Machine Learning Methods for Knowledge Discovery in Medical Data on Atherosclerosis

José Ignacio Serrano[1], Marie Tomečková[2], Jana Zvárová[2]
1. Instituto de Automática Industrial, CSIC, Madrid, Spain,
2.Department of Medical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

Machine learning techniques are methods that given a training set of examples infer a model for the categories of the data, so that new (unknown) examples could be assigned to one or more categories by pattern matching within the model. The data from follow-up studies with repeated collection of the same type of data are very suitable for this analysis. Machine learning algorithms belonging to a variety of paradigms have been applied to knowledge discovery on medical data. All the used algorithms belong to the supervised learning paradigm. Several algorithms have been tested, trying to cover most of the kinds of supervised learning. Two kinds of experiments have been carried out. The first is intended to discover associations between attributes. The second kind is intended to test prediction of future disorders. For the experiments in this paper the data used was from the twenty years lasting primary preventive longitudinal study of the risk factors (RF) of atherosclerosis in middle aged men. Study is named STULONG (LONGitudinal STUdy). The results show that some methods predict some disorders better than others, so it is interesting to use all the algorithms at a time and consider the result confidence based upon the known tendency of each method. The machine learning algorithms have been also used in the prediction of death cause, obtaining poor results in this case, maybe due to the small amount of information (entries) of this type in the dataset.

Keywords: knowledge discovery, supervised machine learning, biomedical data mining, risk factors of atherosclerosis

1. Introduction

Machine learning techniques [1] are methods that given a training set of examples infer a model for the categories of the data, so that new (unknown) examples could be assigned to one or more categories by pattern matching within the model.

Machine learning techniques have been applied successfully to a high variety of problems and data for prediction tasks. The main objective is to research how to apply machine learning algorithms to this data in order to discover relationships between attributes and to make predictions that could be useful for decision support. Medical data is a special kind of data, because many different kinds of features are involved in the collections. Moreover, the medical data have several known problems: missing, incorrect and sparse information and temporal data. Machine learning methods are very suitable for this kind of data [2]. There exists several KDD works attempting to deal with large-scale medical information. In [3], authors try to detect type of hepatitis by extracting short sequential patterns from the temporal features. In [4], simple rules are discovered using 4ft-miner (i.e. statistical tables of two rows and two columns), in order to temporally characterize, by differences, the hepatitis types B and C. Authors in [5] attempt to discover rules of singles boolean features that can be able to predict the liver fibrosis stage. The same application appears in [6] but, in this case, extracted patterns are clustered and then these clusters are assigned to fibrosis stages depending on the covered examples. They also applied this technique to atherosclerosis risk [7]. The data from follow-up studies with the repeated collection of the same type of data are very suitable for this analysis. Additional examples of data mining on biomedical data are presented in [8] and [9]. For the experiments in this paper the data used was from the twenty years lasting primary preventive longitudinal study of the risk factors (RF) of atherosclerosis in middle aged men. The study is named STULONG (LONGitudinal STUdy) [10]. The main target of this study is to validate machine learning as a way of association mining and to validate classification performance as a measurement of the salience for the discovered association. It is also intended to test machine learning algorithms in the prediction of far future disorders.

In the next section, the details of the STULONG dataset are presented. In section 3, the machine learning algorithms tested are described. Section 4 presents the measures for evaluation and Section 5 describes validation experiments. Finally, concluding remarks and future work is presented in Section 6.

2. Description of the Study and Data Set

The STULONG (http://euromise.vse.cz/challenge2004/index.html) [10], [11] data were collected by the 2nd Department of Internal Medicine, 1st Faculty of Medicine of Charles University in Prague and General University Hospital, Prague and transferred to the electronic form and analysed by statistical methods by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University in Prague and the Academy of Science of the Czech Republic, Prague.

The main aims of the study were:

1. To identify the prevalence of risk factors (RF) of atherosclerosis in a population generally considered as the most endangered by possible complications of atherosclerosis, i.e. middle aged men.
2. To follow the development of these RF and their impact on the examined men health, especially with respect to atherosclerotic cardiovascular diseases.
3. To study the impact of complex intervention of RF on their development and cardiovascular morbidity and mortality in men.

Men born in 1926–1937 and living in Prague 2 were selected from the Prague 2 election lists in year 1975. For the first examination, 1419 of 2370 invited men came. Entry examinations were performed in the years 1976–1979. The invitation for examination included a short explanation of the aims of the study, of the first examination purpose, procedure and later observations and asked for co-operation. At that time, no informed signature of the respondent was required. Should the man react to the first invitation for the examination, we considered that a sufficient agreement with the examination itself, observation and results processing. Should man fail to react to the first invitation, we would send two more invitations, minimally.

The risk factors were defined according to the definitions at that time as follows:

- hypertension – blood pressure BP ≥ 160/95 mm Hg or men under the hypotensive medication,
- hypercholesterolemia – cholesterol ≥ 260mg% (6,7 mmol/l),
- hypertriglyceridemia – triglycerides ≥ 200mg% (2,2 mmol/l),
- smoking: ≥ 15 cig./day currently or smoking of the same number of cigarettes within 1 year prior to the beginning of the study (pipe or cigar smokers were considered non-smokers),
- overweight: Brocka index > 115 % (Brocka index: height in cm minus 100 = 100 %),
- positive family case history: death of father or mother from coronary artery disease, or vascular stroke before reaching 65 years of age.

Men were divided according to presence of risk factors (RF), overall health conditions and ECG result into following groups:

- **NG** = group of men without RF defined above, without manifestation of the atherosclerotic diseases or other serious illnesses making their ten-year-long observation impossible, and without ECG changes,
- **RG** = group of men with at least one RF defined above, without manifestation of the atherosclerotic diseases or other serious illnesses making their ten-year-long observation impossible, and without ECG changes,
- **PG** = group of men with a manifested cardio-vascular atherosclerotic diseases or other serious diseases making their ten-year-long observation impossible (e.g. malignant illness, advanced failure of liver or kidneys, serious neurological or psychological problem). The pathologic group included also men with diabetes treated with orally administered anti-diabetics or insulin, and men with pathologic ECG, according to the Minnesota ECG code.

Long-term observation of patients was based on their division into the groups stated above:

- The risk group **RG** was randomly divided into two sub-groups designated as **RGI** (intervened risk group) and **RGC** (control risk group). The patients in the **RGI** group were invited for check up minimally twice a year. Following pharmacological intervention, they were invited as necessary. The patients in the **RGC** group received a short written notice including their laboratory results and ECG description and a recommendation to take these results to their physician; possible intervention of RF was left to the decision of these physicians. At the first examination, no significant difference in age, socio-economic factors or RF occurrence was demonstrated between the RGI and RGC groups.
- 10 % of men in the **NG** group was examined minimally once a year just as the risk group – (they are denoted **NGS**); In this group of men, similarly to the risk group, intervention was initiated as soon as a RF was identified and confirmed (hyperlipidemia, arterial hypertension). The remaining men of the **NG** were invited for a control check up 10–12 years later.
- The men from the **PG** group were excluded from further observation.

Intervention was the key problem of the study and was based on non-pharmacological influence. We tried to modify and to optimize RF.

- *Non-pharmacological intervention:* interviews on lifestyle, i.e. diet, physical activity, suitability or necessity to stop smoking and reduce weight. The interviews were repeated during each control and except for general instructions, they focused also on specific RF of a given man.
- *Pharmacological intervention:* treatment of arterial hypertension and hyperlipoproteinemia – was very limited in the initial stages of the study and may be mostly used only in the last years of the study. The pharmacological therapy was recommended with respect to the overall risk of a given man and his possible other diseases.

Four data files have been used for the analysis:

1. The file *ENTRY* contains values of 244 attributes obtained from entry examinations for each man; these attributes are either codes or results of size measurements of different variables or results of their transformations (identification of man, family and personal history, social factors – education, physical activities, smoking, eating habits, alcohol, after them anthropometric measurements – height, weight, skin folds, physical examination with measurement of blood pressure, pulse, laboratory values and coding of ECG).
2. The file *CONTROL* contains results of observation of 66 attributes recorded during control examinations. There are attributes corresponding to identification, to habit changes, to personal history, physical examination and biochemical values, and data about hypertension, hypercholesterolemia, hypertrigliceridemia and other coronary and oncological diseases. This file consists of 10,572 records of long observation.
3. Additional information about health status of 403 men dropped out in the time of the study was collected by the postal questionnaire. Resulting values of 62 attributes are stored in the *LETTER* file.
4. There are 5 attributes concerning death of 389 patients. Values of these attributes are stored in the *DEATH* file. It contains attributes for the identification of the patients and the date and cause of death.

3. Description of the Used Methods

All the used algorithms belong to the supervised learning paradigm. That is, a learning stage is needed in order to build a model over the training examples and then use this model to predict the category of unknown examples. Several algorithms have been tested, trying to cover most of the kinds of supervised learning. Each of the used methods is very briefly explained next:

3.1 Naive Bayes

Naïve Bayes [12] calculates, for each pair attribute-value, for example *(education, university)*, the probability of belonging to each category, by dividing the number of examples of the target category where the pair appears by the total number of examples where the pair appears. Thus, each pair will have a probability for each tentative category. Naive Bayes is based on the assumption that every pair attribute-value within an example is independent on each other. Thus, when an unlabeled example is classified, the probability for each category of the example is the multiplication of the probability for the corresponding category of each of the pairs that form the example. The predicted category is the one with the highest probability.

3.2 Multilayer Perceptron

The classification model of the Multilayer Perceptron Neural Network [13] is composed of a certain number of layers of neurons interconnected between them. The architecture used for this dataset is presented in Figure 1.

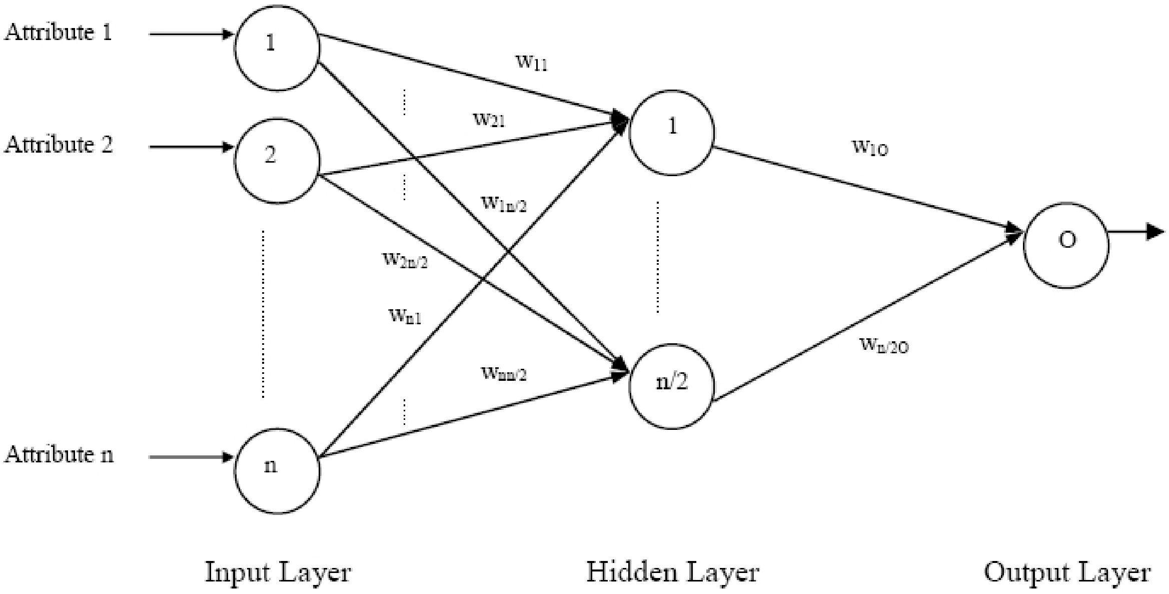


Fig. 1. Architecture of the Multilayer Perceptron Neural Network used.

