrollo Electrónico hacia la Comunidad" (PROCDEC)*

[10] Navarro, J.M. and G. Casas, *Estudio del riesgo cardiovascular en el municipio de Santa Clara*

utilizando el método de Regresión Categórica. Operation Research, 2008. **29(3)**.

[11] Eitrich, T., et al., *Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive.* American Chemical Society: J. Chem. Inf. Model, 2007: p. 92-103.

[12] Ye, N., *The Handbook of Data Mining*, in *Lawrence Erlbaum Associates*. 2003: New Jersey.

[13] Chávez, M.C., et al., *Predicción de mutaciones en secuencias de la proteína transcriptasa inversa del VIH usando nuevos métodos para Aprendizaje Estructural de Redes Bayesianas.* Revista Avances en Sistemas e Informática, 2008: p. 77-85.

[14] Denoda Pérez, L., J.L. Morales Martínez, and G. Casas Cardoso, *Software para Análisis Estadístico Borroso (efuzzy)*. 2011: La Habana.

[15] *IBM SPSS Statistics for Windows,* 2012: NY: IBM Corp.

[16] Arbuckle, J.L., *User's Guide, Version 21.0, IBM® SPSS® Amos™ 21.* IBM Corp. Released 2012. IBM SPSS Statistics for Windows.

[17] SPSS, M.d.S., 2003.

[18] Hogg, R.V., *Probability and Statistical Inference.* Maxwell Mac-millan International. New York, 1993.

[19] Hair, J.F., ed. *Análisis multivariante*. Quinta Edición ed., ed. P. Hall. Vol. 1. 1999.

[20] Devore, J. and R. Peck, eds. *Statistics: The Exploration and Analysis of Data.* Third Edition ed. 1997, Wadsworth Publishing Company: California, U.S.A.

[21] DeGroot, M.H., ed. *Probability and Statistics*. ed. Addison-Wesley. 1987.

[22] Kuncheva, L.I.W., C. J, *Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. . Machine Learning*, 2003. **51**: p. 181-207.

[23] Kuncheva, L.I., *Combining Pattern Classifiers: Methods and Algorithms.* New York, NY, Wiley Interscience., 2004.

[24] Skalak, D.B., *The Sources of Increased Accuracy for Two Proposed Boosting Algorithms.* 1996.

[25] Giacinto, G.R., F., *Design of effective neural network ensembles for image classification purposes.* 2001.