Each connection has an associated weight. The input to each neuron is the weighted sum, using the association weights, of all the incoming values. The output of each neuron is the result of applying a function. In this case, a typical sigmoid function is implemented in all the neurons. Figure 2 shows the function expression and representation.

$$f(x) = \frac{1}{1 + e^{-x}}$$

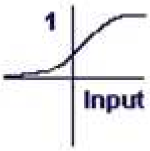


Fig. 2. Expression and representation of sigmoid function.

Thus, each of the attribute values from a sample of the dataset is entered in the corresponding neuron of the input layer, and the values spread through the network to the output layer, where the output value of the neuron is the predicted class.

The training phase consists of, given a set of initial weights values, entering each of the labelled examples of the training dataset into the model and comparing the output value with the expected class. Depending on the error of the predicted class, the back propagation algorithm changes weights from the output layer to the input layer, in order to make the predicted value to be more similar to the expected one. This process is carried out a certain number of epochs or iterations. In this case, this number is equal to 500. The amount the weights are changed in back propagation, so called learning rate, is 0.3, and the momentum applied to the weights during updating is 0.2. If the back propagation algorithm does not reach a good approximation to the expected output after one iteration, then it resets the model and causes the learning rate to decrease.

3.3 Support Vector Machines (SVM)

Support Vector Machines [14] try to separate examples, based on their category, in the *n*-dimensional space, being *n* the number of attributes or features, by hyper planes of the form ***w* + *b***, so that

$$x\,w + b \geq +1 \rightarrow category = true$$

$$x\,w + b \geq -1 \rightarrow category = false$$

x being the example represented as a vector of *n* components. Here, ***w*** is the support vector, perpendicular to the hyper plane, and correspond to examples that are beyond or over the limits of their category (see Figure 3).

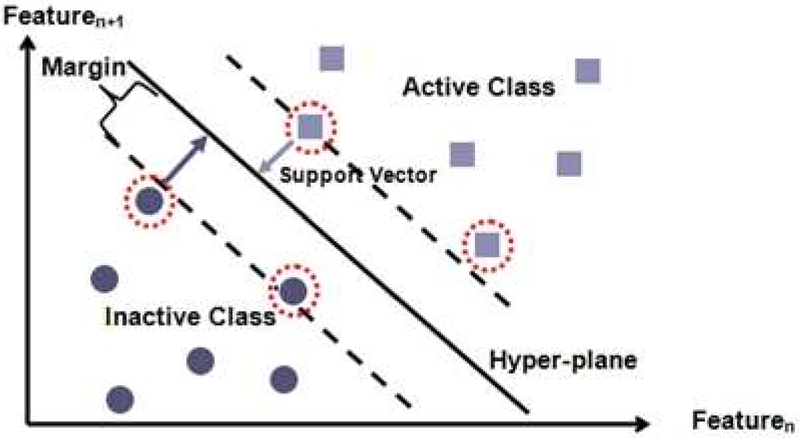


Fig. 3. Support vectors scheme.

The support vector also defines, by its module, a margin of one between the hyper plane and the first positive and negative examples (that the reason for +1, -1 thresholds). For each category, the algorithm tries to find ***w*** maximizing the margin. To classify an unlabeled example the algorithm simply applies the expression above. This is a simple implementation of the method and the one used in the experiments, but there are other more sophisticated implementations and techniques.

3.4 K-Nearest Neighbour

KNN is a memory-based algorithm [15], with the background idea that past experiences can help us to solve present ones by analogy. It considers each example as a vector of *n* components, being *n* the number of attributes or features. It does not need a learning stage. To predict the class of an unlabeled example, the algorithm compares the input example with each of the examples in the training data or memory, by calculating the distance between them. Then, the majority class of the *K* most similar training examples is the one predicted for the input example. The distance used in the experiments is the Euclidean distance between vectors. However, there are more possibilities in the literature.

3.5 ID3 and C4.5 Decision Trees

The model produced by this algorithm is a tree [16], where each node corresponds to an attribute and each arc of the node corresponds to a possible value of the node attribute.

The learning algorithm constructs the tree from the training data. The selection of the attribute that will form a node, at each moment, is carried out by calculating the entropy of the data after the selection of the node. That is, for each attribute, the entropy of the remaining data without the attribute, separated by the different values of the node attribute, is calculated. Thus, the attribute that produces the minimum entropy is the selected for the node. The process goes on until there is no more attributes or the number of remaining examples under a node is lower than a certain threshold. In the former case, the majority class of these remaining examples is the one settled under the node. In Figure 4, we can see an example:

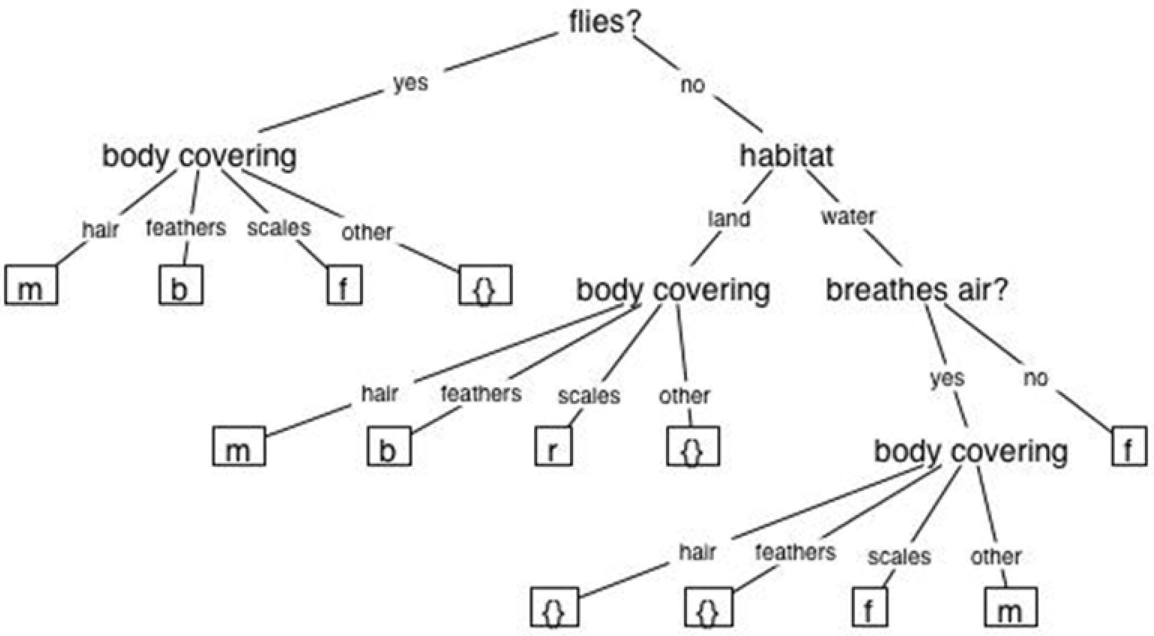


Fig. 4. Example of a decision tree.

In the example there are 4 attributes: *flies*, *body covering*, *habitat* and *breathes air*, and four possible categories, *m*, *b*, *f* and *r* . Here, the first attribute is *flies* because it is the one that produces the division on the data with minimum entropy at that level, and so on. To classify an unlabeled example you only have to follow the tree top-down, and the final leaf is the predicted category. The pathways from the root node to the leaf node can be viewed as rules, where the condition is formed by AND operation of the terms *(node=arc)*.

C4.5 is an extension of ID3 that allows continuous numerical attributes, accounts for missing values and carries out a pruning process in order to reduce the tree size for dealing with larger amount of data. The J48 tree used in the experiments is an implementation of C4.5.

3.6 Ridor Rules Learner

Ridor stands for the RIpple-DOwn Rule learner [17]. It generates the default rule first and then the exceptions for the default rule with the least (weighted) error rate when it is used to classify the training data. Then it generates the "best" exceptions for each exception and iterates until pure. Thus it performs a tree-like expansion of exceptions and the leaf has only the default rule but no exceptions. The exceptions are a set of rules that predict the classes other than class in the default rule. IREP is used to find out the exceptions. IREP algorithm constructs rules by gradually adding one term in the condition at a time so that the error rate is minimized. The rule condition terms are like *(attribute {=,≠,≤,≥} value).*

4. Evaluation

The evaluation processes and measures are the same for all the experiments. Given the data, a part of the collection is considered as a training set and the remaining as a test set. So the models learn from the training set and try to predict the values of examples in the test set. Since the category of test set examples is known, we can check the predictions. Three different typical measures are calculated for each category: precision, recall and F-measure [18]. Precision is the percentage of predictions of one category that were correct. Equation 1 presents the precision expression.

$$Precision\ (category_i) = \frac{number\ of\ correct\ predictions\ as\ category_i}{total\ number\ of\ predictions\ as\ category_i} \qquad (1)$$

Recall is the percentage of all the examples of the test set belonging to a category that were correctly predicted. The expression is presented in Equation 2.

$$Recall\ (category_i) = \frac{number\ of\ predictions\ as\ category_i}{total\ number\ of\ examples\ of\ category_i} \qquad (2)$$

F-measure is a combination of the former measures. It accounts, someway, the intersection between the examples involved in precision and recall, normalized by the sum of both. Equation 3 shows the F-measure expression.

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

Thus, these three measures are calculated for each category of the test set. As said before, given the collection it is needed to divide the data in training and test sets. A common way of evaluation is cross-validation. The collection is divided into *n* equally sized parts. Then, each *n-1* part combination is considered as training and the remaining part as test, so the algorithm is run *n-1* times, and the final results are the average of this *n-1* executions. For all the experiments described below, *n* has a value of 3, so training is always 66 % of the data and test stands for 33 %, running each algorithm three times. Usually, the value of *n* is greater than 3, typically of 10, but in this case we have very few examples of some categories, and a greater value of *n* could produce test sets with no representation of the mentioned categories, what is not desirable.

5. Experiments

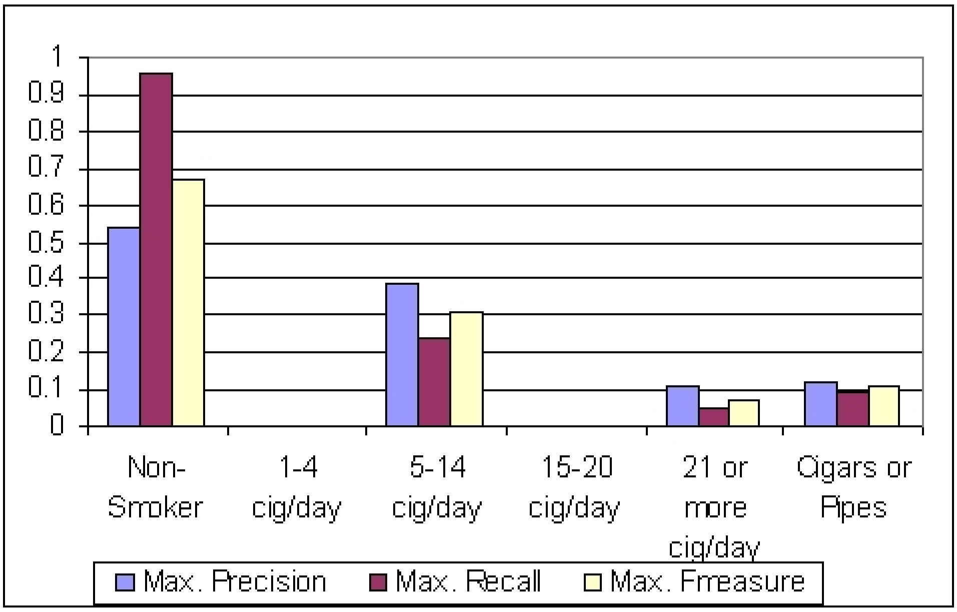
Two kinds of experiments have been carried out. The first is intended to discover associations between attributes by considering the classification performance as an indicator of the association strength. The second kind is intended to test the prediction of future disorders.

It is needed to remark that the observations in the data set with missing values were not removed nor imputed, because the implementations of the learning algorithms are able to deal with missing data. These implementations are the ones included in the WEKA environment [19], used with default parameters to perform the experiments above.

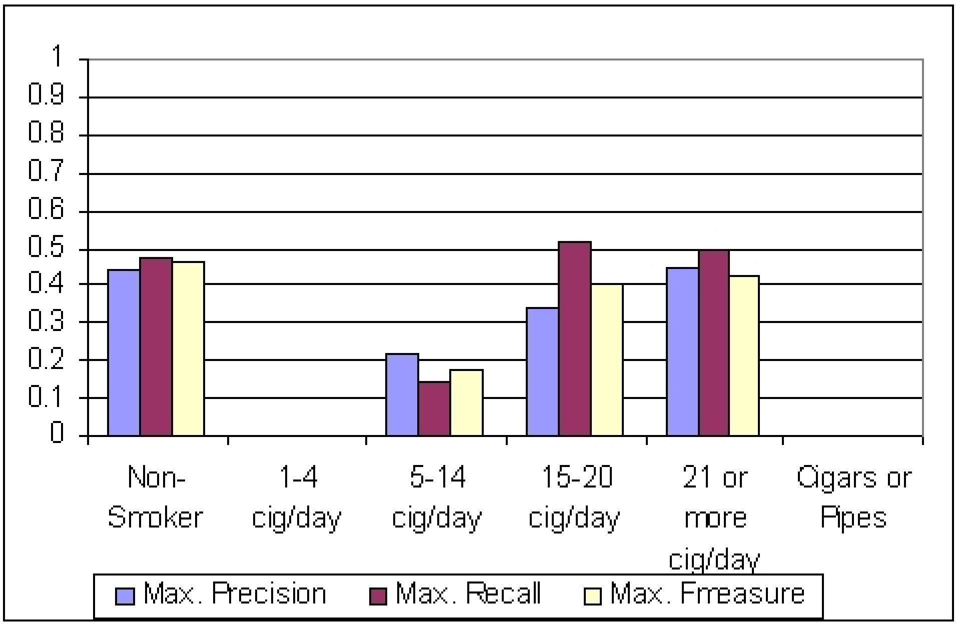
5.1 Finding Answers

The first experiments are related to the analytical questions proposed for the Discovery Challenge of ECML/PKDD 2004 conference, specifically the ones related to the Entry collection. These tasks consist of finding relations in three different groups of patients: normal group, risk group and pathologic group. These groups correspond to the risk level of atherosclerosis – see above, and will be referenced as level groups. Specifically, the target relationships are between social factor features and physical activities features, alcohol features, smoking features, body mass index, blood pressure and HDL cholesterol, and then between physical activities and the remaining and between alcohol and the remaining. So, machine learning algorithms are applied to the data of each different group, trying to predict the value of each of the features of one group given the features of the other, viewing the possible values as the considered categories. Thus, for example, given the four social factor attributes as training factors the algorithms are run in order to predict the value of each of the four physical activities attributes and so on with the other features groups. For each relationship, the maximum values over all the different algorithms results are calculated in order to compare between level groups. So, if the prediction accuracy is good, we could say that there is a strong relationship, in a degree equal to the accuracy, between the features used for training and the feature whose values are predicted, and also we can compare prediction measures between features and level groups to state which relations are stronger than others.

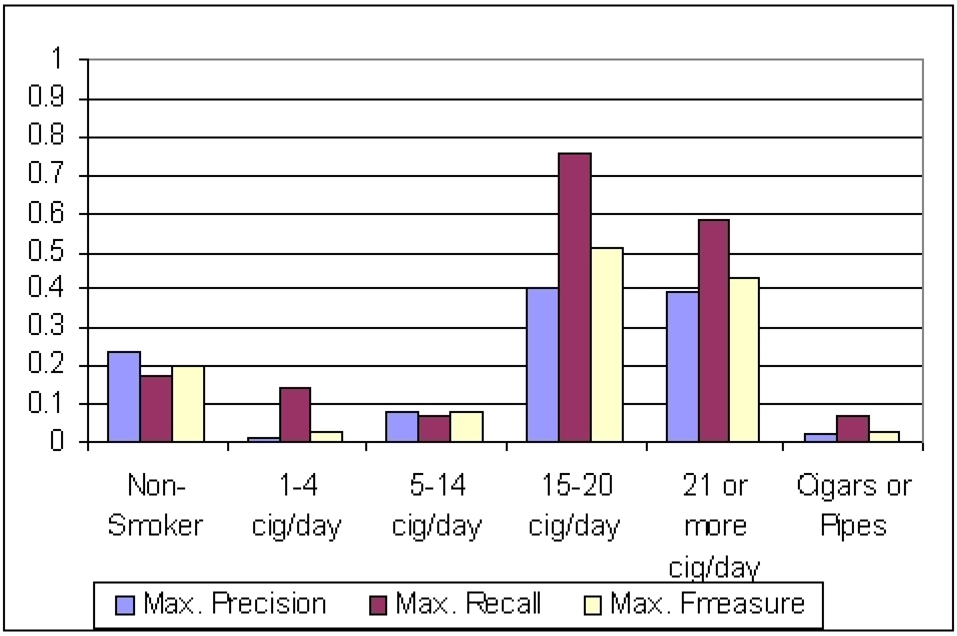
Due to paper length limitations, only some of the most representative results are presented. In Figure 5, the maximum precision, recall and F-measure predictions results for "Smoking", given social attributes, are showed for each of the level groups, a) Normal, b) Pathologic and c) Risk, respectively, and given physical activity attributes for each of the level groups, d) Normal, e) Pathologic and f) Risk, respectively. As can be seen, in the Normal group, either from social factors or physical activity, the best prediction is reached for non-smoking people, being not significant for the remaining values of the "Smoking" attribute. It seems that the relationship between social factors and smoking is slightly stronger than physical activity and smoking, because it produces better results for all the values of the "Smoking" attribute. In both Pathologic and Risk groups, the relationship between the training factors and the non-smoking value is stronger for physical activity factors, being in particular high in the Pathologic group. In the latter groups, people who smoke 15 or more cigarettes a day are better predicted than in the Normal group but non-smokers are much worse detected than in the Normal group.



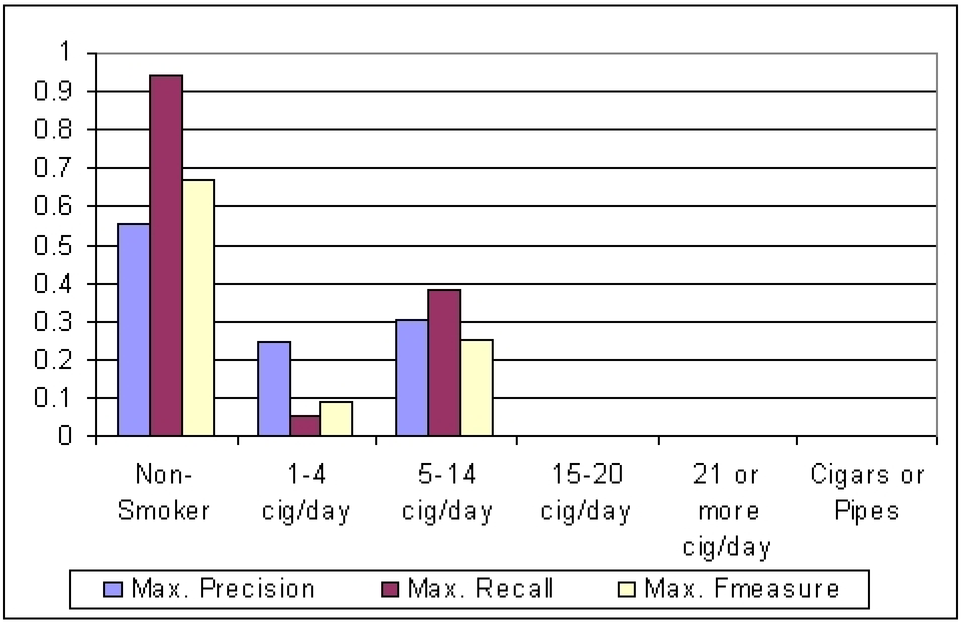
a)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Non-Smoker
1-4 cig/day
5-14 cig/day
15-20 cig/day
21 or more cig/day
Cigars or Pipes
Max. Precision
Max. Recall
Max. Fmeasure

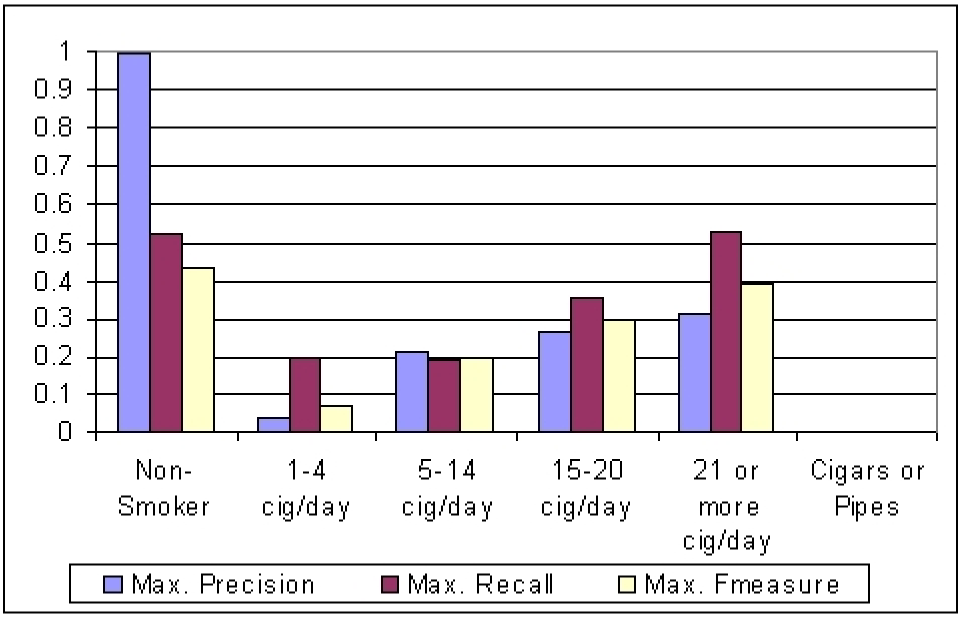
b)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Non-Smoker
1-4 cig/day
5-14 cig/day
15-20 cig/day
21 or more cig/day
Cigars or Pipes
Max. Precision
Max. Recall
Max. Fmeasure

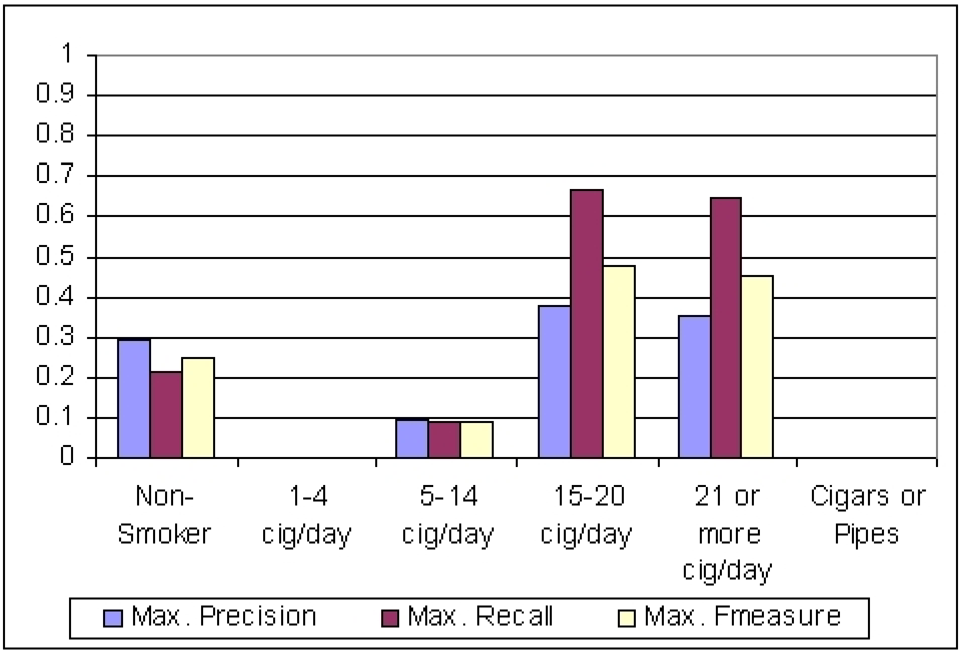
c)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Non-Smoker
1-4 cig/day
5-14 cig/day
15-20 cig/day
21 or more cig/day
Cigars or Pipes
Max. Precision
Max. Recall
Max. Fmeasure

d)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Non-Smoker
1-4 cig/day
5-14 cig/day
15-20 cig/day
21 or more cig/day
Cigars or Pipes
Max. Precision
Max. Recall
Max. Fmeasure

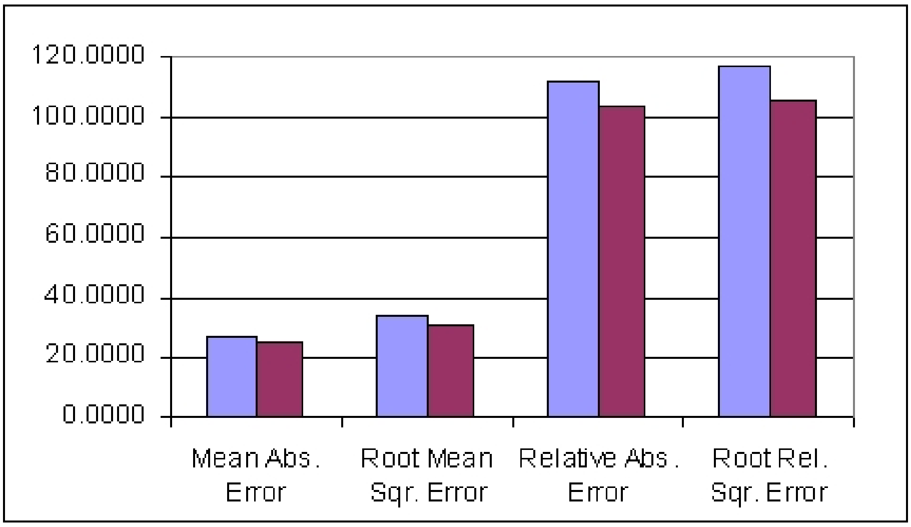
e)



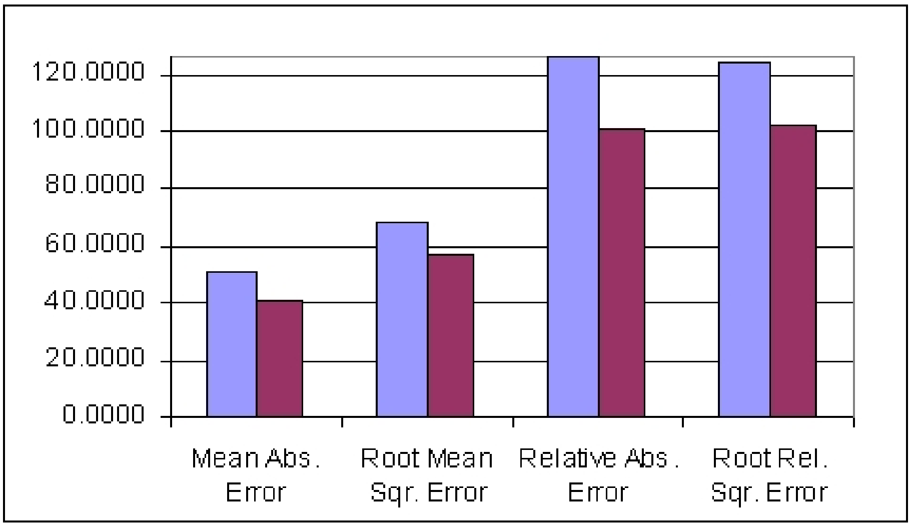
f)

Fig. 5. Maximum precision, recall and F-measure values over all the algorithms for the prediction of "Smoking" attribute, given only the social factors on a) Normal group, b) Pathologic group and c) Risk group, and given only the physical activity factors on d) Normal group, e) Pathologic group and f) Risk group.

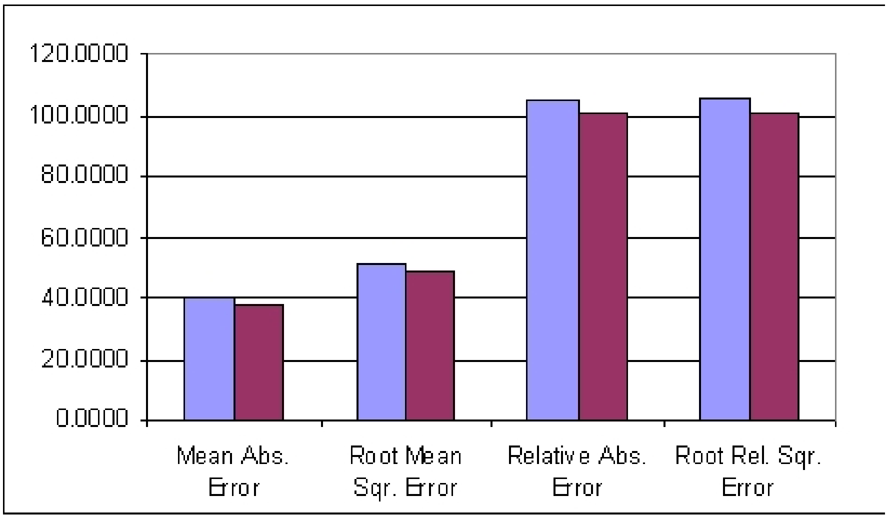
Let us see another representative example. Figure 6 presents the results of the prediction of cholesterol level from social factors, a), b), and c), and from physical activity factors, d), e) and f), for each of the level groups, respectively. In this case, the prediction results are very similar for the relationship between social factors and cholesterol, and physical activity and cholesterol, in all level groups, so we can conclude that the strength of the relationships is similar, too. However, it varies among level groups. In the Normal group, the mean absolute prediction error is about 24, being about 50 and 40 in Pathologic and Normal groups, respectively, concluding that it is easier to predict cholesterol, from both social factors and physical activity as training, for people in the Normal group. This fact denotes a strong relationship between the training factors and the cholesterol level in the latter group.


120.0000
100.0000
80.0000
60.0000
40.0000
20.0000
0.0000
Mean Abs. Error
Root Mean Sqr. Error
Relative Abs. Error
Root Rel. Sqr. Error

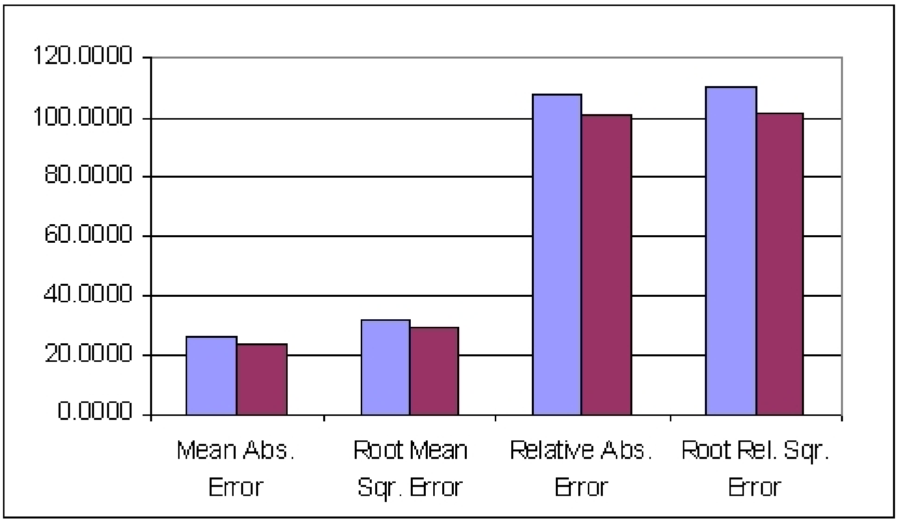
a)


120.0000
100.0000
80.0000
60.0000
40.0000
20.0000
0.0000
Mean Abs. Error
Root Mean Sqr. Error
Relative Abs. Error
Root Rel. Sqr. Error

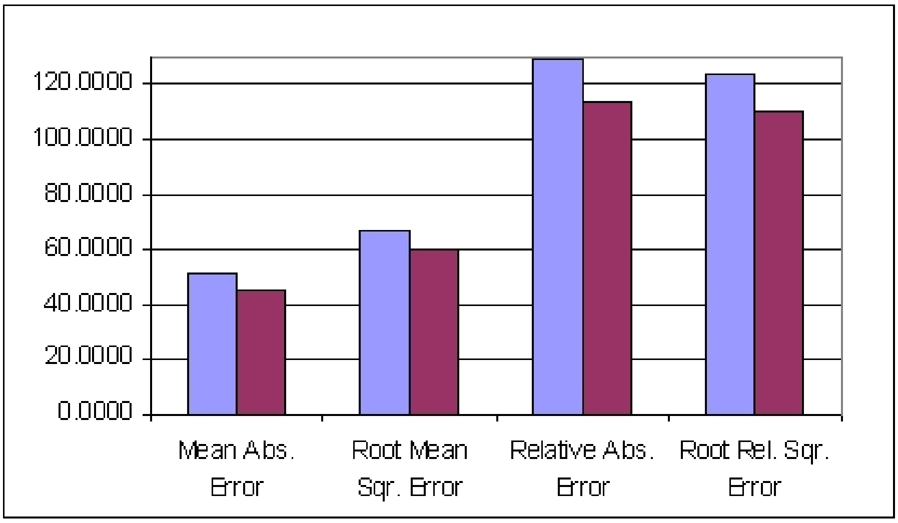
b)


120.0000
100.0000
80.0000
60.0000
40.0000
20.0000
0.0000
Mean Abs. Error
Root Mean Sqr. Error
Relative Abs. Error
Root Rel. Sqr. Error

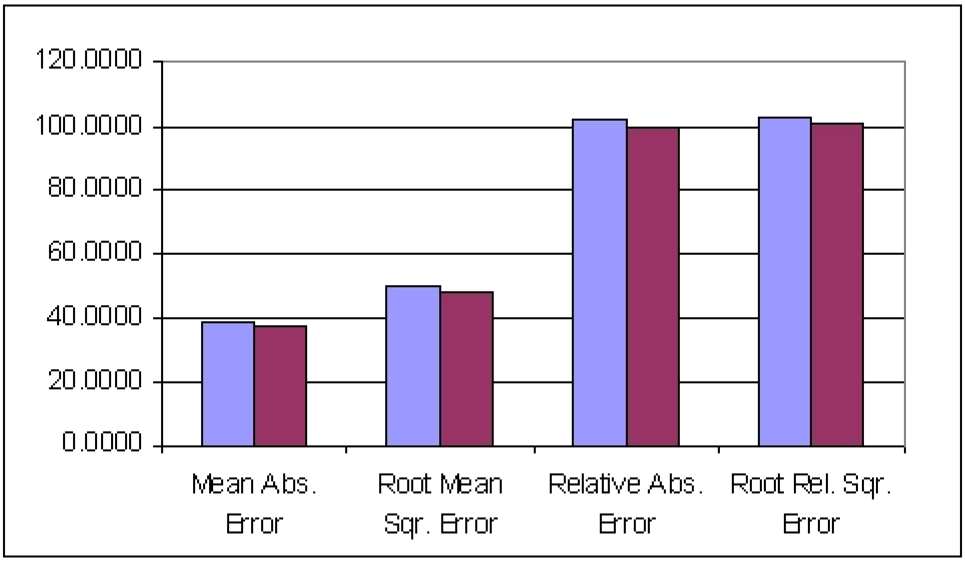
c)


120.0000
100.0000
80.0000
60.0000
40.0000
20.0000
0.0000
Mean Abs. Error
Root Mean Sqr. Error
Relative Abs. Error
Root Rel. Sqr. Error

d)


120.0000
100.0000
80.0000
60.0000
40.0000
20.0000
0.0000
Mean Abs. Error
Root Mean Sqr. Error
Relative Abs. Error
Root Rel. Sqr. Error

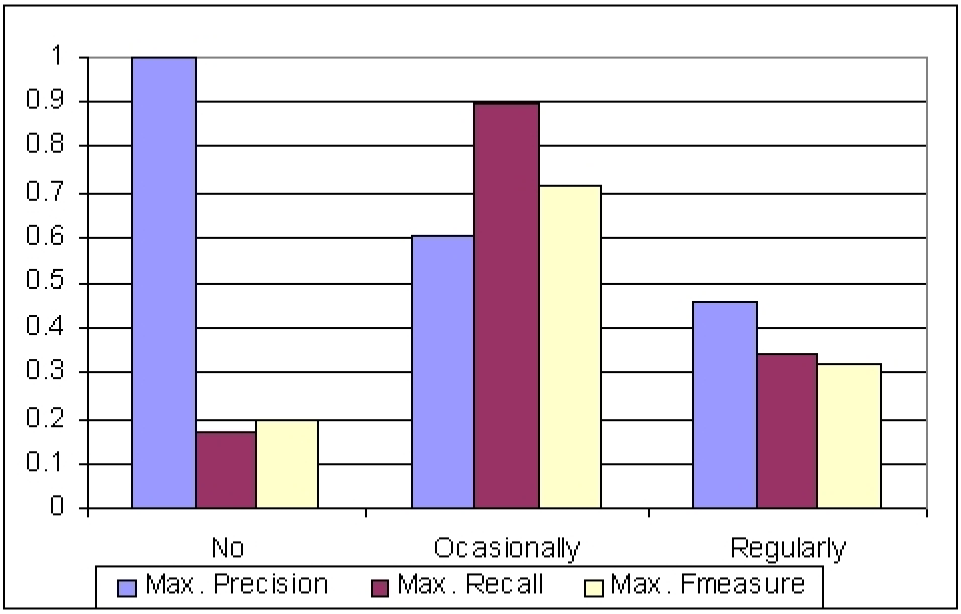
e)



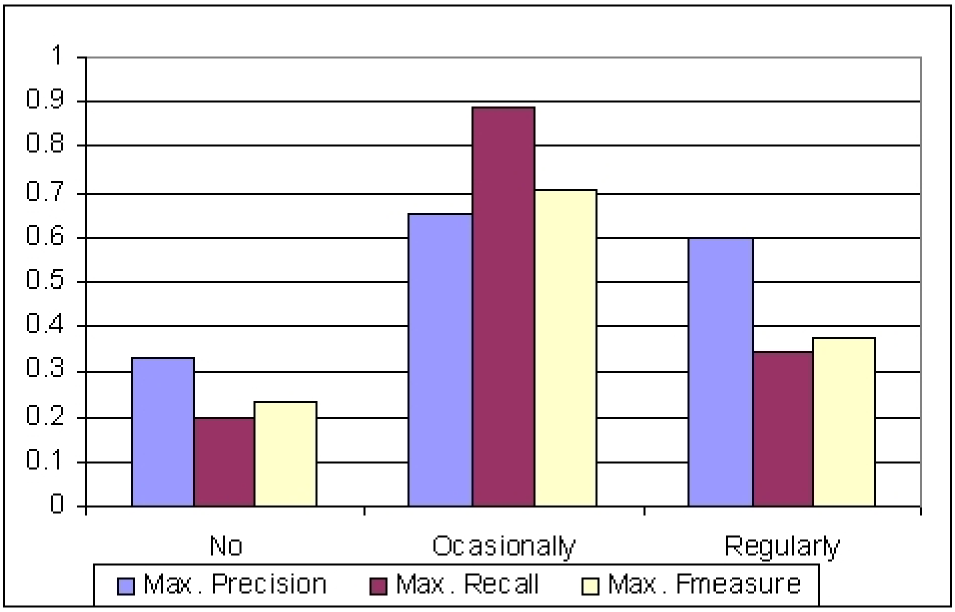
f)

Fig. 6. Average and maximum mean absolute error, root mean square error, relative absolute error and root relative square error values over all the algorithms for the prediction of cholesterol level attribute, given only social factors on a) Normal group, b) Pathologic group and c) Risk group, and given only the physical activity factors on d) Normal group, e) Pathologic group and f) Risk group.

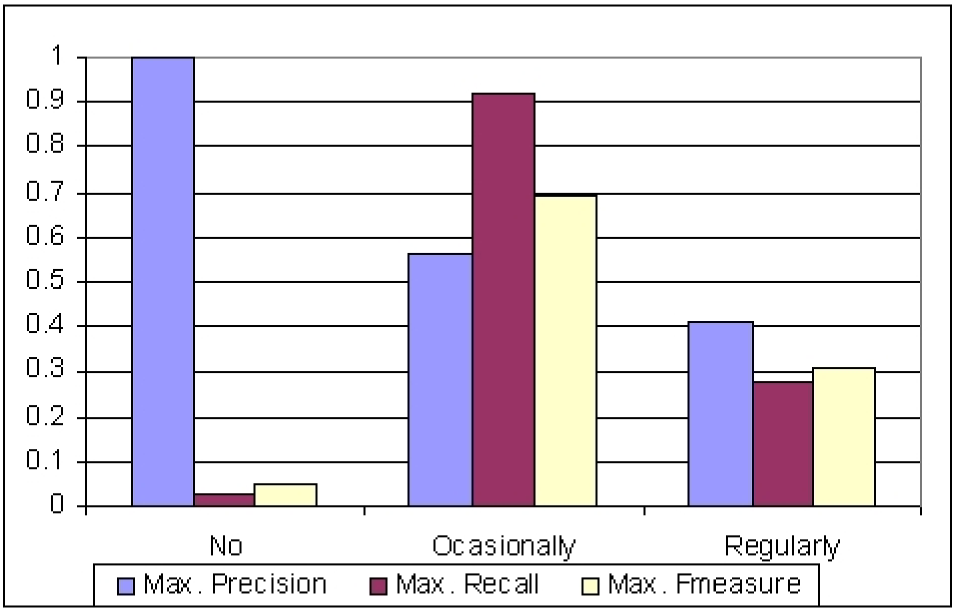
Finally, Figure 7 shows the results for the prediction of alcohol attribute values, separately from social factors and physical activity factors as training, for each level group, analogous to above.



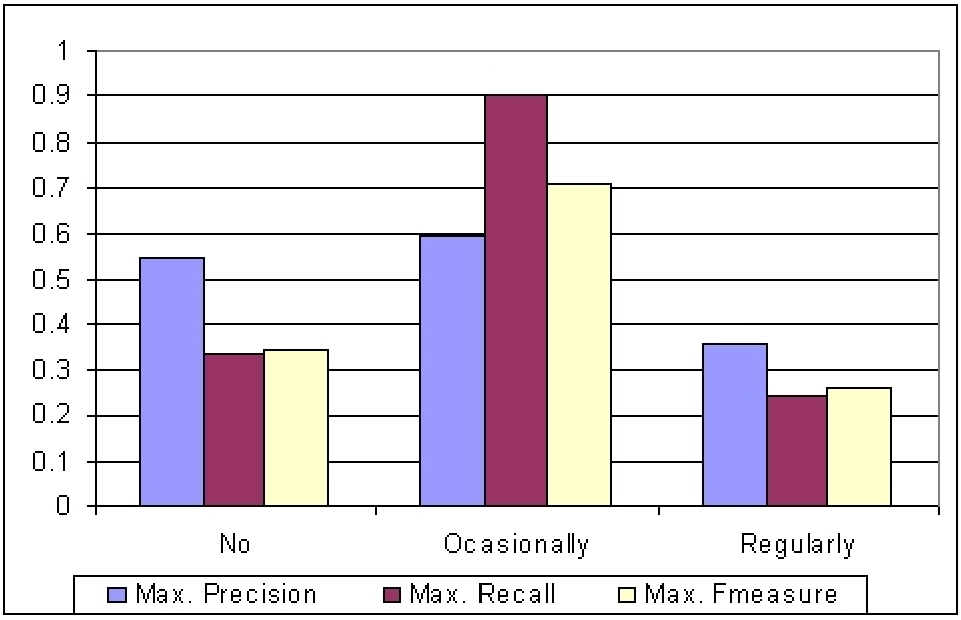
a)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
No
Ocasionally
Regularly
Max. Precision
Max. Recall
Max. Fmeasure

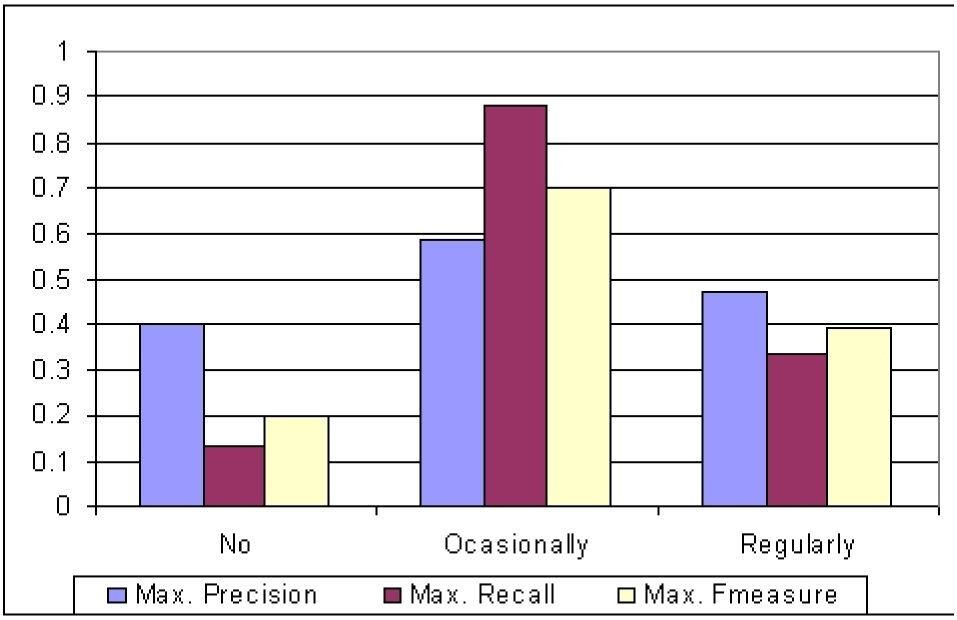
b)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
No
Ocasionally
Regularly
Max. Precision
Max. Recall
Max. Fmeasure

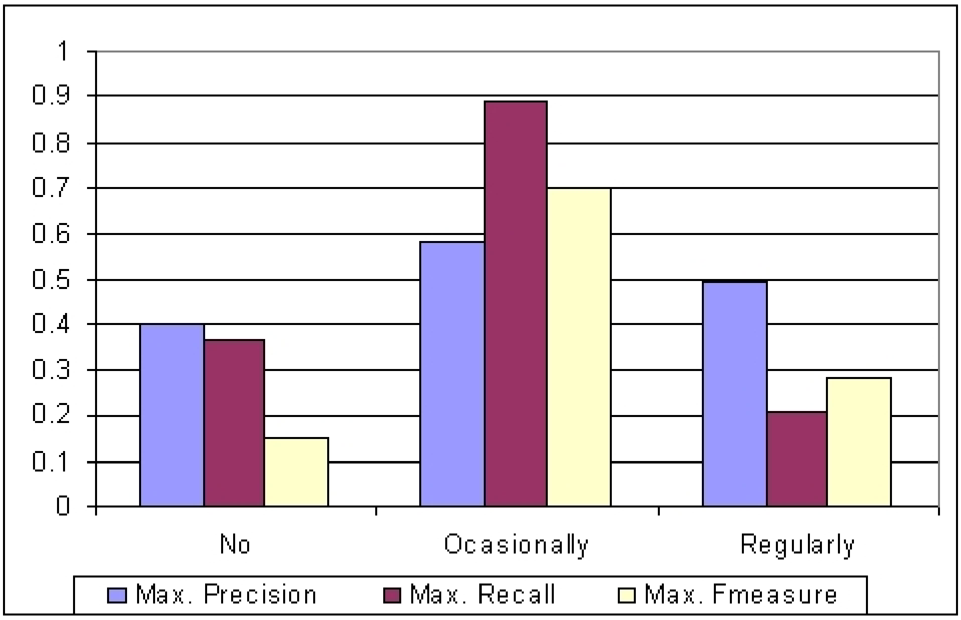
c)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
No
Ocasionally
Regularly
Max. Precision
Max. Recall
Max. Fmeasure

d)


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
No
Ocasionally
Regularly
Max. Precision
Max. Recall
Max. Fmeasure

e)



f)

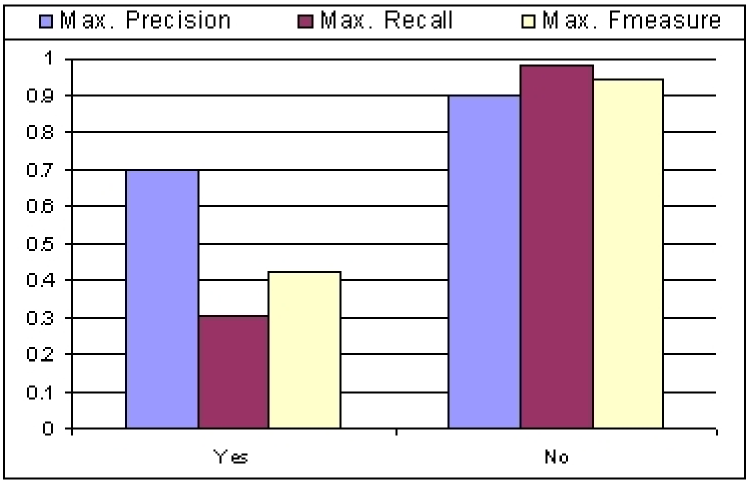
Fig. 7. Maximum precision, recall and F-measure values over all the algorithms for the prediction of the "Alcohol" attribute, given only the social factors on a) Normal group, b) Pathologic group and c) Risk group, and given only the physical activity factors on d) Normal group, e) Pathologic group and f) Risk group.

The results in Figure 7 show that there is a clear relationship in all levels of groups between the training factors and the people who drink alcohol occasionally. People who drink regularly are more difficult to detect and predict from the training factors, resulting in a light relationship that is a little bit stronger in the Pathologic group. The same can be said about people who never drink alcohol in relation to physical activity factors. However, the prediction precision is significantly increased for social factors in Normal and Risk groups. People who never drink are accurately identified from their social factors in the latter groups, what denotes a significant relationship between the involved attributes.

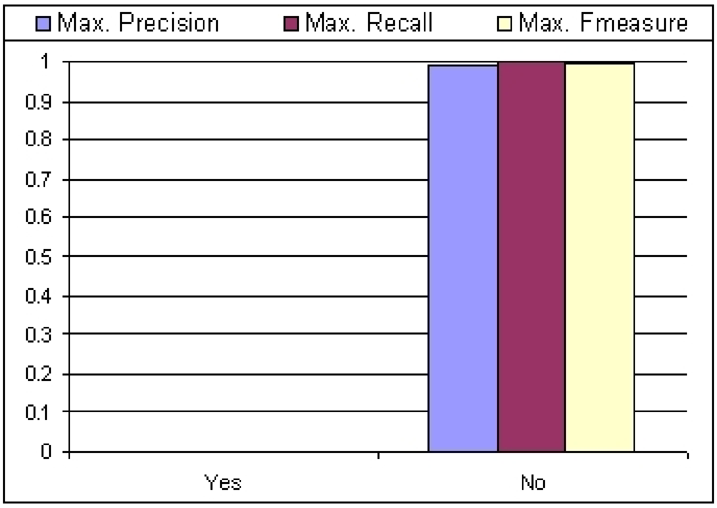
The training features groups are taken with all the attributes at once. From the medical point of view it is also interesting to separate these attributes and to try subsets of them. So, it was tried to predict the value of the physical activity in the job attribute given all possible combinations of social factors attributes, for example. The results show that, for Normal and Risk groups, the "Education" feature alone obtains much better prediction results than any other combination of social factors attributes. In the Pathologic group it is similar, but the difference is not so high as in the other groups, being "Age + Education" the best combination.

5.2 Predicting Future Disorders

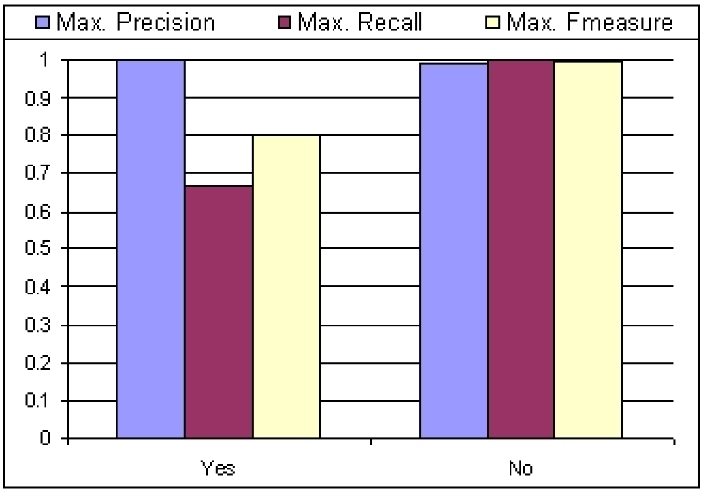
The main objective of the next experiments is to test the prediction accuracy of the algorithms. The Entry collection is not the only one used but also the Control collection is considered. First of all, the patients who have a control record in the Control collection, after ten years from their entry in the study, were selected. Then using their Entry attributes it was tried to predict whether they will have some disorders in ten years. These disorders correspond to systolic-diastolic hypertension, systolic hypertension, diastolic hypertension, hypercholesterolemia and hypertriglyceridemie. The possible values for these disorder attributes are true or false. The same has been done for twenty years records. The results show that the multilayer perceptron was the best algorithm, reaching values near 85 % of precision and 65 % of recall in the detection of all the disorders. The risk of future hypertension in the Risk group is 0 for many men, while some patients in this group were hypertensive since the beginning of the study. From the medical point of view it is more interesting to carry out the experiments only on the Normal group. The same process has been done separately for this group, for ten and twenty years. The results for the different mentioned disorders are presented in Figure 8, a) to e), respectively, for ten years prediction and f) to j), respectively, for twenty years prediction. For each disorder, the maximum values over all the different algorithms results are presented. In this case, results show that there is not one best algorithm. Depending on the disorder to predict and also on certain categories, one algorithm fits better than others (the maximum values presented correspond to different algorithms), so it will be interesting to use all the algorithms and make decisions based on the results from all of them. As a comment, we pointed that the prediction accuracy is much higher than when entries of the three levels of groups are considered all together, confirming the early interest in the Normal group.



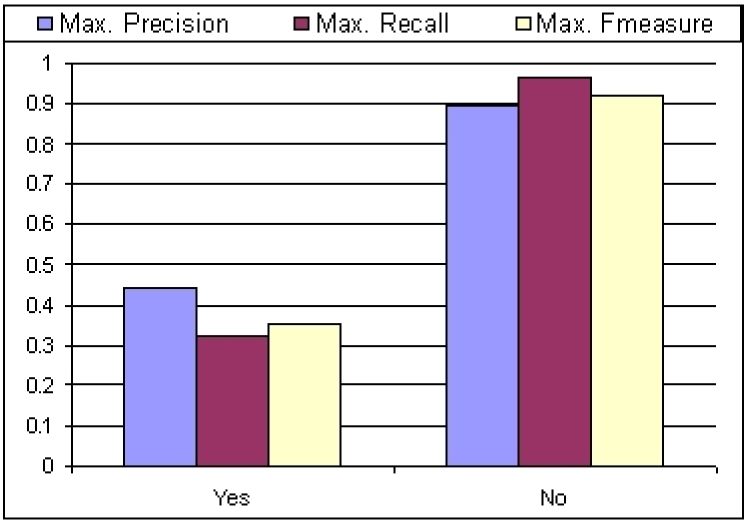
a)


Max. Precision
Max. Recall
Max. Fmeasure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Yes
No

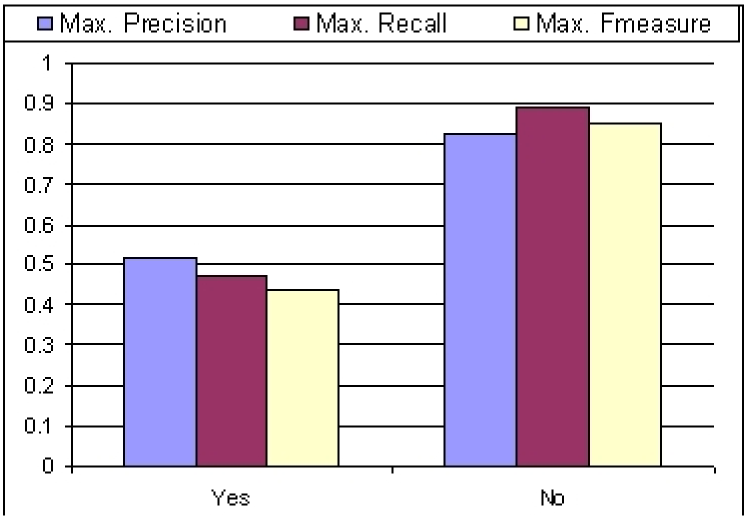
b)


Max. Precision
Max. Recall
Max. Fmeasure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Yes
No

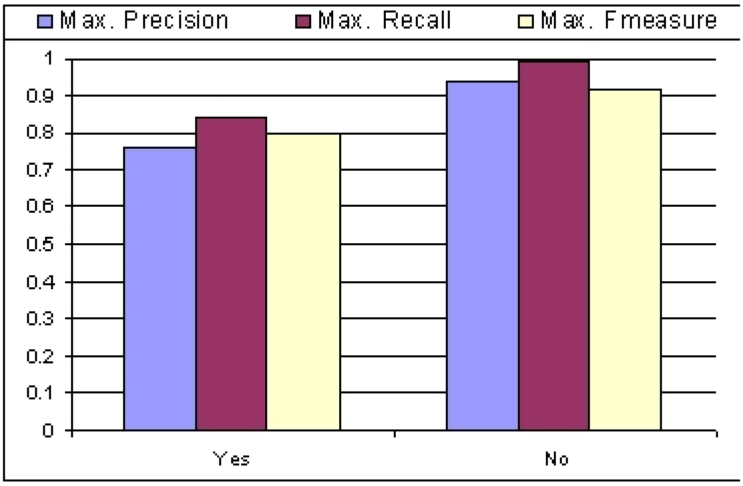
c)


Max. Precision
Max. Recall
Max. Fmeasure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Yes
No

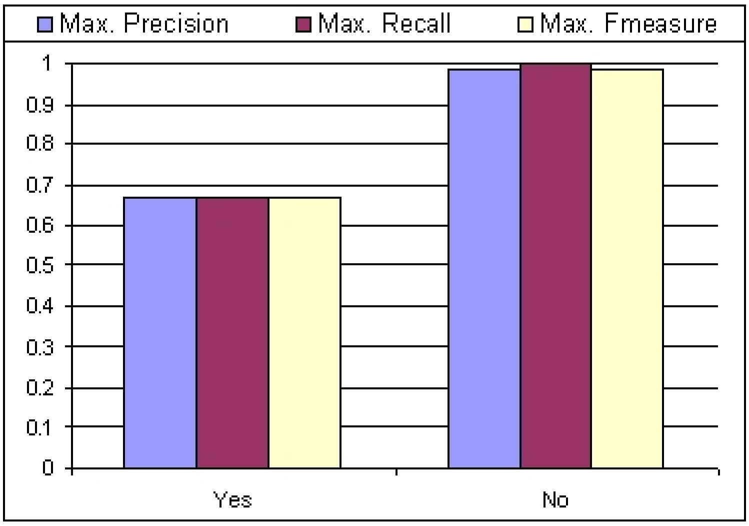
d)


Max. Precision
Max. Recall
Max. Fmeasure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Yes
No

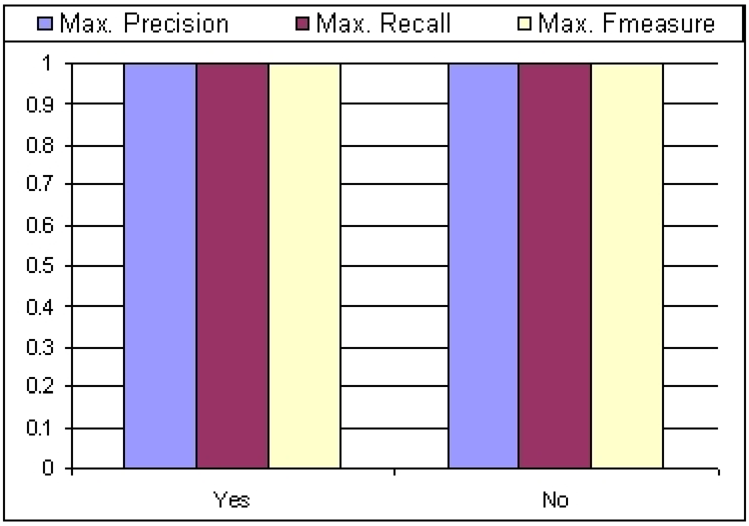
e)


Max. Precision
Max. Recall
Max. Fmeasure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Yes
No

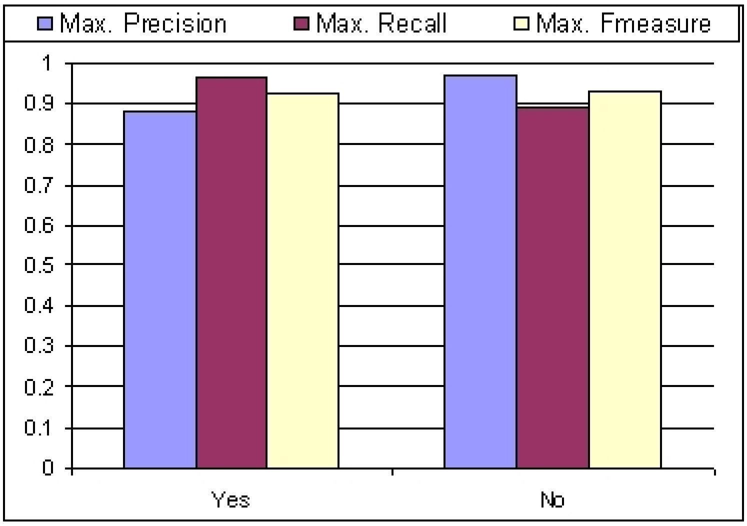
f)


Max. Precision
Max. Recall
Max. Fmeasure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Yes
No

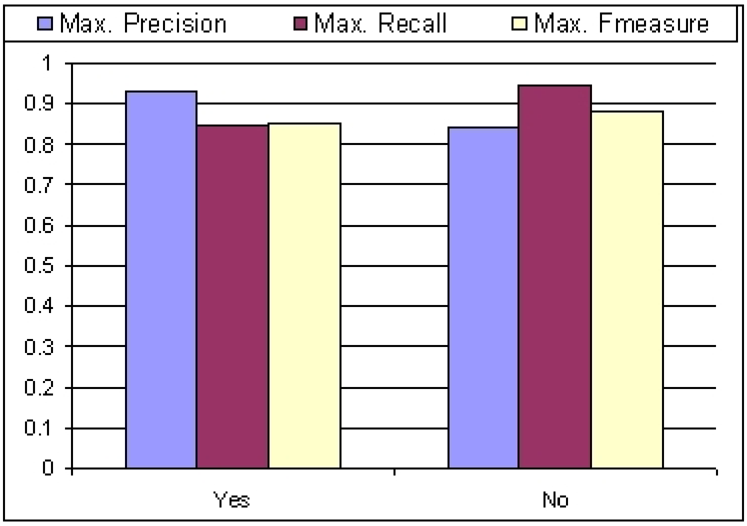
g)


Max. Precision
Max. Recall
Max. Fmeasure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Yes
No

h)


Max. Precision
Max. Recall
Max. Fmeasure
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Yes
No

i)



j)

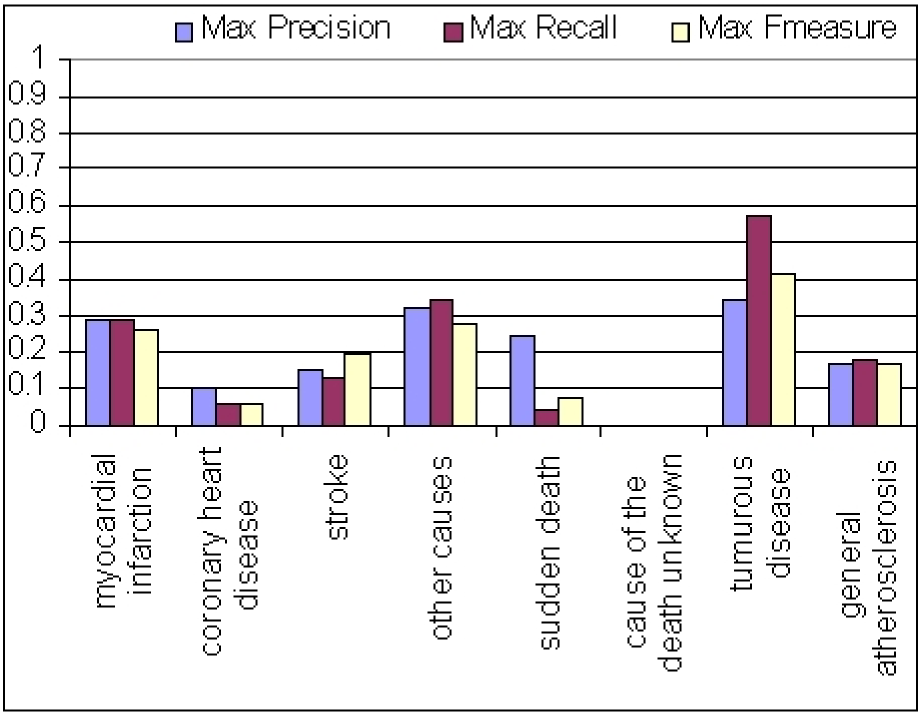
Fig. 8. Maximum precison, recall and F-measure values for the prediction of a) systolic-diastolic hypertension, b) systolic hypertension, c) diastolic hypertension, d) hypercholesterolemia and e) hypertriglyceridemie in ten years, and f) systolic-diastolic hypertension, g) systolic hypertension, h) diastolic hypertension, i) hypercholesterolemia and j) hypertriglyceridemie in twenty years.

The values of Figure 8 show that it is more accurate to predict disorders in twenty years than to predict them in ten years, specifically the prediction of the presence of the disorders, which is accurately inferred in twenty years but very poorly predicted in ten years in all the disorders but diastolic hypertension. The non-presence of the disorders is equally well-predicted for both ten and twenty years. Among all the disorders the best detected is diastolic hypertension, obtaining prediction values near to 100 % accuracy for the presence and non-presence of the disorder. The worst predicted disorder is systolic hypertension, with the presence of the disorder non-detectable at all in ten years.

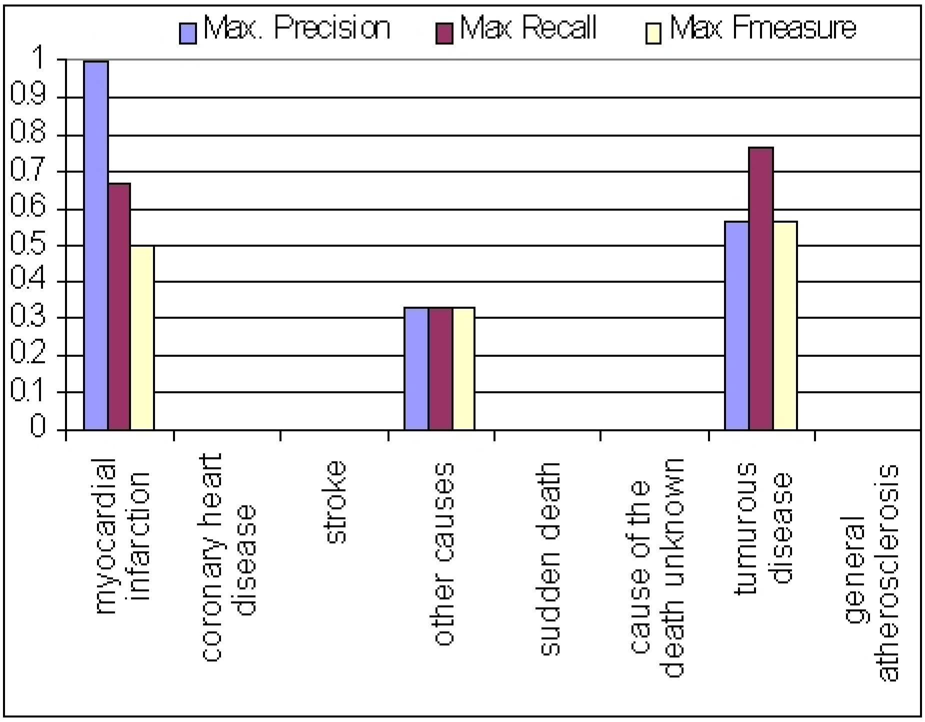
It was also tried to predict some other diseases, like angina pectoris, myocardial infarction, cerebron-vascular accident and so on, but there is a small number of observations with these features, so the results are not relevant.

5.3 Predicting Death Cause

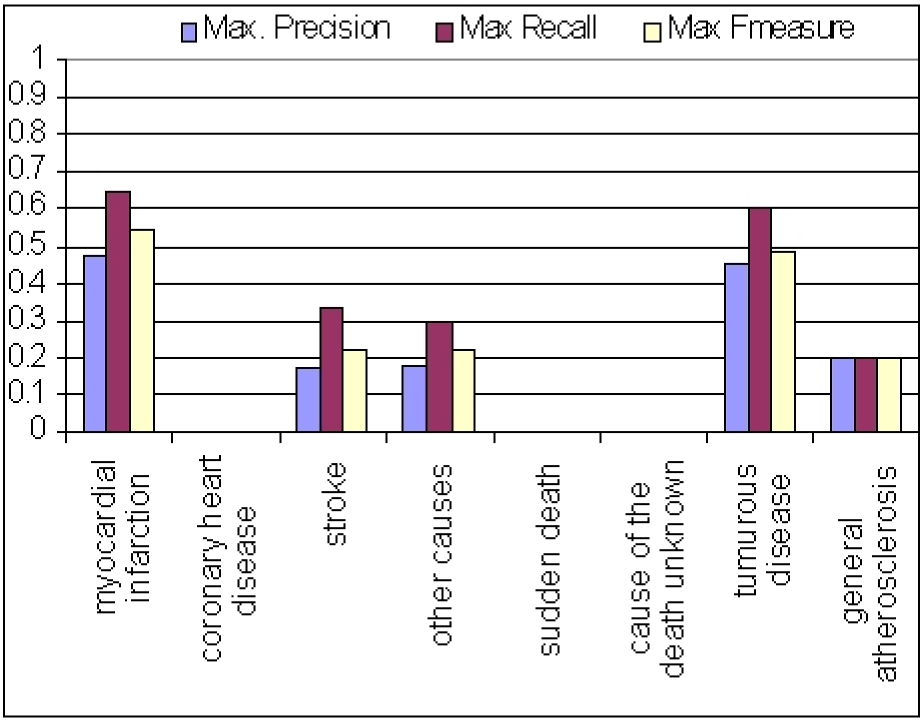
This experiment is analogous to the last one, but now what is tried is the prediction of the cause of death rather than diseases or disorders. Thus, the Death collection is used. The algorithms were trained with the data in the Entry collection for the patients of the Death collection. The experiments were carried out for the three levels of groups separately and for all the entries of all the groups together. The results are presented in Figure 9. In the Normal group, Figure 9b), the best predicted causes were tumour disease and other causes. In the Risk group, Figure 9d), the best prediction was for other causes but also for myocardial infarction and coronary heart disease, that were not predicted at all in the Normal group. In the Pathologic group, Figure 9c), the best predicted causes were tumour disease and myocardial infarction, but stroke and general atherosclerosis could be poorly predicted, too, obtaining much lower results for these latter causes in the other groups. In general, Figure 9a), the results of prediction of death cause are very poor, concluding that data from the Entry collection has not enough information to predict death and/or also maybe more observations are needed. But, what is sufficient for it?



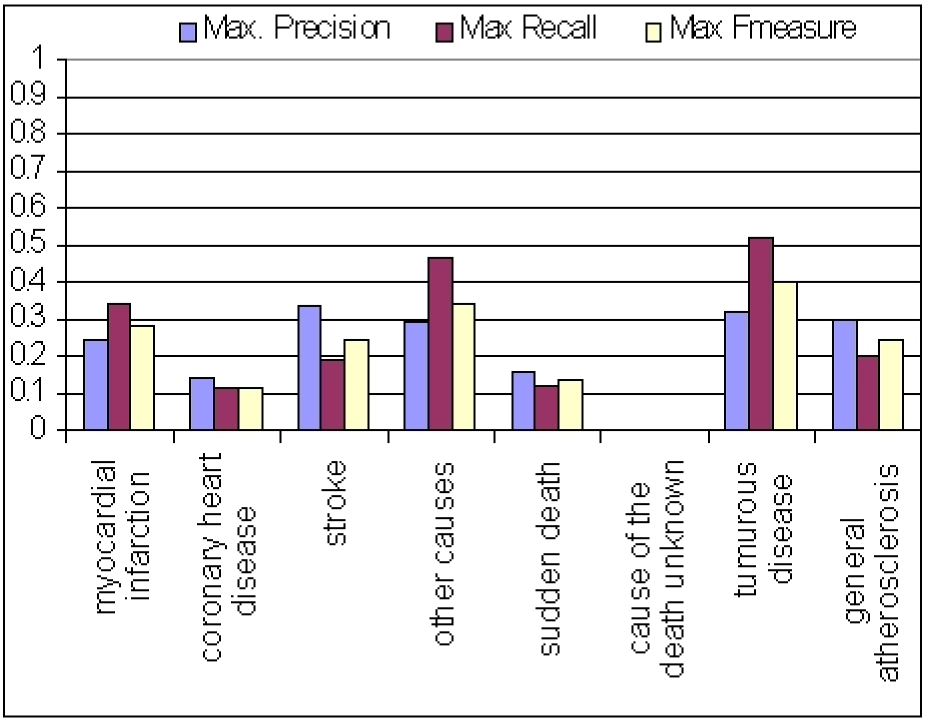
a)



b)



c)



d)

Fig. 9. Maximum values of precision, recall and F-measure, in the prediction of death causes, for a) all the level groups as one, b) Normal group, c) Pathologic group and d) Risk group.

6. Conclusions

Machine learning algorithms belonging to a variety of paradigms have been applied to knowledge discovery on medical data in two different ways: firstly, the methods have been used in order to predict the value of one attribute of the patient database, given a subset of other attributes as training features, proposing the maximum accuracy among all the algorithms as a measure of the strength of the relationship between those training features and the target attribute. This measure has been proven useful also for comparing the relationships between attributes in different groups of patients.

Secondly, the learning techniques have been applied to the prediction of future disorders. The results show that some methods predict some disorders better than others. Then, it is interesting to use all the algorithms at a time and consider the result confidence based upon the known tendency of each method. All the tested methods perform better for twenty years prediction than for ten years predictions, reaching excellent results for some of the disorders that make the methods suitable for decision support. The machine learning algorithms have been also used in the prediction of death cause, obtaining poor results in this case, maybe due to the small amount of information (entries) of this type in the dataset.

It would be interesting for the future to finely tune the parameters of the algorithms and to test more techniques. It is also intended to integrate all methods with the degree of significance and usefulness discovered in this work in order to build an expert system, and the derivation of rules understandable by humans from the results of the system will be also researched.

Acknowledgement

Research was partially supported by the Research Plan of ICS AS CR AV0Z10300504 and the “María Bueno” Research Visits Plan of the Spanish Council for Scientific Research together with the support of the Industrial Automation Institute.

Reference

[1] Mitchell, T.: Machine Learning. McGraw Hill, 1997.

[2] Lavraĉ, N.: Selected Techniques for Data Mining in Medicine. Artificial Intelligence in Medicine, vol. 16 (1), pp. 3-23, 1999.

[3] Aseervatham, S. and Osmani A.: Mining Short Sequential Patterns for Hepatitis Type Detection. ECML/PKDD Discovery Challenge, 2005.

[4] Aubrecht, P., Kejkula, M., Kremen, P., Novakova, L., Rauch, J., Simunek, M., Stepankova, O.: Mining in Hepatitis Data by LISp-Miner and SumatraTT. ECML/PKDD Discovery Challenge, 2005.

[5] Pizzi, L.C., Ribeiro, M.X., Vieira, M.T.P.: Analysis of Hepatitis Dataset using Multirelational Association Rules. ECML/PKDD Discovery Challenge, 2005.

[6] Durand, N., Soulet, A.: Emerging Overlapping Clusters for Characterizing the Stage of Liver Fibrosis. ECML/PKDD Discovery Challenge, 2005.

[7] Durand, N., Cleuziou, G., Soulet, A.: Discovery of Overlapping Clusters to Detect Atherosclerosis Risk Factors. ECML/PKDD Discovery Challenge, 2004.

[8] Cios, K. J.: Medical data mining and Knowledge Discovery. Physica – Verlag, 2001.

[9] Chen, H., Fuller, S. S., Friedman, C. and Hersh, W.: Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Integrated Series in Information Systems (2), Springer Science and Business Media Inc., 2005.

[10] Boudik F., Reissigova J., Hrach K., Tomeckova M., Bultas J., Anger Z., Aschermann M., Zvarova J.: Primary Prevention of Coronary Artery Disease Among Middle Aged Men in Prague: Twenty-year Follow-up Results. Atherosclerosis. 2006 Jan;184(1):86-93.

[11] Tomeckova, M.: The Challenge on Atherosclerosis Data Viewed by the Experts. ECML/PKDD Discovery Challenge, 2004.

[12] Rish, I.: An Empirical Study of the Naive Bayes Classifier. IJCAI-01 Workshop on Empirical Methods in AI, 2001.

[13] Haykin, S.: Neural Networks: A comprehensive Foundation (2nd edition). Pearson Education, 1998.

[14] Scholkopf, B., Smola, A. J., Mtiller, K.-R., Burges, C. J. C., and Vapnik, V.: Support Vector Methods in Learning and Feature Extraction. In Down, T., Frean, M., and Gallagher, M., editors. Proceedings of the Ninth Australian Congress on Neural Networks, Brisbane, Australia. University of Queensland, 1998.

[15] Teknomo, K.: K-Nearest Neighbors Tutorial. http:people.revoledu.comkardi tutorialKNN, 2004.

[16] Quinlan, J. R.: C4.5: Programs for Machine Learning. Morgan Kauffman, 1993.

[17] Compton, P., Edwards, G., Kang, B., Malor, R., Menzies, T., Preston, P., Srinivasan, A. and Sammut, S.: Ripple Down Rules: Possibilities and Limitations. Boose, J.H. & Gaines, B.R., Ed. Proceedings of the Sixth AAAI Knowledge Acquisition for Knowledge-Based Systems Workshop. pp.6-1-6-20. Calgary, Canada, University of Calgary, 1991.

[18] Van Rijsbergen, C. J.: Information Retrieval. Butterworths, London, 1979.

[19] Witten, I. H. and Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.